THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Reconstructing the HIV-1 CRF02_AG and CRF06_cpx Epidemics in Burkina Faso and West Africa using early samples

Citation for published version:
Yebra, G, Kalish, ML & Leigh Brown, AJ 2016, 'Reconstructing the HIV-1 CRF02_AG and CRF06_cpx Epidemics in Burkina Faso and West Africa using early samples' Infection, Genetics and Evolution, vol. 46, pp. 209-218. DOI: 10.1016/j.meegid.2016.03.038

Digital Object Identifier (DOI):
10.1016/j.meegid.2016.03.038

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Infection, Genetics and Evolution

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 05. Apr. 2019

RECONSTRUCTING THE HIV-1 CRF02_AG AND CRF06_CPX EPIDEMICS IN BURKINA FASO AND WEST AFRICA USING EARLY SAMPLES

Gonzalo YEBRA, [a] Marcia L. KALISH [b] and Andrew J. LEIGH BROWN [a]

[a] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

[b] Institute for Global Health, Vanderbilt University, Nashville TN, USA

Corresponding author:

Dr. G. Yebra
Institute for Evolutionary Biology
Ashworth Building, University of Edinburgh
West Mains Road
Edinburgh
EH9 3JT
UK
E-mail: Gonzalo.Yebra@ed.ac.uk

ABSTRACT

Background. HIV-1 circulating recombinant forms (CRFs) represent viral recombinant lineages that play a significant role in the global epidemic. Two of them dominate the epidemic in Burkina Faso: CRF06_cpx (first described in this country) and CRF02_AG. We reconstructed the phylodynamics of both recombinant viruses in Burkina Faso and throughout West Africa.

Methods. We analysed CRF06_cpx and CRF02_AG sequences (protease/gp41) from early samples collected in Burkina Faso in 1986 together with other GenBank sequences (1984-2013) in 4 datasets: African CRF06_cpx (210/60); down-sampled CRF06_cpx (146/45); Burkina Faso CRF02_AG (130/39) and West/Central African CRF02_AG (691/298). For each dataset, we analysed both protease and gp41 jointly using the BEAST multilocus analysis and conducted phylogeographic analysis to reconstruct the early migration routes between countries.

Results. The time to the most recent common ancestor (tMRCA) of CRF06_cpx was 1979 (1973-1983) for protease and 1981 (1978-1983) for gp41. The gp41 analysis inferred the origin of CRF06_cpx (or at least its parental subtype G lineage) in the Democratic Republic of Congo but migrated to Burkina Faso soon after (1982). Both genes showed that CRF06_cpx radiated to the rest of West Africa predominantly after around 1990. These results were robust to the oversampling of Burkina Faso sequences as they were confirmed in the down-sampled dataset. The tMRCA of the Burkina Faso CRF02_AG lineage was 1979 (1977-1983) for protease and 1980 (1978-1981) for gp41. However, we reconstructed its presence in West Africa much earlier (mid-1960s), with an initial origin in Cameroon and/or Nigeria, and its phylogeographic analysis revealed much interconnection within the region with a lack of country-specific phylogenetic patterns, which prevents tracking its exact migration routes.

Conclusions. Burkina Faso presents a relatively young HIV epidemic, with the diversification of the current in-country CRF02_AG and CRF06_cpx lineages taking place around 1980. This country represents the main source of CRF06_cpx in West Africa. The CRF02_AG epidemic started at least a decade earlier and showed much interchange between West African countries (especially involving coastal countries) suggesting a great population mobility and an extensive viral spread in the region.

KEYWORDS

HIV; phylodynamics; phylogeography; recombinant; Burkina Faso; West Africa

1. INTRODUCTION

Burkina Faso, a landlocked country in West Africa and bordered by Côte d'Ivoire, Mali, Niger, Togo, Benin and Ghana, had an HIV prevalence among the adult population of 0.9% in 2013, with a total of 110,000 infected people (UNAIDS, 2015). Both HIV-1 and HIV-2 are present in the region and two HIV-1 circulating recombinant forms (CRFs) dominate the epidemic in Burkina Faso: CRF02_AG and CRF06_cpx (Tebit et al., 2006; Tebit et al., 2009). CRFs represent recombinant HIV-1 genomes that have infected three or more persons who are not epidemiologically related, so they can be assumed to have an epidemiologically relevant contribution to the HIV-1 M group epidemic (LANL, 2016).

CRF02_AG is the most prevalent recombinant worldwide, being especially prevalent in Central-West Africa (Hemelaar et al., 2011; Lihana et al., 2012). It was first described from a sample collected in Nigeria in 1994 (Carr et al., 1998), but its geographical origin was inferred to be in the Democratic Republic of Congo (DRC) (Faria et al., 2012) using *pol* sequences from different Central African countries. These authors applied a BEAST multilocus analysis to a second CRF02_AG dataset which included *gag*, *pol* and *env* sequences from Cameroon (1996-2004), and estimated (after correcting with data about HIV prevalence), its date of origin in 1973 (1972-1975). This date is close to the estimates by Abecasis and colleagues (Abecasis et al., 2009) using independent CRF02_AG *pol* (1976 [1970-1981]) and *env* (1975 [1969-1980]) sequences sampled between 1990 and 2005. However, no studies so far have helped in understanding how CRF02_AG spread to West Africa, where this strain currently dominates the HIV epidemic.

CRF06_cpx was first described in 1996 in a patient from Burkina Faso residing in Australia (Oelrichs et al., 1998); it has since been detected at a lower prevalence in other West African countries (Mali, Côte d'Ivoire, Ghana, Benin, Senegal, etc.) as well as in Nigeria and Niger. It is this African region where its current distribution is mainly restricted. However, this strain has been occasionally detected in developed countries worldwide (LANL, 2015) and, importantly, in an outbreak among injecting drug users (IDUs) first described in Estonia around 2000 (Zetterberg et al., 2004). Later, this variant was identified in other European countries like Sweden, Norway, Finland, Russia and Belarus.

Delatorre and Bello (Delatorre and Bello, 2013) estimated the origin of CRF06_cpx to be in Burkina Faso in 1979 (1970-1985) using phylodynamic and phylogeographic analyses of *pol* sequences sampled in West (n=174) and Central (n=6) Africa between 1995 and 2010. They showed a great level of phylogenetic intermixing from different countries, with Burkina Faso as the epicentre from which CRF06_cpx spread to neighbouring countries. However, phylogeographic analyses are very susceptible to uneven sample size and in that study the overwhelming majority of CRF06_cpx sequences were sampled in Burkina Faso.

A recent paper described early sequences of CRF02_AG and CRF06_cpx derived from serum samples taken in Ouagadougou, Burkina Faso in 1986 (Fonjungo et al., 2014), the same year as the first case of AIDS was officially reported in the country. Analysis of these samples showed that these two recombinants accounted for more than 90% of HIV-1 samples from that year, suggesting that they dominated the HIV epidemic in Burkina Faso from its beginning. These sequences represent the earliest ones described so far in West Africa, although one study (Kalish et al., 2004) reported several sequences sampled in 1984 in the current DRC which, though initially described as pure subtypes, could be part of the CRF02_AG and CRF06_cpx lineages. The inclusion of early sequences in phylogenetic analyses allows a better reconstruction of the history of HIV (Worobey et al., 2008; Yebra et al., 2015).

Here, we use phylodynamic and phylogeographic analyses to shed light on previously unanswered questions regarding the history of the distinctive Burkina Faso HIV epidemic, such as the origin and spread of CRF06_cpx and the dissemination of CRF02_AG from Central to West Africa, where it now dominates the HIV epidemic. To do that, we have included sequences sampled in the period 1984-2013 and collected from public databases. We apply a multilocus approach which allows us to jointly analyse two different genes (partial *pol* and *env*), by incorporating their genetic information but allowing them to have independent phylogenies. Analyses of multiple unlinked loci improve the precision and power to reconstruct the past population dynamics of a population by providing additional information in the same time-frame (Faria et al., 2012; Gill et al., 2013). Finally, we estimate the geographic origin of CRF06_cpx, adjusting for the oversampling of Burkina Faso sequences in the phylogeographic analyses.

2. MATERIAL AND METHODS

2.1. Early samples

We used HIV-1 sequences amplified from serum samples collected in 1986 in a hospital in Ouagadougou, Burkina Faso, and described by Fonjungo et al. (Fonjungo et al., 2014). They consisted of 52 protease (PR) and 37 gp41 sequences for CRF06_cpx and 15 PR and 25 gp41 for CRF02_AG, although the limited length of the sequences (297nt for PR and 366nt for gp41) can complicate the correct identification of the recombinants. To these sequences we added CRF06_cpx and CRF02_AG samples from Los Alamos National Laboratory HIV Database (LANLdb (LANL, 2015)).

2.2. Sequence processing pipeline

In brief, the pipeline (**Supplementary Figure 1**) involved the exclusion of any sequences that either: i) represented a duplicate, ii) were not included in a monophyletic cluster with reference sequences of both the recombinant and the parental variants (see below) in ML trees (**Supplementary Figures 2** and **3**), or iii) presented unusually high evolutionary rates as seen in a Path-O-Gen analysis. In addition, various sub-sampling strategies (described below) were applied to the different datasets in an attempt to obtain an unbiased phylogeographic analyses (Faria et al., 2014; Lemey et al., 2014).

2.3. CRF06_cpx dataset

We downloaded all CRF06_cpx PR (n=388, after removing duplicates using ElimDupes (http://www.hiv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html)) and gp41 (n=157) sequences with available sampling date from LANLdb, to which we added the early sequences. Preliminary trees were constructed with FastTree under the GTR+Γ model of substitution, in which clusters of closely related sequences sampled outside sub-Saharan Africa (e.g. Korea, Algeria) were present and down-sampled. Since CRF06_cpx sequences from both PR (protease) and *env* gp41 fall within the subtype G phylogenetic lineage, we ensured these samples were correctly classified by verifying that the CRF06_cpx sequences were monophyletic in Maximum Likelihood (ML) trees that included reference LANLdb sequences of subtype G (n=115 for PR, n=32 for gp41) trimmed from full-length genomes (**Supplementary Figure 2**). Non-monophyletic sequences were discarded (52 PR, whose trees were worse resolved, and 3 gp41). These ML trees were constructed with PhyML (Guindon et al., 2010) using the GTR+Γ model of substitution and testing the branch support using the approximate Likelihood-Ratio Test (aLRT). Finally, the datasets were analysed with Path-O-Gen v1.4 (tree.bio.ed.ac.uk/software/pathogen) and any sequence that represented outliers in the root-to-tip divergence versus sampling time regression analysis was removed.

In these ML trees most Burkina Faso CRF06_cpx sequences from 1986 formed clustered closely, especially for the PR gene (**Supplementary Figure 2**), so they were down-sampled and only the most

divergent ones in terms of genetic distances (n=2 for PR, n=6 for gp41) were kept in the definitive datasets. The fact that these sequences were amplified from samples taken in a single hospital during the same year could account for their genetic similarity, especially with limited sequence length and a high degree of genetic conservation (such as the PR gene). The rest of Burkina Faso sequences appeared interspersed along the ML trees, with a lack of country-specific clusters in general. This suggests a great degree of interchange between all countries (even from within and outside Africa).

For the evolutionary analysis, we selected sequences from sub-Saharan Africa only (see **Table 1** for the countries included) and, because of its epidemiological importance, a sub-sample (selecting the 4 most genetically divergent sequences for each gene) of the monophyletic Estonian/European outbreak among injecting drug users (to which approximately half of the CRF06_cpx gp41 sequences in LANLdb belonged). Excluding the early sequences from Burkina Faso sampled in 1986, the CRF06_cpx dataset contained sequences sampled between 1995 and 2013 for PR and between 1984 and 2010 for gp41.

As seen in **Table 1**, almost half the sequences (42% for both PR and gp41 datasets) came from Burkina Faso. To test the reliability of our phylodynamic and phylogeographic analysis, we randomly down-sampled the sequences from this country according to sampling year to avoid the bias that may arise from oversampling a particular location (Faria et al., 2014; Lemey et al., 2014). We therefore created a CRF06_cpx down-sampled dataset in which the Burkina Faso sequences (n=89 for PR, n=25 for gp41) were reduced to the number of sequences of the next most represented country: Mali for PR (n=25) and Ghana for gp41 (n=10).

2.4. CRF02_AG dataset

We downloaded all CRF02_AG PR and gp41 sequences in LANLdb from West and Central Africa and removed duplicates (yielding 1,247 PR / 194 gp41 from West Africa and 1,478 PR / 320 gp41 from Central Africa), to which we added the early Burkina Faso sequences (**Supplementary Figure 1**). As described for CRF06_cpx, we excluded both PR and gp41 sequences that were not monophyletic in ML trees containing samples from the parental subtypes (for CRF02_AG these are subtype G and A1 for PR and gp41, respectively) (**Supplementary Figure 3**). Also, to reduce the large numbers of sequences (especially for PR), COMET (Struck et al., 2015) was used to confirm their classification as CRF02_AG and misclassified sequences were removed.

This dataset was also analysed with Path-O-Gen v1.4 (see above). In addition, as happened with CRF06_cpx, the early CRF02_AG sequences from Burkina Faso clustered closely and were therefore down-sampled by keeping the ones with the most divergent genetic distance (4 PR and 9 gp41 in the definitive analysis). The West African PR sequences downloaded from LANLdb were also down-sampled by random selection of sequences within each sampling year and country, in order to have a

more balanced representation of sequences across the years and across countries while keeping a manageable number of them.

For the evolutionary analyses, we prepared 2 datasets: i) for a first analysis focused on Burkina Faso, we selected only sequences sampled in-country (130 PR and 39 gp41); ii) in a second analysis to explore the CRF02_AG spread throughout West Africa, we used all the selected sequences from West Africa together with the 5 closest sequences to each of them from Central Africa selected using ViroBLAST (Deng et al., 2007). **Table 1** shows the number of CRF02_AG sequences included for each country in the definitive dataset, which included sequences sampled between 1993 and 2010 for the PR and between 1984 and 2009 for the gp41 gene.

2.5. Evolutionary analyses of the CRF02_AG and CRF06_cpx datasets

We applied Bayesian MCMC inference to 4 datasets (complete CRF06_cpx and down-sampled CRF06_cpx, Burkina Faso CRF02_AG and African CRF02_AG) using BEAST v1.7.5 (Drummond et al., 2012). We employed a multilocus approach (Lemey et al., 2004) in order to analyse both PR and gp41 at the same time as dataset partitions. This feature allowed us to use the same models of nucleotide substitution, and demographic model to both partitions, while using different molecular clock models, so each partition can have different phylogenetic histories and coefficients of variation for the evolutionary rates. We used an uncorrelated log-normal relaxed molecular clock with the SRD06 model of nucleotide substitution (Shapiro et al., 2006) and the Bayesian Skygrid as the demographic model (Gill et al., 2013). The datasets ran for 100 million generations (200 million for the African CRF02_AG dataset) sampling evolutionary parameters every 10,000th generation. Convergence of the chains was manually identified using Tracer v1.5 (http://tree.bio.ed.ac.uk/software/tracer/). The posterior tree distributions were used to generate a maximum clade credibility (MCC) tree after removing a 10% burn-in using TreeAnnotator v1.7.5 (http://beast.bio.ed.ac.uk/TreeAnnotator). A Bayesian Skygrid plot was generated using Tracer v1.5 to represent the changes of the viral effective population size over time (Gill et al., 2013).

2.6. Phylogeographic analyses of CRF02_AG and CRF06_cpx

To test the phylogeographic structure of the datasets, the association index (AI), which indicates the strength of correlation between the phylogeny and the location of its tips, was calculated for the CRF06_cpx and CRF02_AG posterior PR and gp41 tree distributions excluding the first 10% of tree states as burn-in using BaTS (Bayesian Tip-association Significance testing (Parker et al., 2008)). BaTS tests the null hypothesis (i.e., no correlation between phylogeny and taxa location) by performing randomization tests across the posterior set of trees, thus incorporating phylogenetic uncertainty when testing for phylogeographic structure. The null set was calculated by 100 random replicates of the state distribution across the phylogeny.

We then conducted a Bayesian phylogeographic analysis using the asymmetric discrete traits analysis implemented in BEAST v1.7.5 to identify migration pathways of CRF02_AG (African dataset) and CRF06_cpx (complete and down-sampled datasets) between the different sampling locations. Each country was considered a location except those sequences belonging to the CRF06_cpx Estonian/Eastern European cluster, which grouped in a single location. Bayesian stochastic search variable selection (BSSVS) was employed to reconstruct the geographical locations of the ancestral states (Lemey et al., 2009) contained in a sub-sample of the posterior tree distribution of the BEAST runs described above. The migration routes indicated by the MCC tree (i.e. the changes of state in the nodes) were visualized using Google Earth (http://earth.google.com) and SPREAD v1.0.6 (Bielejec et al., 2011). SPREAD was also used to identify statistically well-supported migration routes using the Bayes Factor (BF).

3. RESULTS

3.1. Data processing and preliminary analyses

The main interest in the analyses lay in the use of early samples collected in Burkina Faso in 1986 together with background sequences from public databases. Due to the limited length of these sequences, we performed a strict data selection to ensure that the sequences would reflect the evolutionary history of the two analysed CRFs (**Supplementary Figure 1**). To reduce the high number of available CRF02_AG samples, additional down-sampling steps were incorporated (**Supplementary Figure 1**).

Finally, the analyses were restricted geographically to answer particular questions about the origin and spread of CRF06_cpx and the dissemination of CRF02_AG from Central to West Africa. For CRF06_cpx we only included sequences collected in sub-Saharan Africa and, due to its epidemiological relevance, a sub-sample from the outbreak which affected IDUs in Estonia and other European countries. For CRF02_AG, we first analysed sequences from Burkina Faso, then extended the analysis to West African sequences together with the closest Central African PR and gp41, in order to estimate the arrival and spread of this strain from the Congo River Basin. These Central African sequences came mostly from Cameroon and, to a lesser extent, Nigeria and Gabon.

3.2. Phylodynamics of CRF06_cpx

For the evolutionary studies in BEAST we analysed both genes jointly using them as partitions within the BEAST multilocus analysis (Lemey et al., 2004) to improve convergence and confidence intervals on the growth dynamics. Although both loci are allowed to have different genealogies, they are assumed to share the same demographic history.

The time to the most recent common ancestor (tMRCA) for the complete CRF06_cpx dataset was 1979 (95% Bayesian Confidence Interval (BCI) 1973-1983) for PR and 1981 (1978-1983) for gp41, with respective evolutionary rates of 5.5 (4.8-6.3) $\times 10^{-3}$ substitutions/site/year (s/s/y) and 6.1 (4.8-7.5) $\times 10^{-3}$ s/s/y.

The evolutionary rate found for CRF06_cpx PR was unusually high (as the usual rate for *pol* ranges between 1-3 $\times 10^{-3}$ s/s/y). The limited length of the PR gene (297nt) might prevent more accurate analyses, especially when applying the SRD06 substitution model which evaluates parameters independently for the first plus second (CP1+2) and the third codon positions (CP3). We examined the relative substitution rates of CP1+2 and CP3 and the former was 1.7-times higher than the latter, which is the opposite of what is expected in a coding region. However, this did not seem to apply to the PR evolutionary rate for CRF02_AG (see below). For CRF02_AG, the relative substitution rate was 1.5-times higher for CP3 than for CP1+2. We hypothesised that this high rate might have resulted from the

adaptation of the CRF06_cpx PR after diverging from the parental subtype G. To investigate this, we compared the evolutionary rates of 2 clades, creating a dataset including the 210 CRF06_cpx PR sequences analysed before as well as 51 subtype G PR reference sequences (from the subtype G sequences used in **Supplementary Figure 2**). We analysed this dataset with BEAST using the same settings as before. However, the MCC tree did not show any notable difference in the evolutionary rate (mean=3.34 × 10^{-3} s/s/y [95% BCI=2.6-4.1]) across its branches regardless the phylogenetic lineage. We then analysed a RAxML tree constructed from this dataset –containing CRF06_cpx plus subtype G– by using RELAX, a program designed to detect differences in selection pressures in a codon-based phylogenetic framework (Wertheim et al., 2015). Given two subsets of branches in a phylogeny, RELAX can determine whether selective strength was relaxed or intensified in one of these subsets relative to the other by calculating and comparing the distribution of the ratio of non-synonymous over synonymous substitutions (ω or dN/dS) along the branches of the tree. Indeed, RELAX found a statistically significant relaxation of selection in CRF06_cpx with respect to the parental G (parameter K=0.84, p=0.027). A relaxation of the selection pressure would account for the observed increase in the mutation rate, particularly in the CP1+2 codon positions.

The Bayesian Skygrid plot (**Figure 2A**) showed an exponential increase in the CRF06_cpx effective population size during the 1980s and 1990s, stabilizing afterwards. Around 1985 the plot shows a temporary increase in population size, probably an artefact due to the sequences' sampling dates: there were no available samples between 1986 and 1994 (for PR) or 1995 (for gp41). However, special care must be taken when interpreting the trends observed in these plots, as they are sensitive to sampling strategies.

3.3 Analysis of spatial structure

We used the program BaTS to test for the presence of statistically significant phylogeographic structure in the CRF06_cpx and CRF02_AG datasets applying the association index (AI). To compare the AI results across different datasets, the ratio between the observed index and the expected one provides a measure of the degree of association, with a value of 0 being complete association and values ≥1 showing no association. **Supplementary Table 1** contains the results of this test showing that, in all cases and all genes, the AI statistic strongly rejected the null hypothesis of random mixing ($p<0.001$). In addition, the index ratio showed that, for both CRF06_cpx and CRF02_AG, the gp41 gene showed a stronger spatial structure than PR; and, comparing the two variants, a stronger spatial structure was observed for CRF06_cpx (0.87 [0.82-0.92] for PR; 0.49 [0.42-0.55] for gp41) than for CRF02_AG (0.94 [0.92-0.96] for PR; 0.78 [0.75-0.80] for gp41). The high index ratio found for CRF02_AG PR, although significantly different than 1, indicates that the results from this dataset's phylogeography analysis should be considered with more caution.

3.4. Phylodynamics of CRF02_AG

The tMRCA for the Burkina Faso CRF02_AG dataset was 1979 (1977-1982) for PR and 1980 (1977-1981) for gp41 with rates of 2.3 (1.8-2.9) × 10^{-3} s/s/y and 5.7 (4.1-7.4) × 10^{-3} s/s/y, respectively. However, gp41 sequences from only 2 sampling times (1986 and 2003) were available.

The African datasets, which included CRF02_AG sequences from West Africa and the closest sequences from Central Africa, showed concordant dates for the tMRCAs of PR (1962 [1954 -1968]) and gp41 (1963 [1958 -1968]). They corresponded to evolutionary rates of 1.4 (1.3-1.6) × 10^{-3} s/s/y for PR and 4.5 (4.0-5.1) × 10^{-3} s/s/y for gp41. Therefore, there was a twenty-year gap between the origin of this CRF in central West Africa and its appearance in Burkina Faso, and that coincided with the appearance of the CRF06_cpx epidemic in that country. The CRF02_AG Bayesian Skygrid plot (**Figure 3A**) which included information from both PR and gp41 genes showed an exponential increase in the viral population size from the late 1960s to the early 1990s. During that time, CRF02_AG presumably travelled from Central to West Africa, and it spread progressively within the West Africa taking advantage of the great population mobility. After approximately 1995 the CRF02_AG population size stabilized.

Most of the Central African CRF02_AG sequences downloaded from LANLdb and selected for being closely related to the West African sequences came from Cameroon (see **Table 1**) and, to a lesser extent, from Nigeria, which was to be expected considering that these 2 countries are located between West Africa and the DRC and are thought to be, respectively, where the origin of CRF02_AG has been inferred and was first detected (Carr et al., 1998; Faria et al., 2012).

3.5. Phylogeography of CRF06_cpx

The gp41 Bayesian MCC tree generated in the phylogeographic analysis showed the tree root to be inferred in the Democratic Republic of Congo (DRC) with a posterior probability of 0.8. This would either show that CRF06_cpx or at least its parental subtype G lineage emerged in the Democratic Republic of Congo (DRC) around 1981 but it was first introduced soon after (1982, 1980-1983) to Burkina Faso (see **Figure 2B**), from which it spread to the rest of West Africa. In addition, in the tree we observed a CRF06_cpx cluster of Cameroonian sequences that also seemed to have come from DRC, although movements from Burkina Faso to Cameroon were also detected, suggesting that Cameroon, perhaps by virtue of its intermediate location, had introductions from these two sources.

There were no PR sequences from DRC and the discrete trait analysis of PR (**Figure 2B**) inferred an origin of CRF06_cpx in Burkina Faso (posterior probability=0.7) also in 1979, from which it spread to the remaining countries included in the analysis. The lack of PR sequences with sampling dates from the DRC meant it was not possible to test whether the origin of CRF06_cpx lay there.

Regardless of the original recombination location of CRF06_cpx, both PR and gp41 showed that this recombinant radiated from Burkina Faso to the rest of West Africa predominantly after around 1990. However, we observed (see trees in **Figure 2B**) a few CRF06_cpx disseminations just before 1990 from Burkina Faso to neighbouring countries: Mali (in both PR and gp41 trees), and Nigeria and Niger (in the gp41 tree). We estimated the statistically significant CRF06_cpx migration events using the Bayes Factor (BF) shown in **Figure 3**. For clarity, only those migrations supported by a BF ≥ 5 are shown. The best-supported migration was from DRC to Burkina Faso (BF=77 for gp41). The rest of the migration events, mainly between West African countries, had lower support. We found that the countries that were most frequently involved in these migration routes were Burkina Faso, Mali, Ghana and Senegal. Analysis of PR and gp41 agreed in the reconstruction of tight links between Burkina Faso, Mali and Ghana, whereas the reconstruction of other migration routes from and towards those countries and Central Africa and/or the West coast of Africa– relied on only one gene.

The tMRCA of the CRF06_cpx monophyletic cluster detected among injecting drug users in Estonia and other European countries (highlighted in the trees in **Figure 1B**) was 1995 (1989-1998) according to the PR Bayesian tree and 1995 (1992-1997) in the gp41 tree. The gp41 dataset included 2 sequences from each of Estonia and Russia. For PR, sequences from Estonia, Russia, Finland, Norway and Belarus (n=1 each) were available and included. As expected, the sequences from Estonia were basal within the cluster in both PR and gp41 analysis, suggesting it was the epicentre of the outbreak in Eastern Europe. To ascertain its source, more specific analyses would be needed; however, the discrete trait analysis pointed weakly to Burkina Faso (probability=0.30 and 0.44 for PR and gp41 respectively) as the ancestral location of this Estonian/European lineage. The second most probable locations were Mali (probability=0.12) and Ghana (probability=0.11) for PR and gp41, respectively.

The down-sampled CRF06_cpx dataset yielded similar results to the complete dataset. The estimated dates to the tMRCAs were 1977 (1976-1983) for PR and 1980 (1978-1982) for gp41, i.e. it was a little older for PR but virtually the same for gp41. The evolutionary rates obtained were 5.2 (3.8-6.9) × 10^{-3} s/s/y for PR and 5.7 (4.5-6.8) × 10^{-3} s/s/y for gp41, which were closer to each other than in the analysis of the complete dataset. Considering the reduced sample size of the down-sampled dataset, which could limit the amount of phylogenetic signal and information to be extracted from the dataset, the results were quite consistent with those from the complete CRF06_cpx dataset. The discrete phylogeographic analysis also inferred the root of the CRF06_cpx phylogenies in the same locations: Burkina Faso for PR (posterior probability=0.6) and DRC for gp41 (posterior probability=0.8). The migration events reconstructed were similar but showed lower statistical supports (data not shown).

Finally, in order to explore the origin of CRF06_cpx more deeply, the most basal (and oldest) sequences for this variant in PR and gp41 were included in ML trees (constructed with RAxML) together with the pure subtype G sequences of these 2 gene regions from sub-Saharan Africa (n=801 for PR, n=127 for

gp41) and found in LANLdb. The oldest CRF06_cpx PR sequence (BF.AY936826.1986) was closest to a subtype G sequence from Togo but within a larger cluster which included multiple sequences from both West Africa (Liberia, Senegal) and Central Africa (Congo, Nigeria, Gabon, Central African Republic and Cameroon). The oldest gp41 sequences (06_cpx.CD.AF260460.1985 and 06_cpx.BF.AY928426.1986, which were both included because of the different countries they were sampled from) showed more distinct clustering, in this case with 3 subtype G sequences from DRC. This high degree of sequence intermixing irrespective of the sampling country, particularly for PR, prevented a clear conclusion about the origin (whether in West or Central Africa) of the subtype G strains that formed CRF06_cpx. The limited length of PR and gp41 also prevented inference of phylogenetic relationships in any higher resolution.

3.6. Phylogeography of CRF02_AG

The Bayesian phylogenetic trees of both PR and gp41 (**Figure 3B**) showed CRF02_AG sequences interspersed regardless of their sampling country with a few exceptions of country-specific clusters. Indeed, the phylogeographic analysis of CRF02_AG in West Africa showed numerous viral migration events between different countries. The origin of CRF02_AG was placed in Cameroon for the PR analysis and in BF for the gp41 analysis, but this was not the primary goal of the CRF02_AG discrete trait analyses and these results could be biased by the number of sequences available for each country and their sampling dates.

Supplementary Figure 4 shows the CRF02_AG migration routes with the highest statistical support (BF≥5) for the PR and gp41 analyses. The most strongly supported routes were, for the PR analysis, from Cameroon to Gabon and Cameroon to Côte d'Ivoire (both with maximal BF support), from Benin to Togo (BF=8.4) and Nigeria to Senegal (BF=6.3). For the gp41 analysis, the most strongly supported routes were Burkina Faso to Côte d'Ivoire (BF=11.6), DRC to Burkina Faso (BF=9.8), Burkina Faso to Cameroon (BF=7.1) and Côte d'Ivoire to the Gambia (BF=7.0). Taken together, the results in **Supplementary Figure 4** show several CRF02_AG migration routes between Central and West Africa and vice versa. They also show routes within West Africa, among which Côte d'Ivoire and Ghana are the most heavily involved countries in these movements, with Burkina Faso playing a more secondary role than in CRF06_cpx transmission. Overall, the analyses of both CRF02_AG genes suggested different migration routes, which is probably explained by the different sampling and availability of sequences. In addition, the CRF02_AG dataset (particularly PR) showed weaker spatial structure than CRF06_cpx, so inferences from the former's phylogeography analysis should be taken with more caution.

4. DISCUSSION

We reconstructed the HIV dynamics of the recombinants HIV-1 CRF06_cpx and CRF02_AG in Burkina Faso and throughout West Africa using a multilocus Bayesian analysis, incorporating protease and gp41 sequences from samples collected in that country in 1986. We have shown that the tMRCA of these two recombinants –the most prevalent HIV-1 variants in Burkina Faso– were reconstructed in the country around 1980, implying a relatively young national epidemic compared to countries from other African regions (see below). Burkina Faso represents the epicentre for the transmission of CRF06_cpx in West Africa, after a possible origin in the Democratic Republic of Congo (DRC). On the other hand, CRF02_AG entered the region 10-15 years before the emergence of CRF06_cpx, and it seems to have experienced an extensive transmission throughout West Africa with a lack of clear patterns according to country of sampling.

Burkina Faso being such a young HIV-1 epidemic is consistent with a later introduction into West Africa. Indeed, the first cases of AIDS in West Africa were reported in 1985 (Benin and Mali), and in Burkina Faso in 1986, later than in East (Serwadda et al., 1985) and South (McCormack et al., 2002) Africa after the origin of HIV-1 in Central Africa. The estimated dates of introduction for both HIV-1 variants coincided with some of the most unstable years in the Burkina Faso's history and it was not until 1988 that HIV/AIDS was prioritized (Banhoro, 2013), which may have potentiated the evolution of the epidemic due to the lack of public health measures focused on stopping its transmission. HIV-2 was also detected (both HIV-2 groups A and B) at a low prevalence in samples from 1986 (Fonjungo et al., 2014).

We found an unusually high evolutionary rate for CRF06_cpx PR (around 5×10^{-3} s/s/y) and tested different explanations. We obtained statistically significant evidence of selection relaxation in the CRF06_cpx PR with respect to the parental subtype G. Relaxation of selection can occur when a new ecological niche is found, which could have happened following a founder effect phenomenon in which a viral lineage enters a fully susceptible population. Alternatively, if this relaxation occurred immediately after the recombination that gave rise to CRF06_cpx, it could have been caused by the fact that the CRF06_cpx PR gene was derived from subtype G while CRF06_cpx gag gene was derived from subtype A1. These two HIV proteins co-evolve due to the fact that the gag is the substrate of PR cleavage (Fun et al., 2012), and a gene shuffle could have required PR to recognise new cleavage sites in gag, which would imply the relaxation of the purifying selection that maintained this co-evolution and the high mutation rates observed in PR regardless its codon position. Our estimate of the time to the most recent common ancestor (tMRCA) for the African CRF06_cpx was 1979 (1973-1983) in PR and 1981 (1978-1983) in gp41, which agrees with the date provided by the only phylogeographic analysis of CRF06_cpx published to date (Delatorre and Bello, 2013), 1979 (1970-1985). Our estimates yielded smaller confidence intervals for PR, while ours is the first evolutionary analysis for the gp41

region. The spread of CRF06_cpx is intimately connected to Burkina Faso, and was first identified in a patient infected in this country (Oelrichs et al., 1998). The PR analysis confirmed the previous observation (which also utilized *pol* sequences) that reconstructed the origin of CRF06_cpx in Burkina Faso (Delatorre and Bello, 2013). On the other hand, our analysis of gp41 raised further possibilities, which are not apparent from single gene analysis, and established the origin of CRF06_cpx in DRC. The hypothesis that CRF06_cpx was formed in Burkina Faso by recombination between pre-existing strains instead of being imported is contradicted by the fact that, whereas CRF06_cpx represents around half the viruses found in the country, subtype G (parental for both CRF06_cpx genes analysed here) accounts for less than 1% of the Burkina Faso sequences deposited in the Los Alamos HIV database. Fonjungo and colleagues (Fonjungo et al., 2014) did not find any subtype G strain in samples from 1986, close to the beginning of the HIV epidemics in West Africa. Likewise, subtypes J and K (which are also involved in the recombination pattern of CRF06_cpx) have been detected in Central Africa but not in West Africa. This makes it less likely that there was ever any co-circulation in Burkina Faso of the putative parental variants that would have been required for the generation of this recombinant. An alternative hypothesis that would reconcile this Central African origin of CRF06_cpx with the fact that *pol* analyses point to Burkina Faso as its place of origin would be the following: CRF06_cpx was formed after a variant coming from Central Africa (that already included subtype G specifically in the *env* gene but also subtypes J and K) experienced successive recombination events with local subtype G viruses in or on its way to West Africa (perhaps in Nigeria or Cameroon) which formed the novel CRF06_cpx *pol* gene –so this CRF would eventually harbour subtype G fragments having different origins and evolutionary histories. The abundance of subtype G in Nigeria/Cameroon could make such a scenario plausible. Subtype A1, the other parental subtype that is part of CRF06_cpx, can be found in East, Central and West Africa, so its incorporation to the CRF06_cpx genome could have happened in either of the latter two regions. The limited availability of full-genome CRF06_cpx sequences is a severe constraint on the analyses that would be needed to ascertain the evolutionary history of the parental strains that formed this definitive recombinant.

Our phylogeographic study of CRF06_cpx was robust to the unbalanced availability of this strain's sequences (for which most samples in GenBank were collected in Burkina Faso) since the ancestral locations were confirmed in a second analysis for which we down-sample the sequences from this country to be comparable in number to other countries. Down-sampling of sequences from highly represented geographic locations or sampling times is needed to avoid potential bias in spatial inference estimates (Faria et al., 2014; Lemey et al., 2014). We have found that CRF06_cpx (regardless of its initial origin) spread throughout West Africa from Burkina Faso, particularly after 1990. This was common to the analyses of both PR and gp41. Our analyses reconstructed statistically significant bidirectional viral migration routes especially between Burkina Faso and Mali and Ghana, with links to Senegal and Côte d'Ivoire also. However, the lack of any HIV sequences from some countries or from

the 1980s prevented us from inferring the gradual movement of CRF06_cpx from Central Africa to Burkina Faso.

Because of its epidemiological importance, we also attempted to identify the origin of the CRF06_cpx clade that caused a continuing outbreak among IDUs in Eastern Europe (Zetterberg et al., 2004). We inferred that this monophyletic lineage, first detected in Estonia in 2004 appeared in that country around 1995 (1989-1998). However, we could not assign the specific origin of this outbreak in Africa with enough confidence.

The second recombinant virus analysed here, CRF02_AG, presents a different history. According to previous reports, it seems to have travelled northwards from DRC, where its had been inferred (Faria et al., 2012) to Cameroon, which acted as the epicentre of its dispersal (Faria et al., 2012; Véras et al., 2011), and eventually to Nigeria where it was first identified (Carr et al., 1998). Our study provides the first analysis of the origin and dispersal of CRF02_AG to and within West Africa, where it is the prevalent HIV-1 variant. We analysed substantially older CRF02_AG sequences than in previous works (including the oldest ones available from the 1980s) and our sub-sampling strategy was designed to provide a representative sample of CRF02_AG diversity throughout West Africa (including the closest lineages from Central Africa which might have been involved in its introduction) to reconstruct the mobility of this recombinant.

The dates that we inferred for the origin of CRF02_AG (1962 for PR and 1963 for gp41) are a decade older than estimates from previous studies (Abecasis et al., 2009; Faria et al., 2012) in which the earliest sequences included in their analyses were from 1990. Faria and colleagues obtained their estimate of 1973 (1972-1975) from samples collected in two cities in Cameroon between 1996 and 2004, whereas Abecasis and colleagues used a convenience sample from the Los Alamos HIV database. The only previous study of CRF02_AG in West Africa used 47 *env* sequences from Guinea-Bissau (in far West Africa) sampled in 1993-2007, and dated the establishment of CRF02_AG in this country through multiple introductions between 1976 and 1981 (Esbjornsson et al., 2011).

Our regional analysis of CRF02_AG shows that this viral variant expanded throughout West Africa showing a high degree of transmission between countries, demonstrated by the lack of clear migration routes and the phylogenetic intermixing of the sequences (also reported by (Véras et al., 2011)). This would agree with a fast spread of CRF02_AG in the region, and with the historical records of migration in West Africa: historically, this region has experienced a great mobility caused by population pressure, environmental conditions, poverty, and endemic conflicts. Migrants regarded this region as an economic unit where people moved freely, and borders between countries were porous and virtually uncontrolled, particularly since the creation of the Economic Community of West African States (ECOWAS) in 1975 –which removed obstacles to the free movement of goods, capital and people (Adepoju, 2005). Mobility traditionally involved labour migration from the landlocked, poorer

countries (including Burkina Faso) to road networks, plantations and mines in the coastal, richer countries –particularly Ghana, Côte d'Ivoire and Senegal. This would account for the fact that these countries, as constant recipients, as well as sources of outward migration, were heavily involved in the migration routes found for CRF02_AG and for CRF06_cpx. Epidemiological and political reasons also probably account for the fact that, despite being having being present in West Africa for 10-15 years, CRF02_AG did not enter Burkina Faso until ~1980 (the time when the CRF06_cpx epidemic started as well).

The inclusion of early sequences in phylodynamic studies can substantially improve estimates of viral evolutionary rates, through better calibration of the molecular clock (Worobey et al., 2008; Yebra et al., 2015). However, in our case the fact that small volume serum samples were almost 3 decades old when RT-PCR was performed meant that only short sequences (PR and gp41) could be generated (Fonjungo, 2014). We tried to compensate for this limitation by adopting a novel multilocus approach to the analysis (Gill et al., 2013; Lemey et al., 2004), by which information from both genes was incorporated in the demographic history reconstruction. However, since this allowed each gene to evolve at different rates and provide different phylogenies to resemble the different evolution of HIV genes, aspects relative to tree topology and the phylogeography analysis did not benefit from the multilocus approach. Another limitation of our study was the restricted and biased availability of sequences for each country, the lack of sequences from all West African countries, and the disparity of sampling for PR and gp41. However, for CRF06_cpx we were able to partially compensate for this limitation by performing further analyses after down-sampling the sequences from Burkina Faso, to prevent the results from being influenced by the higher availability of samples from this country.

Putting all our results together, we can conclude that Burkina Faso presents a relatively young HIV history, with the national epidemic caused by CRF02_AG and CRF06_cpx beginning around 1980. CRF06_cpx might have originally diversified (at least one of its parental variants) in Central Africa in the late 1970s, but we lack enough samples to confirm it. Regardless of its origin, it seems that CRF06_cpx radiated from Burkina Faso to the rest of West Africa from ~1990. CRF02_AG seems to have been introduced in West Africa from Cameroon/Nigeria much earlier, during the 1960s-early 1970s, where it expanded explosively perhaps fuelled by the outstanding mobility in the region, which in turn makes it very difficult to track its exact migration patterns or its origin/first place of introduction.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous refereees for providing very constructive comments that improved the original manuscript. This work was supported by the ICONIC project (supported by the Health Innovation Challenge Fund T5-344, a parallel funding partnership between the Department of

Health and Wellcome Trust), by the PANGEA_HIV Consortium (with support provided by the Bill & Melinda Gates Foundation) and by NIH GM110749.

REFERENCES

Abecasis, A.B., Vandamme, A.M., Lemey, P., 2009. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. J. Virol. 83, 12917-12924.

Adepoju, A., 2005. Migration in West Africa. A paper prepared for the Policy Analysis and Research Program of the Global Commission on International Migration. Available at: http://www.gcim.org/attachements/RS8.pdf.

Banhoro, Y., 2013. Vicissitudes of AIDS policies in Burkina Faso from 1985 to 2001: a historical perspective, in: T., F., Heaton, M.H. (Eds.), HIV/AIDS, illness, and African well-being. University of Rochester Press, pp. 287-308.

Bielejec, F., Rambaut, A., Suchard, M.A., Lemey, P., 2011. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27, 2910-2912.

Carr, J.K., Salminen, M.O., Albert, J., Sanders-Buell, E., Gotte, D., Birx, D.L., McCutchan, F.E., 1998. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology 247, 22-31.

Delatorre, E., Bello, G., 2013. Spatiotemporal dynamics of the HIV-1 CRF06_cpx epidemic in Western Africa. AIDS 27, 1313-1320.

Deng, W., Nickle, D.C., Learn, G.H., Maust, B., Mullins, J.I., 2007. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics 23, 2334-2336.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969-1973.

Esbjornsson, J., Mild, M., Mansson, F., Norrgren, H., Medstrand, P., 2011. HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: origin, demography and migrations. PLoS ONE 6, e17025.

Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem, A.J., Sousa, J.D., Arinaminpathy, N., Pepin, J., Posada, D., Peeters, M., Pybus, O.G., Lemey, P., 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 346, 56-61.

Faria, N.R., Suchard, M.A., Abecasis, A., Sousa, J.D., Ndembi, N., Bonfim, I., Camacho, R.J., Vandamme, A.M., Lemey, P., 2012. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. Infect. Genet. Evol. 12, 453-460.

Fonjungo, P.N., Kalish, M.L., Schaefer, A., Rayfield, M., Mika, J., Rose, L.E., Heslop, O., Soudre, R., Pieniazek, D., 2014. Recombinant viruses initiated the early HIV-1 epidemic in Burkina Faso. PLoS ONE 9, e92423.

Fun, A., Wensing, A.M., Verheyen, J., Nijhuis, M., 2012. Human Immunodeficiency Virus Gag and protease: partners in resistance. Retrovirology 9, 63.

Gill, M.S., Lemey, P., Faria, N.R., Rambaut, A., Shapiro, B., Suchard, M.A., 2013. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol. Biol. Evol. 30, 713-724.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307-321.

Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS 25, 679-689.

Kalish, M.L., Robbins, K.E., Pieniazek, D., Schaefer, A., Nzilambi, N., Quinn, T.C., St Louis, M.E., Youngpairoj, A.S., Phillips, J., Jaffe, H.W., Folks, T.M., 2004. Recombinant viruses and early global HIV-1 epidemic. Emerg. Infect. Dis. 10, 1227-1234.

LANL, 2015. Los Alamos HIV database. Available at: http://www.hiv.lanl.gov.

LANL, 2016. HIV Circulating Recombinant Forms (CRFs). Available at: http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html.

Lemey, P., Pybus, O.G., Rambaut, A., Drummond, A.J., Robertson, D.L., Roques, P., Worobey, M., Vandamme, A.M., 2004. The molecular population genetics of HIV-1 group O. Genetics 167, 1059-1068.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C.A., Smith, D.J., Pybus, O.G., Brockmann, D., Suchard, M.A., 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 10, e1003932.

Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5, e1000520.

Lihana, R.W., Ssemwanga, D., Abimiku, A., Ndembi, N., 2012. Update on HIV-1 diversity in Africa: a decade in review. AIDS Rev. 14, 83-100.

McCormack, G.P., Glynn, J.R., Crampin, A.C., Sibande, F., Mulawa, D., Bliss, L., Broadbent, P., Abarca, K., Ponnighaus, J.M., Fine, P.E., Clewley, J.P., 2002. Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. J. Virol. 76, 12890-12899.

Oelrichs, R.B., Workman, C., Laukkanen, T., McCutchan, F.E., Deacon, N.J., 1998. A novel subtype A/G/J recombinant full-length HIV type 1 genome from Burkina Faso. AIDS Res. Hum. Retroviruses 14, 1495-1500.

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

Serwadda, D., Mugerwa, R.D., Sewankambo, N.K., Lwegaba, A., Carswell, J.W., Kirya, G.B., Bayley, A.C., Downing, R.G., Tedder, R.S., Clayden, S.A., et al., 1985. Slim disease: a new disease in Uganda and its association with HTLV-III infection. Lancet 2, 849-852.

Shapiro, B., Rambaut, A., Drummond, A.J., 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23, 7-9.

Struck, D., Lawyer, G., Ternes, A.M., Schmit, J.C., Bercoff, D.P., 2015. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 42, e144.

Tebit, D.M., Ganame, J., Sathiandee, K., Nagabila, Y., Coulibaly, B., Krausslich, H.G., 2006. Diversity of HIV in rural Burkina Faso. J Acquir Immune Defic Syndr 43, 144-152.

Tebit, D.M., Sangare, L., Tiba, F., Saydou, Y., Makamtse, A., Somlare, H., Bado, G., Kouldiaty, B.G., Zabsonre, I., Yameogo, S.L., Sathiandee, K., Drabo, J.Y., Krausslich, H.G., 2009. Analysis of the diversity of the HIV-1 pol gene and drug resistance associated changes among drug-naive patients in Burkina Faso. J Med Virol 81, 1691-1701.

UNAIDS, 2015. Burkina Faso country report. Available at: http://www.unaids.org/en/regionscountries/countries/burkinafaso.

Véras, N.M., Santoro, M.M., Gray, R.R., Tatem, A.J., Lo Presti, A., Olearo, F., Cappelli, G., Colizzi, V., Takou, D., Torimiro, J., Russo, G., Callegaro, A., Salpini, R., D'Arrigo, R., Perno, C.F., Goodenow, M.M., Ciccozzi, M., Salemi, M., 2011. Molecular epidemiology of HIV type 1 CRF02_AG in Cameroon and African patients living in Italy. AIDS Res. Hum. Retroviruses 27, 1173-1182.

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., Scheffler, K., 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Mol. Biol. Evol. 32, 820-832.

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.J., Kabongo, J.M., Kalengayi, R.M., Van Marck, E., Gilbert, M.T., Wolinsky, S.M., 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 455, 661-664.

Yebra, G., Ragonnet-Cronin, M., Ssemwanga, D., Parry, C.M., Logue, C.H., Cane, P.A., Kaleebu, P., Leigh-Brown, A.J., 2015. Analysis of the history and spread of HIV-1 in Uganda using phylodynamics. J. Gen. Virol. 96, 1890-1898.

Zetterberg, V., Ustina, V., Liitsola, K., Zilmer, K., Kalikova, N., Sevastianova, K., Brummer-Korvenkontio, H., Leinikki, P., Salminen, M.O., 2004. Two viral strains and a possible novel recombinant are responsible for the explosive injecting drug use-associated HIV type 1 epidemic in Estonia. AIDS Res. Hum. Retroviruses 20, 1148-1156.

Table 1: HIV-1 sequences included in the evolutionary analysis using BEAST.

		CRF06_cpx complete dataset		CRF02_AG African dataset	
Country	**Code**	**PR**	**gp41**	**PR**	**gp41**
West Africa		168	44	381	170
Burkina Faso	BF	89	25	38	38
Mali	ML	25	3	67	0
Côte d'Ivoire	CI	6	1	25	25
Ghana	GH	11	8	89	28
Togo	TG	17	1	32	0
Benin	BJ	8	0	50	0
Senegal	SN	12	6	78	62
Guinea-Bissau	GW	0	0	0	2
Liberia	LR	0	0	1	0
The Gambia	GM	0	0	1	15
Central Africa		34	12	311	128
Angola	AO	0	0	5	0
Chad	TD	1	0	0	0
Cameroon	CM	4	5	207	120
Central African Republic	CF	1	0	1	0
D. R. Congo	CD	0	3	5	2
Gabon	GA	2	2	32	1
Equatorial Guinea	GQ	1	0	10	0
Nigeria	NG	25	1	51	5
Niger	NE	0	1	0	0
Southern Africa		3	0	-	-
Madagascar	MG	2	0	-	-
South Africa	ZA	1	0	-	-
Estonia/Europe	EE	5	4	-	-
Total		**210**	**60**	**691**	**298**

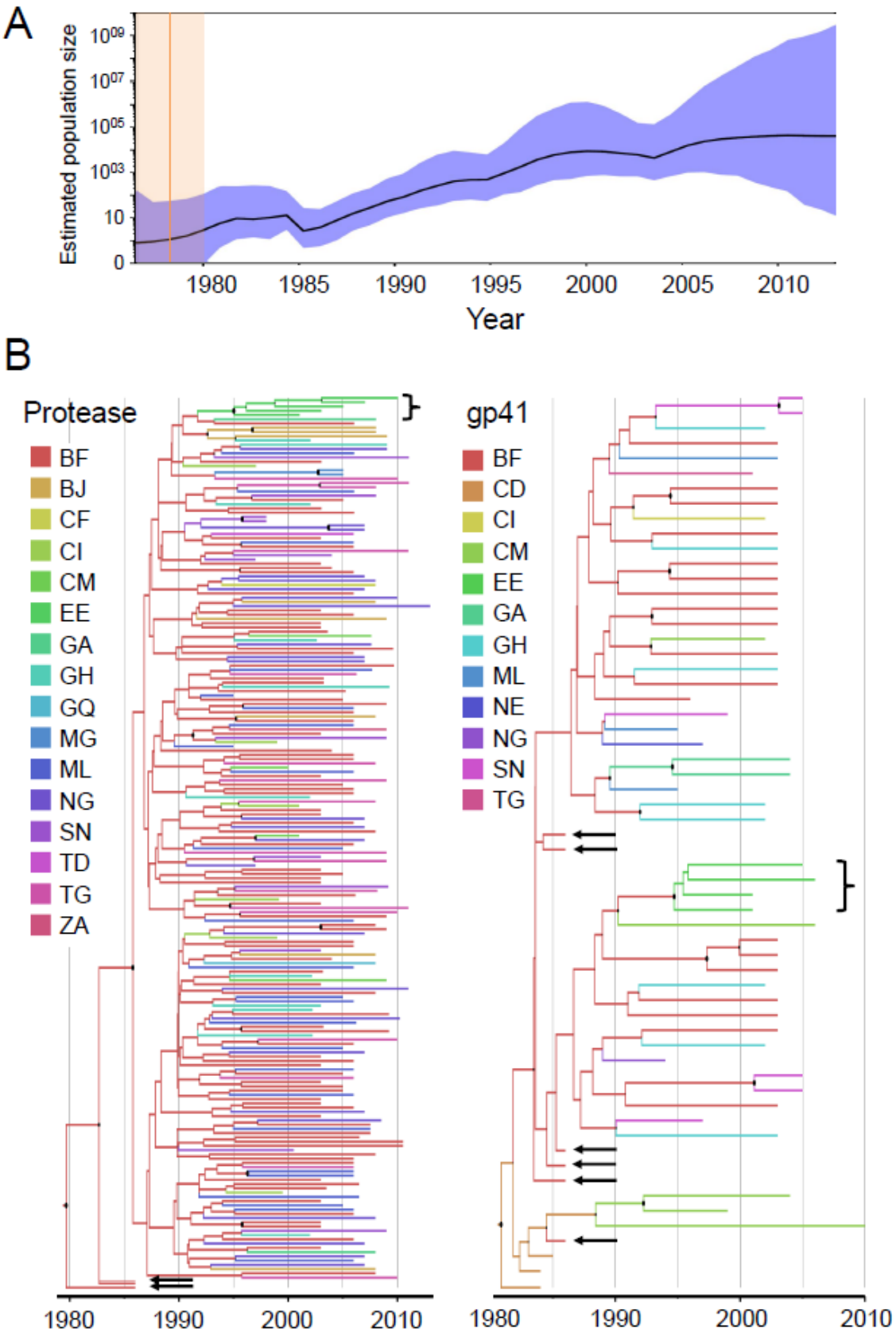
PR, protease. Not shown (zero sequences): Guinea, Sierra Leone, Mauritania (West Africa), Congo and Angola (Central Africa).

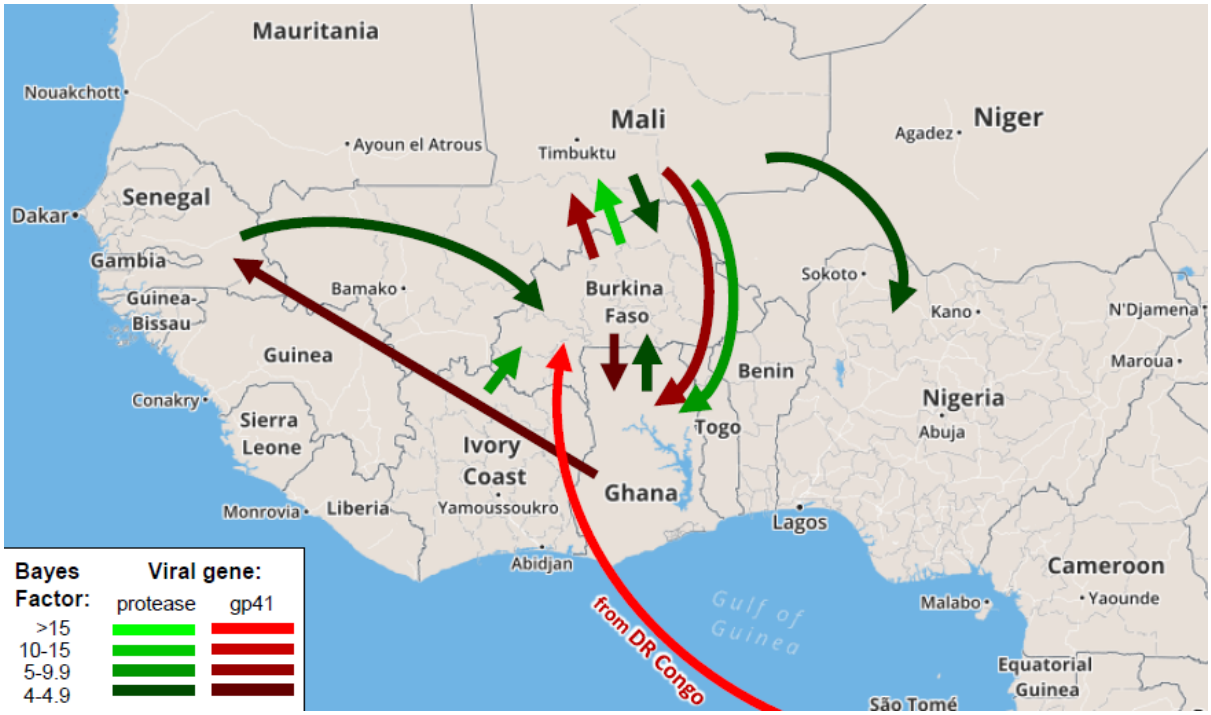
FIGURE CAPTIONS

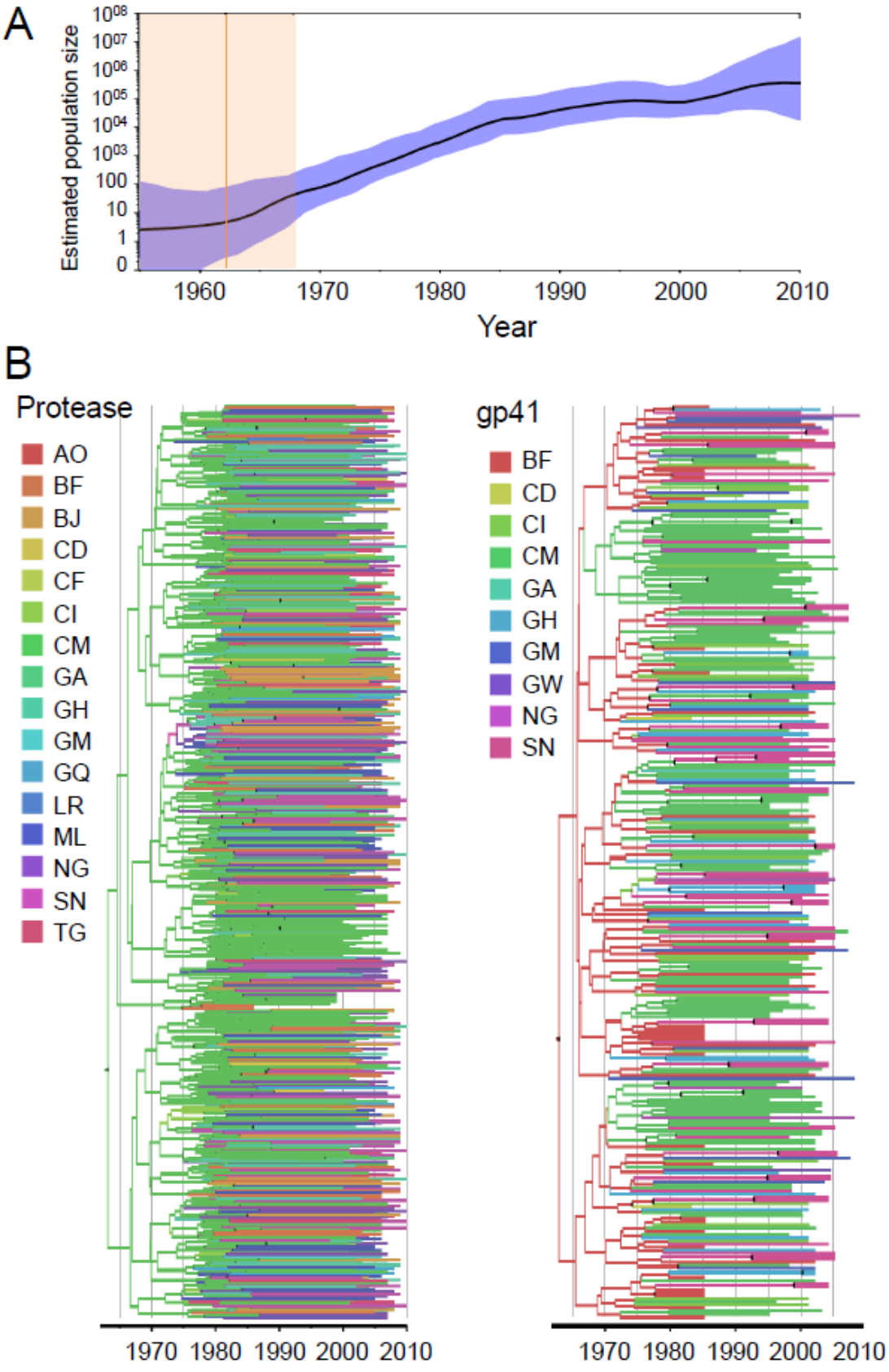
Figure 1: A) Bayesian Skygrid plot of the multilocus analysis showing the changes in the estimated viral effective population size for CRF06_cpx across time. The solid black line indicates the median population size and the purple shading the 95% Bayesian confidence interval (BCI). The solid vertical orange line indicates the date of the most recent common ancestor, and the orange shading its 95% BCI. **B) Time-stamped maximum clade credibility trees of the CRF06_cpx protease (left) and gp41 (right) discrete trait analysis with BEAST**. Branches are coloured (see legend for each tree) according to the location of the sequences (terminal branches) or inferred ancestral location (internal branches). The arrows highlight the Burkina Faso sequences from 1986. Country codes are shown in **Table 1**. Nodes supported by a posterior probability ≥0.7 are indicated by a black circle, and the braces highlight the Estonian/European cluster. Horizontal axes of both panels are expressed in calendar years.

Figure 2: Migration routes of CRF06_cpx inferred from the phylogeographic analyses of protease (in green) and gp41 (in red) genes. For clarity, only those routes statistically supported by a Bayes Factor ≥4 are shown. The colour of the arrows is proportional to the statistical support (see legend). This figure was elaborated using maps from MapBox (www.mapbox.com).

Figure 3: A) Bayesian Skygrid plot of the multilocus analysis showing the changes in the estimated viral effective population size for African CRF02_AG across time. The solid black line indicates the median population size and the purple shading the 95% Bayesian confidence interval (BCI). The solid vertical orange line indicates the date of the most recent common ancestor, and the orange shading its 95% BCI. **B) Time-stamped maximum clade credibility trees of the CRF02_AG protease (left) and gp41 (right) discrete trait analysis with BEAST**. Branches are coloured (see legend for each tree) according to the location of the sequences (terminal branches) or inferred ancestral location (internal branches). The arrows highlight the Burkina Faso sequences from 1986. Country codes are shown in **Table 1**. Nodes supported by a posterior probability ≥0.7 are indicated by a black circle. Horizontal axes of both panels are expressed in calendar years.


A
Estimated population size
10^09
10^07
10^05
10^03
10
0
1980
1985
1990
1995
2000
2005
2010
Year
B
Protease
BF
BJ
CF
CI
CM
EE
GA
GH
GQ
MG
ML
NG
SN
TD
TG
ZA
1980
1990
2000
2010
gp41
BF
CD
CI
CM
EE
GA
GH
ML
NE
NG
SN
TG
1980
1990
2000
2010


Mauritania
Nouakchott
Mali
Ayoun el Atrous
Timbuktu
Agadez
Niger
Senegal
Dakar
Gambia
Bamako
Burkina Faso
Sokoto
Kano
N'Djamena
Guinea-Bissau
Guinea
Maroua
Benin
Conakry
Nigeria
Abuja
Sierra Leone
Togo
Ivory Coast
Ghana
Monrovia
Liberia
Yamoussoukro
Lagos
Abidjan
Cameroon
Yaounde
Malabo
Gulf of Guinea
from DR Congo
Equatorial Guinea
São Tomé
Bayes Factor:
Viral gene:
protease
gp41
>15
10-15
5-9.9
4-4.9


A
Estimated population size
10^{08}
10^{07}
10^{06}
10^{05}
10^{04}
10^{03}
100
10
1
0
1960
1970
1980
1990
2000
2010
Year
B
Protease
AO
BF
BJ
CD
CF
CI
CM
GA
GH
GM
GQ
LR
ML
NG
SN
TG
1970 1980 1990 2000 2010
gp41
BF
CD
CI
CM
GA
GH
GM
GW
NG
SN
1970 1980 1990 2000 2010