Published in final edited form as:

Infect Genet Evol. 2012 March ; 12(2): 453–460. doi:10.1016/j.meegid.2011.04.028.

Phylogenetics of the HIV-1 CRF02_AG clade in Cameroon

Nuno Rodrigues Faria¹, Marc A Suchard², Ana Abecasis³, J. D. Sousa¹, Nicaise Ndembu⁵, R.J. Camacho³, Anne-Mieke Vandamme¹, Martine Peeters⁶, and Philippe Lemey¹

¹Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium ²Departments of Biomathematics and Human Genetics, David Geffen School of Medicine; Department of Biostatistics, School of Public Health; University of California, Los Angeles, United States ³Centro de Malária e outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical, Lisboa, Portugal ⁴Laboratório de Biologia Molecular, Centro Hospitalar de Lisboa Ocidental, Lisboa, Portugal ⁵MRC/UVRI Uganda Research Unit on AIDS, Entebbe, Uganda ⁶UMR145, Institut de Recherche pour le Développement (IRD) and University of Montpellier I (UMI), Montpellier, France

Abstract

Evolutionary analyses have revealed an origin of pandemic HIV-1 group M in the Congo River basin in the first part of the XXth century, but the patterns of historical viral spread in or around its epicentre remain largely unexplored. Here, we combine epidemiologic and molecular sequence data to investigate the spatiotemporal patterns of the CRF02_AG clade. By explicitly integrating prevalence counts and genetic population size estimates we date the epidemic emergence of CRF02_AG at 1973.1 (1972.1, 1975.3 95% CI). To infer their phylogeographic signature at a regional scale, we analyze *pol* and *env* time-stamped sequence data from 8 countries using a Bayesian phylogeographic approach based on a discrete asymmetric model. Our data confirms a spatial origin of this clade in the Democratic Republic of Congo (DRC) and suggests that viral dissemination to Cameroon occurred at an early stage of the evolutionary history of CRF02_AG. We find considerable support for epidemiological linkage between neighbour countries. Compilation of ethnographic data suggests that well-supported viral migration was related with chance exportation events rather than by sustained human migratory flows. Finally, using sequence data from 15 locations in Cameroon, we use relaxed random walk models to explore the spatiotemporal dynamics of CRF02_AG at a finer geographical detail. Phylogeographic dispersal in continuous space reveals that at least two distinct CRF02_AG lineages are circulating in overlapping regions that are evolving at different evolutionary and diffusion rates. Altogether, by combining molecular and epidemiological data, our results provide a time scale for CRF02_AG, place its spatial root within the putative root of group-M diversity and propose a scenario for the spatiotemporal patterns of a successful HIV-1 lineage both at a regional and country-scale.

Corresponding author. Philippe Lemey, Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, Tel: +32-16-332160, Fax: +32-16-332131, philipe.lemey@uz.kuleuven.be.

Competing interest

The authors declare no competing interests.

Introduction

Since the identification of HIV-1 in 1983, there has been a growing understanding of the emergence of this virus. The closest simian immunodeficiency virus (SIV) to HIV-1 group M (Major) were identified in *Pan troglodytes troglodytes* that live in the equatorial forests in the southeast corner of Cameroon [1]. However, the highest sequence diversity of group M has been found in the Democratic Republic of Congo (DRC) [2,3,4]. The explosive urbanization of its capital Kinshasa and the high prevalence of genital ulcer diseases have likely created a unique ground for the generation of a plethora of HIV-1 clades within the first part of the XXth century [5,6], giving rise to what we know today as subtypes and recombinant forms (circulating and unique recombinant forms, CRF/URF). CRFs are mosaic genomes deriving from two different strains that have been found in at least 3 epidemiologically unrelated individuals, whereas URF have only been identified in one patient (<http://www.hiv.lanl.gov/>). Globally, subtype C causes the highest number of infections (50%), followed by subtype A (12%), subtype B (10%), CRF02_AG (5%) and CRF01-AE (4.8%)[7].

CRF02_AG is currently the predominant clade in the Congo River basin (DRC, Angola, Republic of Congo, Central African Republic, Gabon, Equatorial Guinea and Cameroon) [8], but its epidemiology there is quite diverse. While this clade is nearly absent in the DRC [9], Republic of Congo [10] [11] and Central African Republic [12] [13], the clade accounts for almost 20% of the infections in Angola [14] and Gabon [15]. Importantly, while half of the infections are caused by CRF02_AG in Equatorial Guinea [16], the highest proportion of infections caused by this clade occurs in Cameroon [17] [18,19]. HIV surveillance in Cameroon started in 1985, when prevalence was low, around 0.5% [20]. Since then, extensive HIV/AIDS surveillance studies have revealed an overall increase in the prevalence to around 6% [21] with approximately 60% of the infections caused by CRF02_AG. This proportion has been stable over time and is identical in urban and rural settings [17] [18] [22]. In addition, in Cameroon, above 10% of the viruses are unique recombinant forms [17] [18] and all HIV-1 groups have been found to date (group N [23], group O [24] and more recently, group P [25]).

Although the molecular epidemiology of HIV-1 strains has been extensively studied in the Congo River basin, the spatial dynamics of successful viral lineages circulating within this region remains to be elucidated. Given geo-referenced sequence data, Bayesian phylogeographic models provide a powerful framework for investigating simultaneously the spatial and temporal dispersion of viral populations [26,27]. While the current nomenclature proposes that the CRF02_AG clade is a subtype A/ G recombinant, phylogenetic analyses of non-recombinant fragments showed a distinct history in which the putative subtype G is actually a recombinant form with CRF02_AG/J being its putative parental lineages [28]. Although the recombinant origin of CRF02_AG is still a matter of debate [29], this study focuses on understanding the evolutionary history of the CRF02_AG lineage using solely sequence data that shares an identical phylogenetic behaviour within the CRF02_AG clade.

To reconstruct the spatiotemporal dynamics of CRF02_AG at a regional and country-scale, we use Bayesian phylogeographic diffusion models that take into account uncertainty both

at the phylogenetic and viral migration level. First, we inferred the epidemiological dynamics of CRF02_AG integrating molecular sequence data with prevalence counts to reconstruct the temporal origins of this HIV-1 clade. Second, using sequence data with known country of sampling, we show that the CRF02_AG originated in the DRC and highlight the most significant routes of viral dispersal throughout Middle Africa. Finally, given a fine-scale geographic sampling coverage, we introduce relaxed random walks to model the spatial diffusion of HIV viral populations in Cameroon.

Results

A previously published multilocus dataset of 336 *gag-pol-env* CRF02_AG nucleotide sequences from the same blood-donor population in the two most populated urban centers in Cameroon, Yaoundé and Douala [17] was used to infer the time of origin and effective population size estimates (N_e) through time for this clade. We assume that the majority of the infections were acquired in Cameroon. We focus on the exponentially growth period of the HIV-1 epidemic in Cameroon (1990–2000) to estimate the lag that maximized the maximumlikelihood (ML) fit between the best-fitting mean estimates for N_e (see also supplementary materials, Table S1) and the HIV-1 prevalence counts from 1990 to 2007 in this country [21] to date the time of the most recent common ancestor (tMRCA) of the CRF02_AG clade (Figure 1). The ML estimate of the lag between N_e and prevalence data was 5.3 (4.3–7.5 years, 95% CI), providing evidence for an origin of the HIV-1 CRF02_AG clade in or around 1973.1 (1972.1, 1975.3 95% CI).

To analyze the spatial spread of CRF02_AG at a regional scale, we compiled CRF02_AG *pol* (n=83, L=692 nt) and *env* (n=37, L=488 nt) molecular sequences from Angola, Democratic Republic of Congo, Central African Republic, Gabon, Equatorial Guinea and Cameroon. We also included nucleotide data from Chad and previously unpublished sequence data from the island nation Sao Tome and Principe. We use two gene regions that according to the current classification of CRF02_AG derive from subtype G and A respectively (<http://www.hiv.lanl.gov/>). Most probable geographic locations throughout the phylogenetic histories were estimated by applying a discrete asymmetric Bayesian phylogeographic approach, such that dispersal rates from one to another location are allowed to be different from the reversal, thus delivering more realistic scenarios of viral realizations [30]. Additionally, we used a Bayesian stochastic search variable selection (BSSVS) approach to identify most relevant non-zero rates and identify epidemiological linkage between locations [26]. To maximize spatial information embedded in both data sets, we conducted a joint analysis of both *pol* and *env* data sets that allowed independent phylogenies to share the same location-exchange matrix. This analysis placed the majority of the posterior root state probability mass in the DRC with posterior root state probabilities of 0.54 for *pol* and 0.53 for *env* phylogenies, compared to a prior probability of 0.XX (Figure 2). The second most probable root location for both phylogenies was Cameroon (with posterior root state probabilities of 0.14 and 0.23 respectively). These estimates are robust to the sampling scheme used here since the majority of sequences for both loci were from Cameroon (n=23, 12) and Gabon (n=17, 8). Moreover, the DRC is the location attaining most of the posterior mass when performing the analyses separately for each

genomic region albeit with lower support for *env* (posterior root state probability of 0.70 and 0.31 respectively, not shown).

The reconstructed phylogenies suggest that the earliest migratory events were directed from DRC to Cameroon and Gabon (Figure 2). However, we only find support for epidemiological linkage between the DRC and Cameroon [Bayes factor (BF) comparing a model with a nonzero rate to one with a zero rate of 5.4] (Suchard et al. 2001) (Figure 3). In an attempt to put the viral migration from the DRC to Cameroon into a historical demographic context, the number of migrants living in Kinshasa was investigated for the period 1967 and 1977 (Supplementary File 1). We find that migrants from Cameroon were nearly absent in the capital of DRC and vice-versa (existing statistics for the periods of 1967 and 1976 suggest that these numbers varied between 100–200 people, see Supplementary File 1). This suggests the viral dissemination from the DRC to Cameroon may have been due to chance exportation of the virus rather than sustained human migratory flows.

Moreover, sequences from the island nation São Tomé and Príncipe, a former Portuguese colony, were found interspersed with sequences from Angola (also a former Portuguese colony), Gabon and Equatorial Guinea (Figure 2). Although this suggests at least three supported independent sources of this clade in São Tomé and Príncipe, the only supported links were from Equatorial Guinea (BF=12.8) and Angola (BF=5.3) (Figure 3). The majority of the sequences from Gabon are descendent from the neighbouring countries Cameroon and Equatorial Guinea, which is confirmed by a high Bayes factor support (BF=16.91 and BF=9.7). Although Cameroon is also bordered by Chad and Nigeria at the north and Central African Republic in the east (and also by the Republic of Congo in the south; yet the presence of CRF02_AG has not been confirmed there [10,11]), we only detected support for viral migration from Cameroon to Gabon and Equatorial Guinea (Figure 3). By 1976, the majority of the foreign population in Cameroon was from Nigeria (n=56.046 from a total of 2.005.223 people), followed by Chad (n=12.176) and Central African Republic (n=7.946) (Supplementary File 1). Human mobility data would suggest viral intermixing between Cameroon and Nigeria. Instead, when performing a similar analysis including sequences from Nigeria, we obtained supported links from Cameroon to Gabon and subsequently to Nigeria (not shown). Altogether, these observations suggest that human mobility alone cannot address the complexity of viral diffusion. Likely, factors such as population growth and accessibility between locations play also an important role in viral spread [31].

Discrete diffusion models offer insights into the origins and epidemiological links within the set of locations from which viruses were sampled. However, given a more fine-scale geographical coverage (n=15 locations sampled from seven out of ten regions in Cameroon), we are able to estimate the unobserved locations of sequence ancestors in continuous space using recently developed relaxed random walk (RWW) models [27]. To first examine whether the Cameroonian sequences show evidence for a single successful viral population, we performed ML analyses including all available Cameroonian sequences with known sampling locations overlapping with the regional *pol* data set (supplementary materials, Figure S2). The majority of the sequences from Cameroon fell within two supported clusters (n=48 and n=28, named as clusters 1 and 2, supplementary materials, Figure S2). While modelling the diffusion process of the Cameroonian epidemics, Brownian diffusion (BD)

models, that assume a constant variance random walk along each branch in the phylogeny, were compared with RRW models, in which dispersion rates are allowed to vary according to distinct prior distributions. In all cases, the RRW models provided a better fit to the data, with the Cauchy distribution attaining the best fit (Table S4). This is consistent with coefficients of variation for both clusters that indicate mean dispersal rates varying among the branches within about 150% of the mean rate (Table 1).

To compare the dynamics of the two CRF02_AG lineages circulating in Cameroon, we consider their evolutionary and geographic diffusion rates (Table 1). Interestingly, the evolutionary rate for cluster 1 are estimated at 1.41 (1.14, 1.71, 95% highest posterior density (HPD) interval) substitutions per site per year (s/s/y) while cluster 2 yields 0.94 (0.75, 1.16, 95% BCI) s/s/y. The values for the diffusion rates were 7.45 (6.00, 9.02, 95% HPD) km per year (km/y) and 5.61 (4.28, 7.13, 95% HPD) km/y respectively. Finally, the dispersal patterns for both lineages circulating in Cameroon were summarized as 95% HPD regions for the roots of clusters 1 and 2 that represent time-slices of the posterior phylogeny distribution (Figure 4). The root location for cluster 1 was inferred in the Centre region, not far from the root location for cluster 2. Importantly, the continuous diffusion inference shows that the CRF02_AG epidemics ignited in the Centre and spread rapidly to the Littoral and West regions in Cameroon, thereafter spreading to the Northeast and Southeast regions; only after that did this clade diffuse to more remote regions of the South and East. These data show that the diffusion of two distinct CRF02_AG lineages ignited in the most populated regions thereafter spreading to more remote regions in Cameroon.

Discussion

We have investigated the spatio-temporal dynamics of HIV-1 CRF02_AG in the Congo River basin with particular focus on Cameroon, to provide a better understanding on the origins and spread of this clade within the roots of group-M diversity. The 336 CRF02_AG *gag-pol-env* sequences from same blood donors in the two most populated cities in Cameroon provided a suitable data set to estimate the dynamics in effective population size (N_e) through time for CRF02_AG in the general population. We subsequently fitted these estimates to the UNAIDS epidemiologic surveillance data to estimate informatively the divergence time of CRF02_AG at 1973.1 (1972.1, 1975.3 95% CI).

By applying Bayesian phylogeographic inference using discrete non-reversible models to *pol* and *env* geo-referenced sequences, we investigated the spatial patterns of this clade at a regional level. Our findings suggest that the CRF02_AG clade originated in the DRC, although infections caused there by this clade are rare [9] [32]. We identified the most significant epidemiological links of CRF02_AG within the Congo River basin; these suggest that the high predominance of this clade in Cameroon [17] [18] might be related with at least two chance exportations of the virus from the DRC to Cameroon in the very early history of this clade. Finally, we explore the CRF02_AG spread in greater detail in Cameroon and show that there are two distinct epidemic lineages of CRF02_AG that seem to have ignited in the most urbanized (Centre) region of Cameroon. These lineages have been spreading at distinct evolutionary and diffusion rates, albeit in somewhat overlapping geographic regions.

Recent studies have shown that the effective population sizes estimated from phylogenetic inference should be interpreted in light of the number of new transmissions rather than the number of infected individuals (or prevalence) [33,34]. However, the authors point out that during the exponential period of an epidemic these entities are linearly correlated. Therefore, we restricted the estimation of the ML fit to the exponential growth period of the HIV-1 epidemic in Cameroon (1990–2000) to achieve higher temporal resolution in our tMRCA estimates. The lag between the N_e estimates and the prevalence counts was calculated to be 5.3 years. That N_e estimations precede serological counts has also been noted previously for another fast evolving virus [35]. It is possible that the lag obtained by our estimates reflects the difference in years between the number of new infections and the total number of infected individuals [33,34]. Nevertheless, ML phylogenetic analyses with the published 336 *gag*, *pol* and *env* sequences [17] and the reference set described herein indicated that 5 (1.48%), 4 (1.19%) and 2 (0.59%) sequences from different patients clustered paraphyletically with respect to the CRF02_AG cluster (not shown). Therefore, the inclusion of such sequence data cannot be ruled-out as an explanation for the lag observed between the estimated N_e and prevalence data. Despite this, we obtained the patterns of N_e over time using a multigene analysis that benefits from higher phylogenetic resolution to estimate the phylodynamic patterns derived from blood donors, the population that most resembles the general population as envisaged by the surveillance counts. In addition, the results were qualitatively similar when analyzing each locus individually with *gag* yielding the closest estimates to the ones obtained using the multilocus data set (not shown). By making use of prevalence counts to infer the tMRCA of the CRF02_AG lineage, the uncertainty on this estimate achieved through our analysis narrows by over 29% compared to (and are included in) the credible intervals delivered by previous estimates [36]. Although CRF02_AG was only identified in 1994 [37], it has been estimated that by this time over 500 thousand people living in the Congo River basin were infected; after, the proportion of infections stabilized [8]. More generally, our results are in line with the timing of this levelling-off, giving support for the divergence time for CRF02_AG estimated here.

Bayesian phylogeographic estimates of the dispersal patterns were applied to two distinct loci. The inference of a spatial root in the DRC was robust to both an analysis sharing a non-reversible diffusion model across unlinked loci or a single gene analysis and also to the sampling scheme used here, since the majority of the sequences were from Cameroon and Gabon. A combined analysis has the potential to more efficiently use the genetic and geographic information in the two loci. In addition, according to the current classification of CRF02_AG the *pol* and *env* data sets used here derive from subtype G and A respectively (<http://www.hiv.lanl.gov/>). Thus, to ensure that parental sequences were not being used we performed a conservative data selection to restrict the phylogeographic inference to sequence data that shared an identical phylogenetic behaviour within the CRF02_AG clade. Nevertheless, and given that the geo-reference sequence data available for analysis is limited, the phylogeographic inference presented here would benefit from a more comprehensive sampling scheme.

For obvious reasons, viral migration has frequently been explained based on human mobility [38]. For example, a study conducted in Yaoundé showed that the risk of HIV infection in

men increased up to five times with more prolonged time intervals away from the town [39]. Our findings suggest that viral migration from the DRC to Cameroon occurred in an early stage of the epidemic. Although the human migration patterns within the Congo River basin during the 1970s are difficult to trace, within 1967–77 the number of people living in Kinshasa grew from 901.520 to 2.440.000 and the great majority of migrants living in this city were from Angola (between 14.98 to 10.73% of the total population; see Supplementary File 1). Demographic surveys show that Cameroonian migrants were nearly absent in the DRC and that the reverse was also true (Supplementary File 1), suggesting that this linkage was due to chance exportation of the virus. Within the inferred intervals estimated for viral flow from the DRC to Cameroon, transnational movement could be accomplished either by waterways (mainly through Congo and Sangha rivers), roads (from Ouessou to Bangui in Central African Republic and from there to Bertoua and Yaoundé) or also by air. In addition, it is likely that temporary labour recruitment from Cameroon might have contributed to shape HIV-1 epidemiology in the mineral-rich Gabon and Equatorial Guinea. For instances, the proportion of CRF02_AG infections in miners working in south-eastern Gabon [40] is similar to the observed in the general population of Cameroon and Equatorial Guinea.

The low dispersal rates within Cameroon are inherent to applying these models to a sample that encompasses a limited geographic range for clades that span over 30 years. If viruses from these clades that were potentially exported to other countries had been included, this would have yielded higher dispersal rates. In fact, despite the large overlap, the somewhat wider sampling range within Cameroon for cluster 1 might explain its higher dispersal rates compared to cluster 2. In general, this demonstrates that such estimates are strongly associated with the sampling range and comparisons of dispersal rates across different sampling ranges are likely to be misleading. In addition, the applicability of continuous diffusion models to human viruses may be limited to confined geographic areas because even relaxed random walk models might be poor approximations for viral diffusion across large geographic ranges. The concomitant differences in evolutionary rates between the clusters remain more difficult to explain. In the absence of information concerning risk groups for the data used in this study we can only speculate that different transmission dynamics in distinct risk groups might be responsible for the rate differences [41]. Alternatively, it may be that the two clusters are spreading in different ethnic groups that have inherently different genetic backgrounds and also distinct mobility patterns. Cameroon is home to an extensive diversity in languages and ethnicity. Analysis of sequence data referenced with patient ethnic information intersected with the complex mobility patterns may provide some insight into viral spread in Cameroon.

Our study sheds light in the emergence and dynamics of an important HIV-1 clade in the Congo River basin, the source location for the HIV-1 group M diversity. The evolutionary history of human viruses can only be fully understood when the intrinsic spatial and temporal components are taken into account. Importantly, understanding the origins and dispersal patterns of successful HIV-1 clades both at a regional and intra-country level not only unites the fragmentary pieces delivered by serological counts but should ultimately become central to improve the characterization and control of HIV spread.

Materials and Methods

Nucleotide datasets compilation

A multilocus alignment of 336 *gag* (HXB2: 1255 to 1682), *pol* (HXB2: 4228–5093) and *env* (HXB2: 7890–8266) comprising CRF02_AG published gene sequences sampled between 1996 and 2004 from blood donors from Yaoundé and Douala [17] was used to investigate in detail the demographic dynamics of the CRF02_AG lineage. *Pol* (HXB2: 2253 to 2944) and *env* (HXB2: 7038 to 7463) HIV-1 CRF02_AG gene sequences sampled in Middle Africa as defined by United Nations (Chad, Cameroon, Central African Republic, Equatorial Guinea, Gabon, Democratic Republic of Congo, Angola and São Tomé and Príncipe) were selected from the 2010 LANLdb (<http://www.hiv.lanl.gov/>) to investigate viral migration patterns at a regional level. To investigate viral migration within Cameroonian locations, we compiled *pol* HIV-1 CRF02_AG gene sequences (HXB2: 2253 to 3275) with known date of sampling and detailed geographical location (city or village) from the 2010 LANLdb database (<http://www.hiv.lanl.gov/>) (for details see Table S1). Sequence data belonged to 7 out of the 10 regions of Cameroon. Although there is no available data from the Adamawa, North and Far North, these regions have the lowest HIV-1 prevalence in the country [21].

Subtype assignment and sequence alignment

A recent study has demonstrated that 4.9% of the original subtype assignments in the LANL database need a revision [42]. Therefore, and because the effect of recombination in phylogeographic inference might affect the results [43], we conducted a stringent procedure to ensure that the data used for the phylogeographic analyses were closely related to the CRF02_AG clade. First, we excluded all sequences that did not cluster monophyletically within CRF02_AG cluster using a reference set with all available full genome sequences from the closest phylogenetic clades to CRF02_AG [28] sampled worldwide from subtypes A (n=29), sub-subtypes A1 (n=87) and A2 (n=3), subtype G (n=26), CRF02_AG (n=27) and subtype J (n=4). Sequence alignments including the reference set and i) the Middle African, or the ii) Cameroonian dataset were created using the multiple alignment with the fast Fourier transform algorithm (MAFFT) [44] and manually edited with Se-AL (<http://tree.bio.ed.ac.uk>). The overlapping regions were then used to perform phylogenetic analyses under a general time reversible model with discrete gamma and invariant among-site rate variation (GTR+4Γ+I) using a ML criterion implemented in PhyML with an approximate likelihood ratio test statistics to search for the best of NNI and SPR tree [46]. Only sequences that grouped monophyletically within the CRF02_AG clade with significant statistical support (bootstrap values above 75%) were used for subsequent analyses. In addition, the subtype assignment was confirmed using the NCBI HIV subtyping tool (<http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi>) and the Rega subtyping tool v.2 [47] [48]. After the subtype assignment steps, Middle African datasets comprised 77 *pol* and 59 *env* sequences with 691 and 488 nucleotides, spanning the genomic regions 2252 to 2943 and 7037 to 7525 (HXB2). These were sampled in Cameroon (23, 12 for *pol* and *env* respectively), Chad (3, 9), Gabon (17, 9), Democratic Republic of Congo (5, 10), Central African Republic (3 *pol*), Guinea- Equatorial (14 *pol*) and Angola (6 *pol*). The Cameroonian dataset consisted in 80 *pol* sequences (HXB2: 2252 to 3319) from a total of 15 locations. In particular, the sequences originated from the Centre (22), Northwest (7), East (18), West (7),

Littoral (10), Southwest (6), and South (9) regions. All data sets for phylogeographic analyses tested negative for inter-subtype recombination using the Phi-test implemented in SplitsTree4 version 4.10 [49]. All sequence alignments are available from the authors upon request.

Timing the introduction of CRF02_AG in Cameroon

To estimate the changes in the effective population size N_e through time for CRF02_AG in Cameroon, we used a Bayesian coalescent approach as implemented in BEAST [50] with BEAGLE to enhance computational speed [Suchard and Rambaut, 2009]. For this purpose, we analyzed 336 *gag*, *pol* and *env* HIV-1 CRF02_AG gene sequences [17] using either a multilocus approach or separate analyses under a GTR+4Γ+I substitution process. We used the uncorrelated lognormal molecular clock model that allows rates to vary among the branches of the inferred phylogenies [51]. For the multilocus analysis we tested constant, exponential and constant-logistic demographic tree prior models. Nucleotide substitution models, molecular clock models and the demographic model were shared among the different partitions, whereas each partition was allowed to have different phylogenetic histories and different coefficients of variations for the molecular clock model [52]. Demographic model fit was assessed by comparison of marginal likelihoods [Suchard et al., 2001 and 53] (supplementary materials, Table S1). For the single loci analyses we used the semi-parametric Skyride demographic tree prior with time aware smoothing [54]. Markov chain Monte Carlo simulations were run for 10 to 25×10^7 chain steps, sub-sampling parameters every 5000 steps. Convergence of the chains was inspected using Tracer.v.1.5. To estimate accurately the introduction of CRF02_AG in Cameroon, we assess the fit of the posterior mean estimates of N_e from BEAST to the HIV-1 seroprevalence data in a similar fashion to previously described for dengue virus [35]. Particularly, the lag between the prevalence counts and the estimates using virus sequence data was selected in order to maximize the likelihood of a linear regression related the counts to N_e translated by a unknown amount of time in the R package. The optimally lagged estimation of the divergence time of CRF02_AG was used as a calibration prior in subsequent analyses.

Bayesian phylogeographic models

Bayesian phylogeographic analyses [Lemey et al., 2009, Lemey et al., 2010] were carried in under an MCMC framework as implemented in BEAST [50] with BEAGLE [Suchard and Rambaut, 2008]. A Bayesian skyride tree prior was used as a coalescent demographic model with time-aware smoothing [54]. The identification of significant migration pathways was performed using discrete non-reversible diffusion models and a BSSVS approach [26] [30]. For the discrete diffusion models, geographic locations were recorded at the tips of *pol* and *env* phylogenies respectively. The unobserved locations of the ancestral nodes until the root were inferred for a posterior distribution of trees. Non-reversible models provided a better-fit than reversible diffusion models as confirmed by a higher BF support (Suchard et al, 2001). To reconstruct the evolutionary history of CRF02_AG in Cameroon explicitly in continuous space, we propose models of continuous diffusion [27]. Exact latitude and longitude for each viral isolate were recorded at the tips of an unknown phylogeny, and the unobserved two-dimensional locations along each node of the posterior distribution of the phylogeny were estimated. Gamma, Cauchy, Lognormal and homogeneous prior

distributions to rescale the variance of the random walk were tested and posterior estimates for the parameters were compared (supplementary materials, Table S5).

Significant non-zero rates obtained by the BSSVS approach were spatially projected and converted into a keyhole markup language (KML) file, which can be viewed with Google Earth (<http://earth.google.com>), are available upon request. The animated continuous phylogeographic descriptions of the epidemics in Cameroon are available at <http://www.phylogeography.org/>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

NRF is supported by Fundação para a Ciência e Tecnologia (grant no. SFRH/BD/64530/2009). MAS is supported by NIH R01 GM86887. The research leading to these results has received funding from the European Commission (EC grant CHAIN 7FP, 223131) and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC Grant agreement n° 260864.

Reference List

1. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*. 2006; 313:523–526. [PubMed: 16728595]
2. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC. Human immunodeficiency virus. Phylogeny and the origin of HIV-1. *Nature*. 2001; 410:1047–1048. [PubMed: 11323659]
3. Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, et al. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J Virol*. 2000; 74:10498–10507. [PubMed: 11044094]
4. Sharp PM, Hahn BH. AIDS: prehistory of HIV-1. *Nature*. 2008; 455:605–606. [PubMed: 18833267]
5. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*. 2008; 455:661–664. [PubMed: 18833279]
6. Sousa JD, Muller V, Lemey P, Vandamme AM. High GUD incidence in the early 20 century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. *PLoS One*. 5:e9936. [PubMed: 20376191]
7. Taylor BS, Hammer SM. The challenge of HIV-1 subtype diversity. *N Engl J Med*. 2008; 359:1965–1966. [PubMed: 18971501]
8. Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infect Dis*. 11:45–56. [PubMed: 21126914]
9. Kita K, Ndembu N, Ekwilanga M, Ido E, Kazadi R, et al. Genetic diversity of HIV type 1 in Likasi, southeast of the Democratic Republic of Congo. *AIDS Res Hum Retroviruses*. 2004; 20:1352–1357. [PubMed: 15650428]
10. Bikandou B, Takehisa J, Mboudjeka I, Ido E, Kuwata T, et al. Genetic subtypes of HIV type 1 in Republic of Congo. *AIDS Res Hum Retroviruses*. 2000; 16:613–619. [PubMed: 10791871]
11. Niama FR, Toure-Kane C, Vidal N, Obengui P, Bikandou B, et al. HIV-1 subtypes and recombinants in the Republic of Congo. *Infect Genet Evol*. 2006; 6:337–343. [PubMed: 16473564]
12. Marechal V, Jauvin V, Selekon B, Leal J, Pelembe P, et al. Increasing HIV type 1 polymorphic diversity but no resistance to antiretroviral drugs in untreated patients from Central African Republic: a 2005 study. *AIDS Res Hum Retroviruses*. 2006; 22:1036–1044. [PubMed: 17067275]

13. Muller-Trutwin MC, Chaix ML, Letourneur F, Begaud E, Beaumont D, et al. Increase of HIV-1 subtype A in Central African Republic. *J Acquir Immune Defic Syndr*. 1999; 21:164–171. [PubMed: 10360809]
14. Bartolo I, Rocha C, Bartolomeu J, Gama A, Marcelino R, et al. Highly divergent subtypes and new recombinant forms prevail in the HIV/AIDS epidemic in Angola: new insights into the origins of the AIDS pandemic. *Infect Genet Evol*. 2009; 9:672–682. [PubMed: 18562253]
15. Pandrea I, Robertson DL, Onanga R, Gao F, Makuwa M, et al. Analysis of partial pol and env sequences indicates a high prevalence of HIV type 1 recombinant strains circulating in Gabon. *AIDS Res Hum Retroviruses*. 2002; 18:1103–1116. [PubMed: 12396449]
16. Djoko CF, Wolfe ND, Vidal N, Tamoufe U, Montavon C, et al. HIV type 1 pol gene diversity and genotypic antiretroviral drug resistance mutations in Malabo, Equatorial Guinea. *AIDS Res Hum Retroviruses*. 26:1027–1031. [PubMed: 20718620]
17. Brennan CA, Bodelle P, Coffey R, Devare SG, Golden A, et al. The prevalence of diverse HIV-1 strains was stable in Cameroonian blood donors from 1996 to 2004. *J Acquir Immune Defic Syndr*. 2008; 49:432–439. [PubMed: 18931623]
18. Carr JK, Wolfe ND, Torimiro JN, Tamoufe U, Mpoudi-Ngole E, et al. HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: evolutionary relics? *Retrovirology*. 2010; 7:39. [PubMed: 20426823]
19. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, et al. Tracing the origin and history of the HIV-2 epidemic. *Proc Natl Acad Sci U S A*. 2003; 100:6588–6592. [PubMed: 12743376]
20. Rémy G. Image géographique de l'infection à VIH-1 en Afrique Centrale: des discontinuités remarquables. *Ann Soc Belg Méd Trop*. 1993; 73:127–142. [PubMed: 8368889]
21. UNAIDS/WHO UNAIDS/WHO. Epidemiological Fact Sheets on HIV and AIDS. 2008 Update.
22. Konings FA, Zhong P, Agwara M, Agyingi L, Zekeng L, et al. Protease mutations in HIV-1 non-B strains infecting drug-naïve villagers in Cameroon. *AIDS Res Hum Retroviruses*. 2004; 20:105–109. [PubMed: 15008125]
23. Bodelle P, Vallari A, Coffey R, McArthur CP, Beyeme M, et al. Identification and genomic sequence of an HIV type 1 group N isolate from Cameroon. *AIDS Res Hum Retroviruses*. 2004; 20:902–908. [PubMed: 15320995]
24. Peeters M, Gueye A, Mboup S, Bibollet-Ruche F, Ekaza E, et al. Geographical distribution of HIV-1 group O viruses in Africa. *AIDS*. 1997; 11:493–498. [PubMed: 9084797]
25. Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, et al. Confirmation of Putative HIV-1 Group P in Cameroon. *J Virol*. 2010
26. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol*. 2009; 5:e1000520. [PubMed: 19779555]
27. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol*. 27:1877–1885. [PubMed: 20203288]
28. Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, et al. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. *J Virol*. 2007; 81:8543–8551. [PubMed: 17553886]
29. Bulla I, Schultz AK, Schreiber F, Zhang M, Leitner T, et al. HIV classification using the coalescent theory. *Bioinformatics*. 26:1409–1415. [PubMed: 20400454]
30. Edwards CJSB, Lemey P, Barnes I, Fulton TL, Barnett R, O'Connell T, Coxon P, Monaghan N, Valdiosera CE, Baryshnikov GF, Thomas MG, Suchard MA, Bradley DG. Out of Ireland: a Hibernian origin for modern polar bear matrilineal preparation.
31. Gray RR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, et al. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS*. 2009; 23:F9–F17. [PubMed: 19644346]
32. Vidal N, Mulanga C, Bazepeo SE, Mwamba JK, Tshimpaka JW, et al. Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002. *J Acquir Immune Defic Syndr*. 2005; 40:456–462. [PubMed: 16280702]
33. Frost SD, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci*. 365:1879–1890. [PubMed: 20478883]
34. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. *Genetics*. 2009; 183:1421–1430. [PubMed: 19797047]

35. Bennett SN, Drummond AJ, Kapan DD, Suchard MA, Munoz-Jordan JL, et al. Epidemic dynamics revealed in dengue evolution. *Mol Biol Evol.* 27:811–818. [PubMed: 19965886]
36. Abecasis AB, Vandamme AM, Lemey P. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J Virol.* 2009; 83:12917–12924. [PubMed: 19793809]
37. Howard TM, Rasheed S. Genomic structure and nucleotide sequence analysis of a new HIV type 1 subtype A strain from Nigeria. *AIDS Res Hum Retroviruses.* 1996; 12:1413–1425. [PubMed: 8893049]
38. Quinn TC. Population migration and the spread of types 1 and 2 human immunodeficiency viruses. *Proc Natl Acad Sci U S A.* 1994; 91:2407–2414. [PubMed: 8146131]
39. Lydie N, Robinson NJ, Ferry B, Akam E, De Loenzien M, et al. Mobility, sexual behavior, and HIV infection in an urban population in Cameroon. *J Acquir Immune Defic Syndr.* 2004; 35:67–74. [PubMed: 14707795]
40. Caron M, Makuwa M, Souquiere S, Descamps D, Brun-Vezinet F, et al. Human immunodeficiency virus type 1 seroprevalence and antiretroviral drug resistance-associated mutations in miners in Gabon, central Africa. *AIDS Res Hum Retroviruses.* 2008; 24:1225–1228. [PubMed: 18788914]
41. Berry IM, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, Bruno W, Leitner T. Unequal Evolutionary Rates in the Human Immunodeficiency Virus Type 1 (HIV-1) Pandemic: the Evolutionary Rate of HIV-1 Slows Down When the Epidemic Rate Increases. *Journal of Virology.* 2007; 81:10625–10635. [PubMed: 17634235]
42. Zhang M, Foley B, Schultz AK, Macke JP, Bulla I, et al. The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology.* 7:25. [PubMed: 20331894]
43. Avise, JC. *Phylogeography: The History and Formation of Species.* Cambridge, MA: Harvard University Press; 2000.
44. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005; 33:511–518. [PubMed: 15661851]
46. Gouy M, Guindon S, Gascue O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27:221–224. [PubMed: 19854763]
47. Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res.* 2009; 37:W634–W642. [PubMed: 19483099]
48. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics.* 2005; 21:3797–3800. [PubMed: 16076886]
49. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 2006; 23:254–267. [PubMed: 16221896]
50. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 2007; 7:214. [PubMed: 17996036]
51. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 2006; 4:e88. [PubMed: 16683862]
52. Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, et al. The molecular population genetics of HIV-1 group O. *Genetics.* 2004; 167:1059–1068. [PubMed: 15280223]
53. Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics.* 2006; 22:2047–2048. [PubMed: 16679334]
54. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* 2008; 25:1459–1471. [PubMed: 18408232]
55. Suchard MA, Rambaut A. Many-core algorithms for statistical phylogenetics. *Bioinformatics.* 2009; 25:1370–1376. [PubMed: 19369496]

Figure 1. Fitting viral effective population size estimates to HIV prevalence data

Effective population size estimates ($N_e \tau$, where τ represents the generation time), that were estimated from a 336 multilocus data set from blood-donor patients in Cameroon [17], were fit to the HIV prevalence in the same country [21]. The optimally lagged tMRCA is bounded by its ML confidence interval. The inferred ML estimate of the lag between N_e and prevalence counts is indicated by an arrow.

Figure 2. Phylogeographic origins and spread of CRF02_AG in Middle Africa

Bayesian maximum clade credibility phylogeographic trees for *pol* and *env* datasets. Each branch is coloured according to the most probable location and the legend for the colours is shown on the right. The ancestral root state probability for *pol* and *env* data sets is shown on the right of the respective colour codes in grey. Country-codes: CD: Democratic Republic of Congo, CM: Cameroon, GA: Gabon, GQ: Equatorial Guinea, AO: Angola, TD: Chad, CF: Central African Republic, ST: São Tomé and Príncipe.

Figure 3. Most significant epidemiological links of CRF02_AG dispersal in Middle Africa
 Most significant links plotted in Middle Africa. Sequence data from both data sets were used in a combined analysis. The putative root of CRF02_AG emergence, Kinshasa, is highlighted with a dashed circle. Only epidemiological links supported by Bayes factor rates above 5 are indicated. Legend for the strength of the Bayes factor rates is shown on the bottom-left.

Figure 4. Spatiotemporal dynamics of the CRF02_AG epidemics in Cameroon

The dispersal patterns of two distinct lineages of CRF02_AG are indicated for 1975, 1990 and 2005. For each cluster, the 95% HPD of the root location is highlighted by a dashed line. Red-blue lines represent older-recent branches of the MCC trees projected in the surface. Transparent polygons represent the 80% HPD uncertainty on the location of the viral lineage. White-red gradients indicate older-recent age of dispersal. The figure is based on images made available by Google Earth (<http://earth.google.com>). A dynamic

visualization of the spatiotemporal process can be examined at <http://www.phylogeography.org/>.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Comparison of the evolutionary parameters between clusters 1 and 2 of CRF02_AG virus in Cameroon. We report posterior mean and 95% HPD intervals except where indicated.

	Cluster 1	Cluster 2
tMRCA	34.1 (30.2, 38.1)	33.8 (30.0, 37.7)
Substitution rate (s/s/y)	1.41 (1.14, 1.71)	0.94 (0.75, 1.16)
Dispersion rate (km/y)	7.51 (5.97, 9.06)	5.41 (3.84, 7.00)
Coefficient of variation	1.58 (1.26, 1.92)	1.38 (1.06, 1.76)
Root area 80% HPD (degrees ²)	14.11	10.97
Total 80% HPD area (degrees ²)	115.00	12.01