Virus Evolution, 2017, 3(2): vex032

doi: 10.1093/ve/vex032
Research article

Divergent HIV-1 strains (CRF92_C2U and CRF93_cpx) co-circulating in the Democratic Republic of the Congo: Phylogenetic insights on the early evolutionary history of subtype C

C.J. Villabona Arenas,[1,*,†,§] N. Vidal,[1,‡,§] S. Ahuka Mundeke,[1,2,3] J. Muwonga,[3,4] L. Serrano,[1] J.J. Muyembe,[2,3] F. Boillot,[5] E. Delaporte,[1] and M. Peeters[1]

[1]Unité Mixte Internationale 233, Institut de Recherche pour le Développement, INSERM U1175, Université de Montpellier, 911 Avenue Agropolis, Montpellier, 34394, France, [2]Institut National de Recherche Biomédicale, Av. De la Démocratie 5345, Kinshasa, Democratic Republic of the Congo, [3]Cliniques Universitaires de Kinshasa, Route de Kimwenza, Kinshasa, Congo, Democratic Republic of Congo, [4]Laboratoire National de Référence du SIDA, Kinshasa, Democratic Republic of Congo and [5]Alter—Santé Internationale and Développement, Montpellier, 34090, France

*Corresponding author: E-mail: christian-julian.villabona-arenas@ird.fr

†http://orcid.org/0000-0001-9928-3968

‡http://orcid.org/0000-0001-7022-2643

§Contributed equally to this study.

Abstract

Molecular epidemiological studies revealed that the epicenter of the HIV pandemic was Kinshasa, the capital city of the Democratic Republic of the Congo (DRC) in Central Africa. All known subtypes and numerous complex recombinant strains co-circulate in the DRC. Moreover, high intra-subtype diversity has been also documented. During two previous surveys on HIV-1 antiretroviral drug resistance in the DRC, we identified two divergent subtype C lineages in the *protease* and partial *reverse transcriptase* gene regions. We sequenced eight near full-length genomes and classified them using bootscanning and likelihood-based phylogenetic analyses. Four strains are more closely related to subtype C although within the range of inter sub-subtype distances. However, these strains also have small unclassified fragments and thus were named CRF92_C2U. Another strain is a unique recombinant of CRF92_C2U with an additional small unclassified fragment and a small divergent subtype A fragment. The three remaining strains represent a complex mosaic named CRF93_cpx. CRF93_cpx have two fragments of divergent subtype C sequences, which are not conventional subtype C nor the above described C2, and multiple divergent subtype A-like fragments. We then inferred the time-scaled evolutionary history of subtype C following a Bayesian approach and a partitioned analysis using major genomic regions. CRF92_C2U and CRF93_cpx had the most recent common ancestor with conventional subtype C around 1932 and 1928, respectively. A Bayesian demographic reconstruction corroborated that the subtype C transition to a faster phase of exponential growth occurred during the 1950s. Our analysis showed considerable differences between the newly discovered early-divergent strains and the conventional subtype C and therefore suggested that this virus has been diverging in humans for several decades before the HIV/M diversity boom in the 1950s.

Key words: HIV-1; subtype C; molecular epidemiology; phylogeny; democratic; Republic of the Congo.

1. Introduction

Human immunodeficiency virus type 1 (HIV-1) strains are divided into four major genetic groups, M, N, O, and P, which represent separate transmissions of simian strains to humans. HIV-1 group M (HIV-1/M) is by far the most prevalent and its pandemic's origins are traced back to the 1920s (with 95% of estimated dates between 1909 and 1930) in the city of Kinshasa, in the modern-day Democratic Republic of the Congo (DRC) (Vidal et al. 2000; Faria et al. 2014). Between the 1920s and the 1950s, multiple factors (i.e. urban growth, strong railway links developed during Belgian colonial rule and disproportionate sex ratios in the early 20th century—a predominance of males in urban settings) led to the fast spread of HIV from Kinshasa to other regions in the DRC, and subsequently across the globe (Faria et al. 2014; Pineda-Pena et al. 2016; Mir et al. 2016).

The spread of particular strains at different time points led to founder effects and, consequently, to distinctive lineages that are conveniently named subtypes (i.e. subtypes A–D, F–H) (Robertson et al. 2000). However, recombination is a hallmark of HIV evolution and inter-subtype recombinants have continuously emerged from patients co-infected with different subtypes (Zhang et al. 2010). Recombinant forms that start a chain of infection are named circulating recombinant forms (CRFs). The prevalence of CRFs is increasing all over the world (Angelis et al. 2015; Lu et al. 2016; Oster et al. 2017), and some reports suggest that they might have better fitness when compared to their parental subtypes (Njai et al. 2006; Kouri et al. 2015; Turk and Carobene 2015).

Several studies suggest the preferential selection of certain drug resistance mutations in subtype C (Loemba et al. 2002; Brenner et al. 2003, 2006; Skhosana et al. 2015; Huang et al. 2016) when compared to subtype B, which is the most prevalent genetic form in high-income countries. However, subtype C accounts for approximately 50 per cent of current HIV infections. Although the majority of subtype C infections are in southern Africa, this subtype has spread to Brazil, China, east Africa, India, and European countries (Bello et al. 2008; Dalai et al. 2009; Hemelaar 2011; Abecasis et al. 2013).

The DRC has been shown to harbor the highest number of co-circulating subtypes (including intra-subtype diversity, i.e. sub-subtypes), complex CRFs and unique recombinant forms (URFs), and basal non-classifiable strains (Vidal et al. 2000; Niama et al. 2006; Rodgers et al. 2017). During two previous surveys on HIV-1 antiretroviral drug resistance in the DRC (Muwonga et al. 2011; Boillot et al. 2016), we identified two subtype C lineages that fell basal (i.e. a monophyletic divergent clade that shared its most common ancestral with subtype C) to available subtype C sequences in the *pol* gene. In this study, we characterize these lineages using eight newly sequenced near full-length genomes and provided further insights on the early epidemic history of HIV/M subtype C.

2. Methods

2.1 Samples

The eight divergent samples were collected as plasma in the capital city of Kinshasa ($n = 1$) and Mbuyi-Mayi (region of Kasai-Oriental, 1,300 km apart from the capital) ($n = 6$) in 2008 or as a dried blood spot ($n = 1$) in a decentralized primary health care facility in the Nord-Kivu province (2,500 km from Kinshasa) in 2012, during previously reported surveillance studies on drug resistance in the DRC (Muwonga et al. 2011; Boillot et al. 2016).

2.2 Sample preparation and sequencing

Nucleic acid extracts were prepared from plasma using the *QIAamp Viral RNA kit* (Qiagen, Courtaboeuf, France) and from dried blood spots using the *NucliSens miniMAG extraction system* (BioMérieux, Craponne, France) according to manufacturer's instructions.

RNA was transcribed with the *Expand Reverse Transcriptase* (Roche Diagnostics, Meylan, France) and reverse primers IN3 (5'-TCTATBCCATCTAAAAATAGTACTTTCCTGATTCC-3', positions 4,212–4,246 on HXB2) and LSIG1 (5'-TCAAGGCAAGCTTTATT GAGGCTTAAGCAG-3', positions 9,599–9,628 on HXB2). DNA amplification of overlapping genomic fragments was done using nested PCR and the *Expand Long Template PCR system* (Roche diagnostics, Meylan, France) as described previously (Vergne et al. 2000).

The amplified fragments were purified using the *Geneclean Turbo Kit* (Q-Biogen, MP-Biomedicals, France) and directly sequenced with *BigDye Terminator version 3.1 sequencing kit* (Life Technologies, Courtaboeuf, France). Electrophoresis and data collection were done on a 3130XL Genetic Analyzer. Sequences from both strands were reconstituted using SeqMan Pro tool from the package DNAstar v11.2.1 (Lasergene, Madison, WI).

2.3 Subtype/CRF determination of the new strains

The new genomic sequences were combined with representatives of each subtype, sub-subtype, and CRFs of HIV-1/M that have been reported in Africa, including the latest reported full-length genome sequences from DRC (Rodgers et al. 2017) and non-classifiable strains from the DRC (Mokili et al. 2002). A multiple sequence alignment (MSA) was obtained using MAFFT v7 (Katoh and Standley 2013), manually checked and end-trimmed. Poorly aligned positions or divergent regions were eliminated with Gblocks (Castresana 2000; Talavera and Castresana 2007).

The final MSA was then used to test every new genomic sequence for recombination using similarity and bootscan plot analysis with Simplot v3.5.1 (Lole et al. 1999). Analyses were done using a window of 400–500 base pairs (bp) with 10- to 20-bp increments. The analysis was later refined with a more restricted group of reference sequences and a varying window length with 10-bp increments to better define breakpoints. Finally, the alignment was cut into different segments and each of them was submitted to phylogenetic analysis to corroborate any recombination event. Each distinctive segment was used for Maximum-Likelihood (ML) phylogenetic analysis of using a GTR + 4Γ+I (general time-reversible plus among-site rate heterogeneity and invariant sites) nucleotide substitution model as implemented in PhyML 3.0 (Guindon et al. 2010). The best topology after SPR (subtree pruning and regrafting) topological moves and nearest neighbor interchanges analyses were selected and approximate likelihood ratios (aLRT) were used to assess confidence of the groups.

The Subtyping Distance tool (SUDI) from Los Alamos database (https://www.hiv.lanl.gov/content/sequence/SUDI/sudi.html) was used to calculate genetic distances and determine if any new distinctive clade should most appropriately be

considered a new sub-subtype. The input alignment consisted of our HIV-1/M reference sequences plus an HIV-1/N sequence as out-group reference (strain N_95CM.YBF30).

Finally, exhaustive local alignment searches using BLAST were done to identify sequences with high similarity to the newly generated near full-length genomes.

2.4 Reconstruction of dated phylogenies

We screened Los Alamos database to generate an HIV-1 subtype C reference dataset. We retrieved all available HIV subtype C full-genome sequences with known year and geographic location and retained one sequence from each patient. We conducted preliminary ML phylogenetic analysis using FastTree v.2 (Price et al. 2009, 2010) and a GTR + 4Γ nucleotide substitution model to identify clonal records, and narrowed the number of sequences to be representative of countries but particularly of years (Supplementary Table S1). This down-sampling reduced computational burden and comprised representative subtype C sequence samples ($n = 92$) isolated between 1986 and 2014. We also included subtype A/CRF02_AG reference sequences ($n = 24$; eight sequences from CRF02_AG, eight sequences from subtype A1, and two sequences from every other A sub-subtype) to account for the similarity of some regions of the novel strains with these subtypes. We generated profile alignments with Mafft v.7 (Katoh and Frith 2012, 2013), followed by rounds of automated and manual refining in Muscle v.3.8.31 (Edgar 2004a,b) and Mesquite v.3.2 (Maddison and Maddison 2002), respectively; this final alignment comprised 124 sequences.

In order to date relevant evolutionary events, time-scaled evolutionary analyses were done using either ML analysis as implemented in PhyML (Guindon et al. 2010)—where branch lengths were converted into units of real calendar time using Least-Squares Dating in LSD-0.3beta (To et al. 2016)—or using Markov chain Monte Carlo (MCMC) sampling, as implemented in BEAST v1.8.3 software package (Drummond et al. 2012). We also used the BEAGLE parallel computation library to enhance the speed of the likelihood calculations (Suchard and Rambaut 2009). Reconstructions were done using *gag*, *pol* and *env* genes—A partitioned analysis (only the demographic model was shared) in BEAST or individually in PhyML. Regression of root-to-tip genetic distance against sampling time was used to explore temporal signal and data quality using TempEst v.1.5 (Rambaut et al. 2016). All datasets were analyzed using a GTR + 4Γ+I nucleotide substitution model. For BEAST, we used a relaxed uncorrelated lognormal molecular clock model in order to infer the timescale of HIV evolution while accommodating among-lineage rate variation (Drummond et al. 2006) and a Bayesian skygrid model as a non-parametric coalescent tree prior (Gill et al. 2013). For each dataset, two to six MCMC chains of 80–200 million steps were computed. Samples were combined and diagnosed using visual trace inspection and calculation of effective sample sizes in Tracer (Rambaut et al. 2014).

We performed further analyses using the partial p51-RT gene in order to include 160 additional sequences from our surveillance studies in the DRC (Vergne et al. 2000; Vidal et al. 2006; Muwonga et al. 2011; Boillot et al. 2016). In this analysis, we placed an informative prior on the root based on our previous findings (this partial p51-RT dataset comprised 284 sequences). These sequences were sampled between 2002 and 2013 from the capital city of Kinshasa ($n = 35$), Kongo-Central ($n = 4$), Kasai-Oriental ($n = 14$), Tshopo and North Kivu ($n = 31$), and Haut-Katanga ($n = 65$) regions; for 11 additional sequences, the specific region of origin in the DRC was unknown.

In addition, we performed analyses excluding the A/CRF02_AG reference datasets and nearly identical sequences from the partial p51-RT dataset to describe only the subtype C viral population dynamics; this down sampling resulted in 245 sequences. Using the subtype C subset, we also performed demographic model testing (exponential, logistic and two-phase exponential-logistic growth) via path sampling and stepping-stone (100 steps with two million iterations each) that resulted in the exponential-logistic growth as the best-fit parametric coalescent tree model (Supplementary Table S2). Therefore, we also used the two-phase exponential-logistic growth to estimate the population growth rate of subtype C for each period and placed an informative prior on the time of transition between them based on previous HIV-1/M findings (Faria et al. 2014).

Finally, given that our preliminary analysis suggested that the envelope gene of some novel sequences was different from those lineages of CRF02_AG described in the literature and also that our reference datasets was only a small representative of this group, we decided to further explore the phylogenetic relationships of the novel strains in the envelope gene by including a bigger set of samples from subtype A ($n = 160$) and CRF02_AG ($n = 121$), representative of countries and years. This alignment comprised 382 sequences and phylogenetic reconstruction was done using a ML analysis plus a Least-Squares dating approach (Guindon et al. 2010; To et al. 2016).

3. Results

3.1 Subtyping of the new HIV-1 strains

The genomic structures of the new strains are depicted in Fig. 1. Bootscans and the corresponding phylogenetic profile findings are given in Supplementary Figs S1–S3. The aLRT values for groups were overall ≥90. There were only two instances of a lower statistic support: partition 1,535–1,750 (215 bp, aLRT = 0.81) and partition 3,630–4,510 (880 bp, aLRT 0.84) for samples 367, 653, and 817 (Supplementary Fig. S3). Overall, the different analyses revealed the presence of three different recombinant patterns. The first recombinant profile (referred as CRF92_C2U) contains large fragments related to subtype C and small unclassified fragments. The second profile is a unique recombinant strain with a predominant CRF92_C2U profile and two additional small regions from different subtypes, that is, one unclassified and one small subtype A fragment. Finally, another group of complex recombinant viruses (referred as CRF93_cpx) contains two fragments of divergent subtype C sequences, which are not conventional subtype C nor the above described C2, plus multiple divergent A/CRF02_AG fragments.

3.1.1 CRF92_C2U: sub-subtype C2 with two small unclassified fragments

Three samples that were collected in Mbuyi-Mayi during 2008 (strains 699, 796, and 819) and one sample that was collected in North-Kivu during 2012 (strain VIR90) always clustered together and basal to subtype C reference strains (Supplementary Fig. S1). The infected individuals did not have an epidemiological link. These sequences were predominantly related to subtype C and our SUDI analyses (Supplementary Fig. S4) showed that they fell within the distance range of inter-sub-subtype distances; henceforth we named them sub-subtype C2.

In addition, two small genomic regions did not group consistently with any known subtype or CRF and SUDI analyses

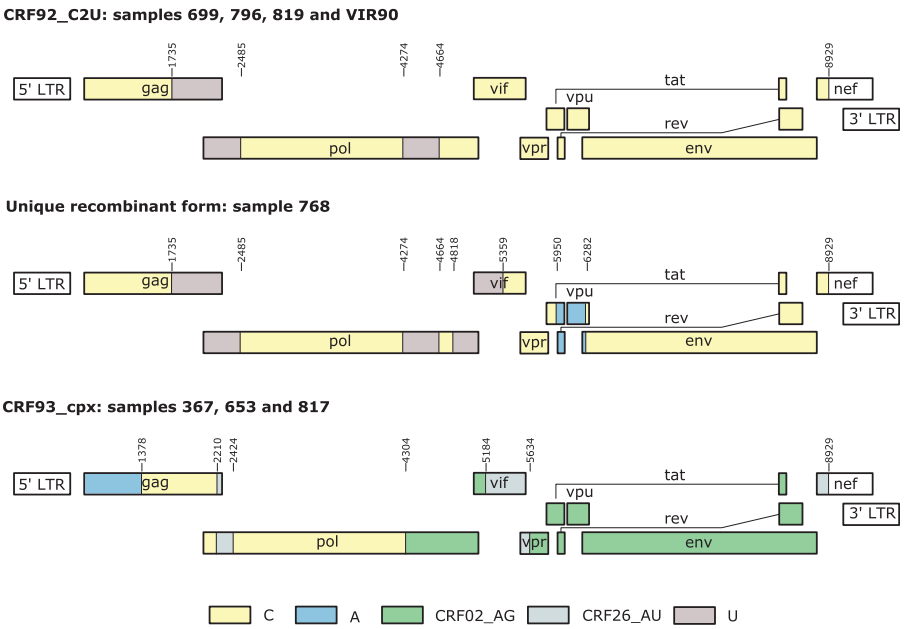

Figure 1. Genomic structure of the newly generated near full-length genomes. Recombinant breakpoint data were mapped to the HXB2 genome using Los Alamos Recombinant HIV-1 Drawing Tool v2.1.0 (https://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html). Colors indicate the subtypes with which recombinant regions have the highest level of identity, however, all these regions are different from conventional subtype and sub-subtype lineages described in the literature. U, non-classified/unknown.

showed that they fell in the range of inter-subtype distances (Supplementary Figs S1 and S4); we referred to them as unknown. The first unknown region (557 bp, from 1,735 to 2,292 on HXB2) covered the end of the *gag* gene (comprising *p7 nucleocapsid*, *p6* and the small spacer peptides, *p1* and *p2*) and part of the protease in the *pol* gene. The second unknown region (390 pb, from 4,274 to 4,663 on HXB2) covered part of the *integrase* in the *pol* gene.

BLAST searches resulted in three similar sequences to CRF92_C2U. The first one was a near-full-length genomic sequence sampled in 2004 from an unknown location (strain LA08SySa, accession KU168263) that was originally labeled as subtype C (Berg et al. 2016). Another near-full-length genomic sequence was sampled in 2002 in the Kwilu region, DRC (strain CG-0151-02V, accession KY392767) and originally described as divergent subtype C (Rodgers et al. 2017). LA08SySa and CG-0151-02V have a similar genomic organization as CRF92_C2U and grouped within the respective sub-subtype C2 clade (Supplementary Fig. S1). Finally, we found an *env* sequence (strain VI1358, accession HQ912711) that was collected in 1994 from an infected individual who regularly attended a clinic in Belgium and who was from sub-Saharan Africa (Balla-Jhagjhoorsingh et al. 2011). Further searches in our laboratory database found eight additional sequences from the DRC, which clustered within the C2 clade in the V3-V5 *env* gene ($n = 6$; aLRT = 0.90) or the partial p51-RT regions ($n = 2$; aLRT = 0.99) (Supplementary Table S3). Five were from Mbuyi-Mayi, the remaining were from Kinshasa, Goma and one unknown location. Therefore, a total of fourteen C2 strains were detected in the DRC from 1997 to 2012. Taking into account all unique subtype C sequences from the DRC deposited in *Los Alamos* database and in our own laboratory database ($n = 278$), we found that sub-subtype C2 represent 5 per cent of the subtype C strains in the DRC.

3.1.2 *Unique recombinant form with a predominant CRF92_C2U backbone*

One sample collected in Mbuyi-Mayi in 2008 (strain 768) was similar to CRF92_C2U. However, it contains an additional unknown region (a 540-bp fragment between the C-terminal portion of the *pol* gene and half of the *vif gene*) and a fragment that resembles subtype A but different from all previously recognized A lineages (a 330-bp fragment covering the *vpu* gene and flanked regions, Supplementary Fig. S2). The strain is thus classified as URF (Fig. 1).

3.1.3 *CRF93_Cpx: a complex circulating mosaic containing divergent subtype C and a fragments*

Three sequences sampled from individuals with no epidemiological link shared a complex mosaic structure: strain 367 from Kinshasa and strains 653 and 817 from Mbuyi-Mayi. Different fragments from this mosaic were related to either subtype C or to A-like genetic forms and the majority of them are different from previously documented lineages (Fig. 1 and Supplementary Fig. S3). The first part of the *gag* gene resembled subtype A, although different from all known A sub-subtypes (i.e. A1 to A5, A-FSU and CRF02_AG). Two fragments in *gag* and *pol* regions were related to subtype C but they did also not correspond to the above-described C2; therefore, this region suggests a potential under-represented C3 lineage. Finally, the end of the *pol* gene and the region comprising the *vpr* gene and *gp*160 sequence were similar to CRF02_AG while a small part of *pol*, and most of *vif* and *nef* genes were similar to A5 (the sub-subtype of subtype A within CRF26_AU).

Extensive blast searches did not show any similarity of deposited HIV-1 strains with the same mosaic form.

3.2 Early lineage diversification of subtype C in the DRC

We investigated the evolutionary history of subtype C using a representative time-stamped dataset ($n = 92$) in different genomic regions (Although we referred to them as either *gag*, *pol* or *env*, we used the non-recombinant regions as delineated in the top-left panel of Fig. 2) and the partial p51-RT. Root-to-tip divergence analysis for each dataset are presented in Supplementary Fig. S5. The estimated mean evolutionary rates for *gag'*, *gag*, *pol* and *env* regions were, respectively, 1.87×10^{-3} (95% HPD (highest posterior density)$= 1.64 \times 10^{-3}$, 2.09×10^{-3}), 1.27×10^{-3} (95% HPD$= 1.03 \times 10^{-3}$ to 1.52×10^{-3}), 1.21×10^{-3} (95% HPD$= 1.10 \times 10^{-3}$ to 1.33×10^{-3}) and 2.45×10^{-3} (95% HPD$= 2.26 \times 10^{-3}$ to 2.65×10^{-3}) nucleotide substitutions per site per year. Table 1 summarizes the estimated time to the most recent common ancestor (TMRCA) using MCMC and least squares techniques; the mean difference between both approaches was 3.7 years (interquartile range, 2–5 years). The MCMC time-scaled evolutionary history from the conventional subtype C group ($n = 89$) clearly places the common ancestor of pandemic HIV-1 subtype C around 1949 (averaged over the four genomic regions estimates), matching that of Novitsky et al. (2010) (1950, 1928–1962). The inclusion of seven sub-subtype C2 sequences results in a TMRCA around 1932 (Fig. 2 and Supplementary Fig. S6, mustard-colored squares), whereas the inclusion of the more divergent subtype C sequences from the three mosaic sequences, CRF93_cpx, results in a TMRCA around 1928 (Fig. 2 and Supplementary Fig. S6, *pol* and partial-RT regions, brick red-colored squares). The TMRCA for the full tree (averaged over the full MCMC genomic regions estimates) was around 1910, overlapping with the estimates of Faria et al. (Faria et al. 2014) that also used Bayesian skygrid and two-phase exponential-logistic models.

3.3 Recombination of ancient and contemporary lineages

Figure 2 shows the detailed phylogenetic relationships of subtype C, sub-subtype C2, subtype A (including representatives of every A sub-subtype plus CRF26_AU) and CRF02_AG using different genomic regions ($n = 124$). Phylogenetic findings were in line with those obtained in preliminary reconstruction after bootscanning analyses. The bottom-right panels of Fig. 2 and Supplementary Fig. S6 ($n = 382$) corroborates that the envelope region of the mosaic strains represent a distinct but recent CRF02_AG variant with a TMRCA around the early 1970s. In addition, a recently reported full-length genome sequence from DRC (strain CG-0373-02V) grouped basal to conventional subtype C but not within sub-subtype C2 clade, representing a different divergent strain.

3.4 Extensive diversity of subtype C in DRC

Further phylogenetic analyses using a comprehensive partial p51-RT sequence dataset for subtype C from our surveillance studies in the DRC ($n = 160$) (Fig. 3, left panel) (Muwonga et al. 2011; Boillot et al. 2016) showed that strains 120109035 and VIR91 (this strain has an epidemiological link with VIR90) grouped within sub-subtype C2 as suggested above (Supplementary Table S3). This analysis also revealed that strain LBTB028 (accession AM041013) grouped with strain CG-0373-02, corroborating our previous observation of the later

Table 1. Estimated time to the most recent common ancestor.

Region	*Gag'*		*Gag*		*Pol*		*Pol* (Partial *p51-RT*)		*Env*	
	MCMC	LS	MCMC	LS	MCMC	LS	MCMC	LS	MCMC	LS
Ref. Subtype C	1946 (1938–1954)	1948 (1943–1954)	1949 (1937–1956)	1956 (1946–1962)	1952 (1947–1957)	1957 (1948–1964)	1953 (1947–1960)	1952 (1941–1960)	1950 (1944–1954)	1948 (1944–1953)
C + C2	1926 (1914–1935)	1921 (1916–1929)	1934 (1922–1946)	1936 (1924–1946)	1935 (1927–1942)	1941 (1929–1950)	1933 (1923–1944)	1932 (1918–1944)	1933 (1926–1940)	1932 (1926–1938)
C + C2 + divergent C[a]	…	…	1928 (1915–1940)	1921 (1911–1929)	1928 (1919–1936)	1933 (1919–1944)	1928 (1914–1938)	1924 (1907–1937)	…	…

MCMC and LS stand, respectively, for Markov chain Monte Carlo and least squares dating. Values in parentheses represent the 95 per cent HPDs.
[a]Subtype C fragments from CRF93_cpx.

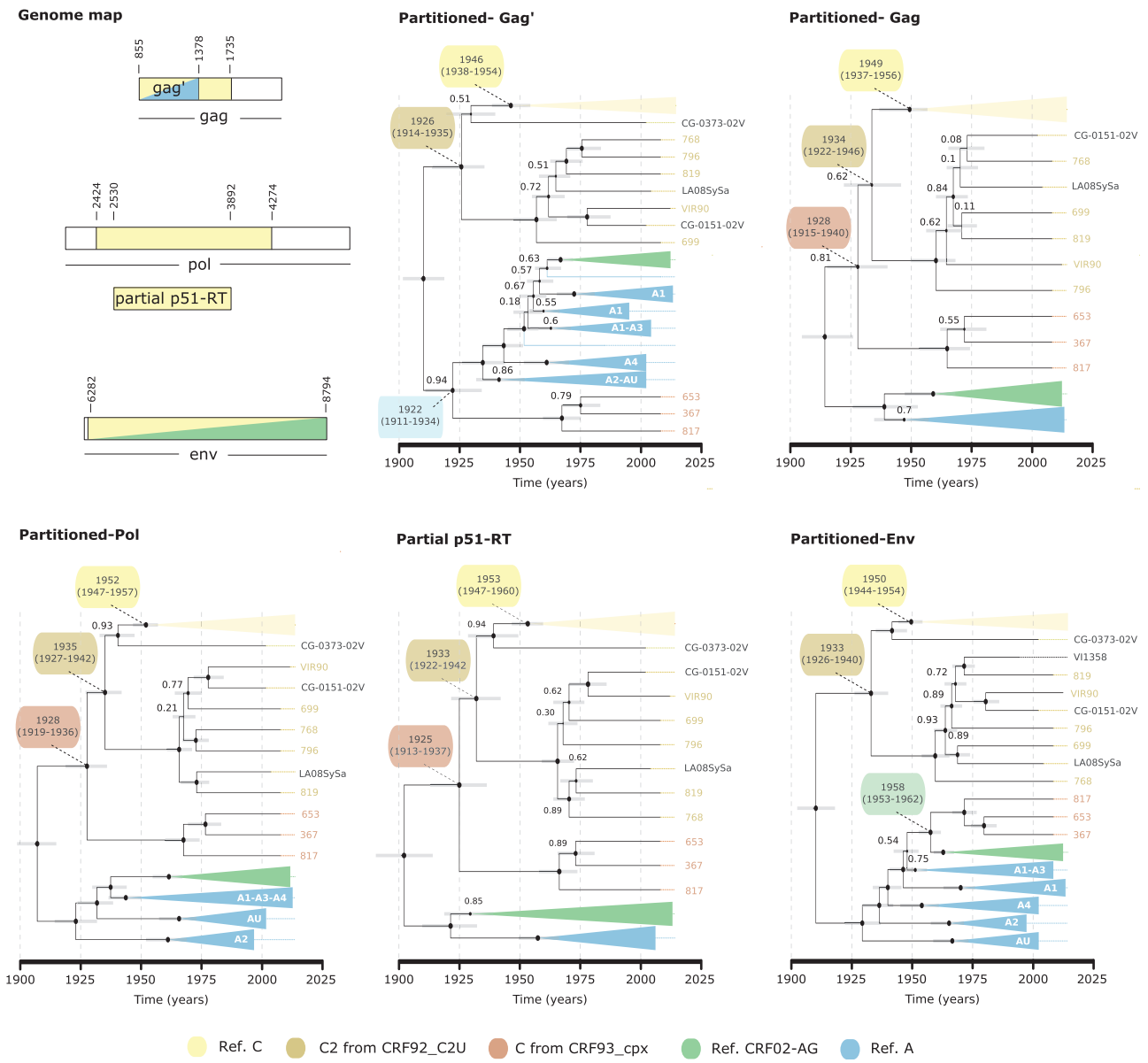

Figure 2. Maximum clade credibility trees for the newly obtained sequence data sampled from the DRC. The top-left panel outlines the non-recombinant sequence region (color coded) used for each gene. Posterior probability values are shown only for groups below 0.95. Distinctive sub-subtypes A and A5 (i.e. from CRF26_AU) are indicated. 95 per cent HPDs for dates are provided as light-gray bars.

representing an additional divergent lineage. Moreover, three strains (899, 120120048 and LBTB046, accessions MF372655, MF372654 and AM041004) grouped separately as a second additional divergent lineage. The distribution of these viruses suggests that all these lineages are potentially spreading across the country, for example in the city of Lubumbashi (region of Haut-Katanga) in the southeastern part of DRC, the second-largest metropolis in the country (Fig. 3, right panel). This analysis also slid the TMRCA of conventional subtype C from 1950 (1928–1962) to around 1942 (1934–1953), given that it includes many other minor genetic variants that also fell basal to subtype C.

A depiction of the relationships between the reference subtype C and subtype C from the DRC (including all identified divergent lineages) is presented in Fig. 4. Strains from DRC that resemble conventional subtype C are predominantly sampled and intermix with the subtype C reference strains sampled in other countries. Subtype C viral population dynamics showed a large increase in diversity starting around the late-fifties (Fig. 4. bottom panel). Using the two-phase exponential-logistic model of population growth, we determined that this transition to a faster phase of growth occurred around 1954 (1946–1962) and that the growth rates greatly increased from 0.06 per year (0.0004–0.12) to 0.25 per year (0.21–0.31). Although the demographic curves indicated a reduction in epidemic growth over the past decade, we underscore that these coalescent methods underestimate growth rates near the present (Hall et al. 2016). The best-fit parametric model and the overall findings were consistent—although with a much higher 95 per cent HPD range—when the analysis was restricted to sequences from the DRC only ($n = 155$) (Supplementary Table S2; growth transition

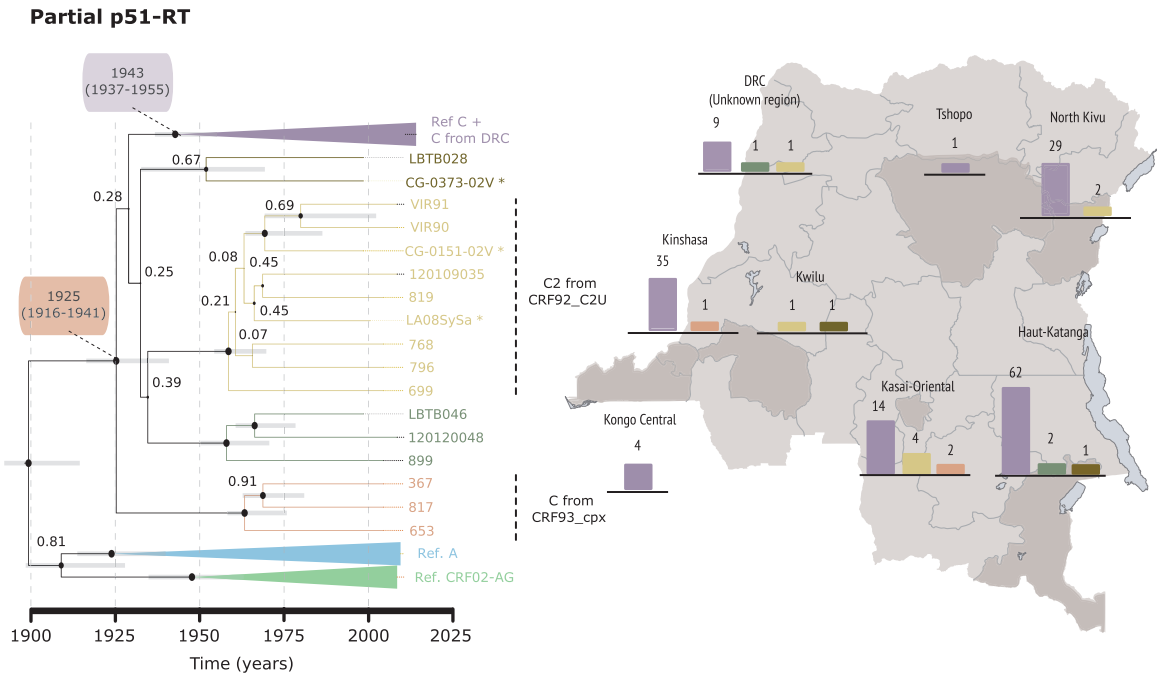

Figure 3. Maximum clade credibility tree using the partial p51-RT region and sampling region of strains from the DRC. Posterior probability values are shown only for groups below 0.95. Distinctive divergent subtype C lineages are colored (the number and size of the bars in the map denote the number or samples from the corresponding colored lineages). Asterisks indicate sequences available in the literature from other studies. 95 per cent HPDs for dates are provided as light-gray bars.

1957 95% HPD 1942–1973, exponential growth 0.06 95% HPD 1×10^{-5}-0.13, logistic growth 0.21 95% HPD 0.03–0.31).

4. Discussion

Given the unparalleled diversity of HIV-1/M strains circulating today, it is expected that a substantial diversity of basal strains exist in the DRC. During surveys on antiretroviral drug resistance in different locations from the DRC in 2008 and 2012, we identified several p51-RT gene sequences that clustered in the base of conventional subtype C clade. We characterized near-full length genomes of these lineages and identified two novel CRFs and one URF. Following the standards of HIV nomenclature, we named these novel lineages as CRF92_C2U and CRF93_cpx. However, we stress that while the CRF02_AG region of CRF93_cpx represents a different variant (TMRCA ~1971), the divergent backbones of both CRFs exemplify ancient diversity that evolved independently and persisted up to date.

Our findings allowed better documenting the early events of the HIV/AIDS epidemic by evidencing extensive diversity of subtype C in the DRC. This diversity already existed before some strain was exported out of the source region around 1950s and began a chain of infections. We argue that additional difficult-classifiable strains that fell basal to conventional groups will emerge for other subtypes, as sampling and sequencing efforts increase for some geographic regions of historical importance (as evidenced by the divergent A-like strains in the *gag*' region in Fig. 1 and Supplementary Fig. S6).

The observed phylogenetic pattern, i.e. the progressive loss of a clear distinction between subtypes the 'deeper' we go in the phylogenetic tree (Fig. 3), resulted from using enough sequences from the putative source of the AIDS epidemic and underscores the rapid evolution of HIV (Worobey, 2008). Historically, the chain of infections by exported strains resulted in distinctive clades that were firstly detected and led to the current HIV nomenclature. New data from the DRC are progressively filling the gaps between subtypes and underscore that HIV nomenclature does not provide an accurate picture of the historical epidemiology and/or evolution of the virus (Abecasis et al. 2007). This reflection is important because CRF92_C2U and CRF93_cpx are new in terms of discovery but their long internal branches reflect an old and independent evolutionary history. Thus, it may be the case that any of these divergent subtype C lineages resemble more an original "pure subtype" but by the standards of nomenclature we named them recombinants forms. Following this reasoning, pure subtypes may represent old recombinant variants that became predominant as discussed previously (Vidal et al. 2009).

Overall, our results illustrate that multiple divergent subtype C lineages that shared a common ancestor—as early as the mid-late 1920s—with conventional subtype C continue to circulate in the DRC, even though at a very low prevalence. We corroborated that a more rapid diversification started in the fifties, in line with previous estimates of HIV/M transition to a faster phase of growth around 1960 (Faria et al. 2014). In the same vein, the distinctive fragments that shared identity with A-like genetic forms clearly suggest that genetic variants related to other subtypes were also present in the DRC (Mir et al. 2016). Some of these lineages may have gone extinct but some still remain to be discovered. For example, the distinctive CRF02_AG region of CRF93_cpx may be related to a different lineage in the DRC (although CRF02_AG is not a widely prevalent genetic form in this country, previous analyses show that local strains fell basal to the conventional CRF02_AG groups; Mir et al. 2016). Even though it is challenging to discern the medical relevant aspects of novel genetic variants that are reminiscent of HIV

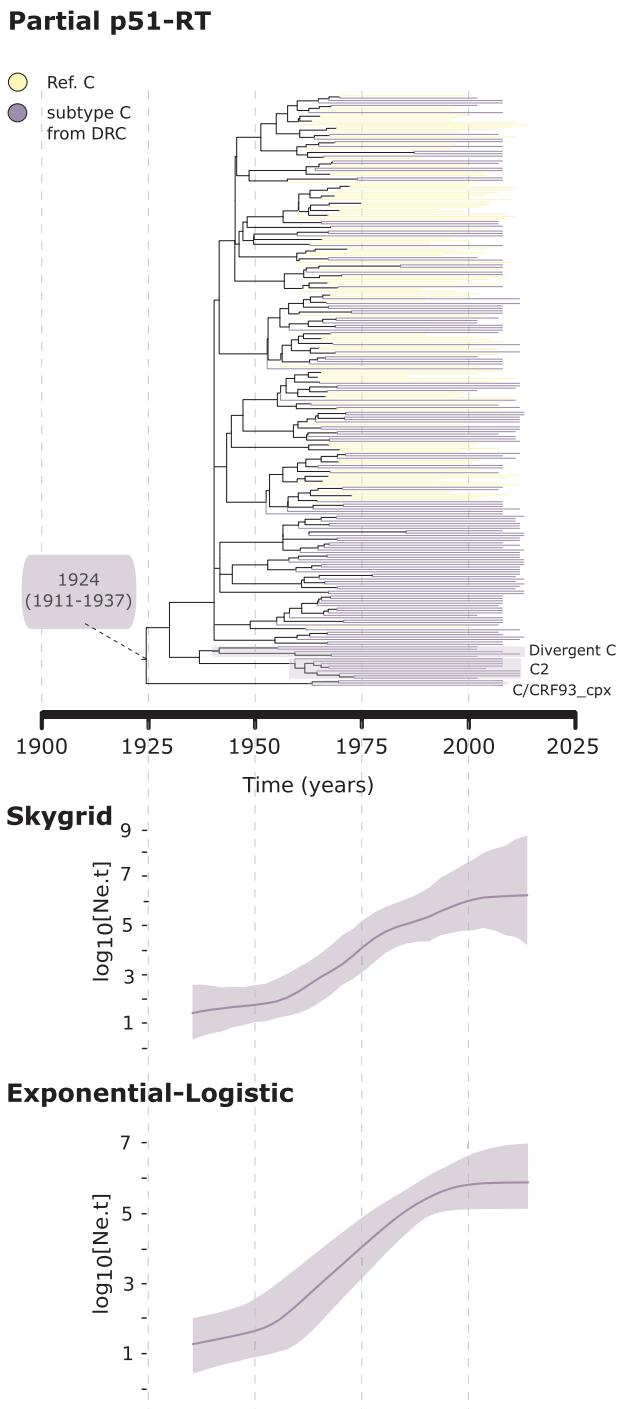

Figure 4. Maximum clade credibility tree of representative subtype C sequence data and corresponding demographic reconstructions. Posterior probability values and tips were removed for clarity. C2, the divergent C region from CRF93_cpx and additional divergent C lineages from Fig. 3 are highlighted. Colored areas in the demographic curves represent the 95 per cent HPDs.

early history, they have the potential to further spread (there is at least one instance of CRF92_C2U reaching Belgium in 1994) and recombine (i.e. the URF and CRF93_cpx).

Overall, although disentangling HIV's ancient history is challenging (Abecasis et al. 2007), we evidenced that the rapid evolution of HIV left behind a very diverse array of genetic lineages that continue to circulate—apparently in low levels—in the DRC. Our analysis showed considerable differences between the newly discovered early divergent strains and the conventional subtype C and therefore suggested that this virus has been diverging in humans for several decades before the HIV/M diversity boom in the 1950s.

Acknowledgements

The authors thank Samuel Edidi from Laboratoire National de Référence du SIDA (Kinshasa, Democratic Republic of Congo) for organizing the processing of the samples, Carlette Doufoundou-Guilengui and Stella Consenza for providing technical assistance during amplification and sequencing of the samples and Brian Foley from the Los Alamos HIV Database for helpful discussions.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Data availability

The new sequences have been deposited in GenBank.

References

Abecasis, A. B. et al. (2013) 'HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics', *Retrovirology*, 10/1: 7.

—— (2007) 'Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form', *J Virol*, 81/16: 8543–51.

Angelis, K. et al. (2015) 'Global dispersal pattern of HIV type 1 subtype CRF01_AE: a genetic trace of human mobility related to heterosexual sexual activities centralized in Southeast Asia', *The Journal of Infectious Diseases*, 211/11: 1735–44.

Balla-Jhagjhoorsingh, S. S. et al. (2011) 'Characterization of neutralizing profiles in HIV-1 infected patients from whom the HJ16, HGN194 and HK20 mAbs were obtained', *PloS One*, 6/10: e25488.

Bello, G. et al. (2008) 'Origin and evolutionary history of HIV-1 subtype C in Brazil', *Aids*, 22/15: 1993–2000.

Berg, M. G. et al. (2016) 'A Pan-HIV strategy for complete genome sequencing', *Journal of Clinical Microbiology*, 54/4: 868–82.

Boillot, F. et al. (2016) 'Programmatic feasibility of dried blood spots for the virological follow-up of patients on antiretroviral treatment in Nord Kivu, Democratic Republic of the Congo', *Jaids-Journal of Acquired Immune Deficiency Syndromes*, 71/1: E9–E15.

Brenner, B. et al. (2003) 'A V106M mutation in HIV-1 clade C viruses exposed to efavirenz confers cross-resistance to non-nucleoside reverse transcriptase inhibitors', *Aids*, 17/1: F1–5.

Brenner, B. G. et al. (2006) 'HIV-1 subtype C viruses rapidly develop K65R resistance to tenofovir in cell culture', *Aids*, 20/9: F9–13.

Castresana, J. (2000) 'Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis', *Molecular Biology and Evolution*, 17/4: 540–52

Dalai, S. C. et al. (2009) 'Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe', *Aids*, 23/18: 2523–32.

Delatorre, E. O., and Bello, G. (2012) 'Phylodynamics of HIV-1 subtype C epidemic in east Africa', *PloS One*, 7/7: e41904.

Djoko, C. F. et al. (2011) 'High HIV type 1 group M pol diversity and low rate of antiretroviral resistance mutations among the uniformed services in Kinshasa, Democratic Republic of the Congo', *AIDS Research and Human Retroviruses*, 27/3: 323–9.

Drummond, A. J. et al. (2012) 'Bayesian phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29/8: 1969–73.

—— (2006) 'Relaxed phylogenetics and dating with confidence', *PLoS Biol*, 4/5: e88.

Edgar, R. C. (2004a) 'MUSCLE: a multiple sequence alignment method with reduced time and space complexity', *BMC Bioinformatics*, 5/1: 113.

—— (2004b) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Res*, 32/5: 1792–7.

Faria, N. R. et al. (2014) 'HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations', *Science*, 346/6205: 56–61.

Gill, M. S. et al. (2013) 'Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci', *Molecular Biology and Evolution*, 30/3: 713–24.

Guindon, S. et al. (2010) 'New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0', *Systematic Biology*, 59/3: 307–21.

Hall, M. D., Woolhouse, M. E., and Rambaut, A. (2016) 'The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: a simulation study', *Virus Evolution*, 2/1: vew003.

Hemelaar, J., Isolation W-UNfH, Characterisation. et al. (2011) 'Global trends in molecular epidemiology of HIV-1 during 2000-2007', *Aids*, 25/5: 679–89.

Huang, A. et al. (2016) 'Global Comparison of Drug Resistance Mutations After First-Line Antiretroviral Therapy Across Human Immunodeficiency Virus-1 Subtypes', *Open Forum Infectious Diseases*, 3/2: ofv158.

Katoh, K., and Frith, M. C. (2012) 'Adding unaligned sequences into an existing alignment using MAFFT and LAST', *Bioinformatics*, 28/23: 3144–6.

——, and —— and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution*, 30/4: 772–80.

Kouri, V. et al. (2015) 'CRF19_cpx is an Evolutionary fit HIV-1 Variant Strongly Associated With Rapid Progression to AIDS in Cuba', *EBioMedicine*, 2/3: 244–54.

Lau, K. A., and Wong, J. J. (2013) 'Current Trends of HIV Recombination Worldwide', *Infectious Disease Reports*, 5/Suppl 1: e4).

Loemba, H. et al. (2002) 'Genetic divergence of human immunodeficiency virus type 1 Ethiopian clade C reverse transcriptase (RT) and rapid development of resistance against nonnucleoside inhibitors of RT', *Antimicrobial Agents and Chemotherapy*, 46/7: 2087–94.

Lole, K. S. et al. (1999) 'Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination', *J Virol*, 73/1: 152–60.

Lu, X. et al. (2016) 'Recombinant Patterns of Nine Novel HIV-1 Recombinant Strains Identified in Hebei Province, China', *AIDS Research and Human Retroviruses*, 32/5: 475–9.

Luo, C. C. et al. (1995) 'HIV-1 subtype C in China', *The Lancet*, 345/8956: 1051–2.

Maddison, W. P., and Maddison, D. R. Mesquite: a modular system for evolutionary analysis. Version 0.991. 2002. <http://mesquiteproject.org> accessed Feb 2017.

Magiorkinis, G. et al. (2016) 'The global spread of HIV-1 subtype B epidemic', *Infection, Genetics and Evolution: journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 46: 169–79.

Mir, D. et al. (2016) 'Phylodynamics of the major HIV-1 CRF02_AG African lineages and its global dissemination', *Infection, Genetics and Evolution: journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 46: 190–9.

Mokili, J. L. et al. (2002) 'Identification of a novel clade of human immunodeficiency virus type 1 in Democratic Republic of Congo', *AIDS Research and Human Retroviruses*, 18/11: 817–23.

Muwonga, J. et al. (2011) 'Resistance to antiretroviral drugs in treated and drug-naive patients in the Democratic Republic of Congo', *Journal of Acquired Immune Deficiency Syndromes*, 57 Suppl 1: S27–33.

Neogi, U. et al. (2012) 'Molecular epidemiology of HIV-1 subtypes in India: origin and evolutionary history of the predominant subtype C', *PloS One*, 7/6: e39819.

Niama, F. R. et al. (2006) 'HIV-1 subtypes and recombinants in the Republic of Congo', *Infection, Genetics and Evolution: journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 6/5: 337–43.

Njai, H. F. et al. (2006) 'The predominance of Human Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness', *Retrovirology*, 3: 40.

Oster, A. M. et al. (2017) 'Increasing HIV-1 subtype diversity in seven states, United States, 2006-2013', *Annals of Epidemiology*.

Pineda-Pena, A. C. et al. (2016) 'On the contribution of Angola to the initial spread of HIV-1', *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 46: 219–22.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009) 'FastTree: computing large minimum evolution trees with profiles instead of a distance matrix', *Molecular Biology and Evolution*, 26/7: 1641–50.

——, ——, and —— et al. (2010) 'FastTree 2–approximately maximum-likelihood trees for large alignments', *PloS One*, 5/3: e9490.

Rambaut, A. et al. (2016) 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2/1: vew007.

—— Tracer v1.6 2014. <http://beast.bio.ed.ac.uk/Tracer> accessed Feb 2017.

Robertson, D. L. et al. (2000) 'HIV-1 Nomenclature Proposal', *Science*, 288/5463: 55d.

Rodgers, M. A. et al. (2017) 'Sensitive next-generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo', *Journal of Virology*, 91/6: e01841-16.

Skhosana, L. et al. (2015) 'High prevalence of the K65R mutation in HIV-1 subtype C infected patients failing tenofovir-based first-line regimens in South Africa', *PloS One*, 10/2: e0118145.

Suchard, M. A., and Rambaut, A. (2009) 'Many-core algorithms for statistical phylogenetics', *Bioinformatics*, 25/11: 1370–6.

Talavera, G., and Castresana, J. (2007) 'Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments', *Systematic Biology*, 56/4: 564–77.

To, T. H. et al. (2016) 'Fast dating using least-squares criteria and algorithms', *Systematic Biology*, 65/1: 82–97.

Turk, G., and Carobene, M. G. (2015) 'Deciphering How HIV-1 Intersubtype Recombination Shapes Viral Fitness and Disease Progression', *EBioMedicine*, 2/2: 188–9.

Vergne, L. et al. (2000) 'Genetic diversity of protease and reverse transcriptase sequences in non-subtype-B human immunodeficiency virus type 1 strains: evidence of many minor drug resistance mutations in treatment-naive patients', *Journal of Clinical Microbiology*, 38/11: 3919–25.

Vidal, N. et al. (2009) 'Genetic characterization of eight full-length HIV type 1 genomes from the Democratic Republic of Congo (DRC) reveal a new subsubtype, A5, in the A radiation that predominates in the recombinant structure of CRF26_A5U', *AIDS Research and Human Retroviruses*, 25/8: 823–32.

—— (2006) 'HIV type 1 pol gene diversity and antiretroviral drug resistance mutations in the Democratic Republic of Congo (DRC)', *AIDS Research and Human Retroviruses*, 22/2: 202–6.

—— (2000) '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', *Journal of Virology*, 74/22: 10498–507.

Wertheim, J. O. et al. (2014) 'The global transmission network of HIV-1', *The Journal of Infectious Diseases*, 209/2: 304–13.

Wilkinson, E., Engelbrecht, S., and de Oliveira, T. (2015) 'History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region', *Scientific Reports*, 5/1: 16897.

——, ——, and —— et al. (2016) 'Origin, imports and exports of HIV-1 subtype C in South Africa: A historical perspective', *Infection, Genetics and Evolution: journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 46: 200–8.

Worobey, M. (2008) 'The origins and diversification of HIV', in Paul Volberding, Warner Greene, Joep Lange, Joel Gallant, Nelson Sewankambo (eds.) *Global HIV/AIDS Medicine*, Philadelphia, PA: Saunders Elsevier, p. 830.

—— et al. (2008) 'Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960', *Nature*, 455/7213: 661–4.

Zhang, M. et al. (2010) 'The role of recombination in the emergence of a complex and dynamic HIV epidemic', *Retrovirology*, 7: 25.