Virus Evolution, 2018, 4(1): vey003

doi: 10.1093/ve/vey003
Research article

Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages

Marcel Tongo,[1,2,3,*,†] Gordon W. Harkins,[4] Jeffrey R. Dorfman,[5,6,‡] Erik Billings,[7,8] Sodsai Tovanabutra,[7,8] Tulio de Oliveira,[1] and Darren P. Martin[2,*,§]

[1]KwaZulu-Natal Research Innovation and Sequencing Platform (Krisp), School of Laboratory Medicine and Medical Sciences, College of Health Sciences, Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban 4041, South Africa, [2]Division of Computational Biology, Department of Integrative Biomedical Sciences and Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town 7925, South Africa, [3]Center of Research for Emerging and Re-Emerging Diseases (CREMER), Institute of Medical Research and Study of Medicinal Plants (IMPM), Yaoundé, Cameroon, [4]South African MRC Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Bellville 7535, South Africa, [5]Division of Immunology, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town 7925, South Africa, [6]Division of Immunology, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 2193, South Africa, [7]U.S. Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, MD 20910–7500, USA and [8]Henry M. Jackson Foundation for the Advancement of Military Medicine Inc., Bethesda, MD 20910–7500, USA

*Corresponding authors: E-mail: marcel.tongo@gmail.com (M.T.); E-mail: darrenpatrickmartin@gmail.com (D.P.M.)

[†]http://orcid.org/0000-0002-5262-892X

[‡]http://orcid.org/0000-0001-9938-8911

[§]http://orcid.org/0000-0002-8785-0870

Abstract

Subtype A is one of the rare HIV-1 group M (HIV-1M) lineages that is both widely distributed throughout the world and persists at high frequencies in the Congo Basin (CB), the site where HIV-1M likely originated. This, together with its high degree of diversity suggests that subtype A is amongst the fittest HIV-1M lineages. Here we use a comprehensive set of published near full-length subtype A sequences and A-derived genome fragments from both circulating and unique recombinant forms (CRFs/URFs) to obtain some insights into how frequently these lineages have independently seeded HIV-1M sub-epidemics in different parts of the world. We do this by inferring when and where the major subtype A lineages and subtype A-derived CRFs originated. Following its origin in the CB during the 1940s, we track the diversification and recombination history of subtype A sequences before and during its dissemination throughout much of the world between the 1950s and 1970s. Collectively, the timings and numbers of detectable subtype A recombination and dissemination events, the present broad global distribution of the sub-epidemics that were seeded by these events, and the high prevalence of subtype A sequences within the regions where these sub-epidemics occurred, suggest that ancestral subtype A viruses

(and particularly sub-subtype A1 ancestral viruses) may have been genetically predisposed to become major components of the present epidemic.

Key words: HIV-1 subtype A; phylogenetic analysis; recombination; diversity; adaptation; evolution

1. Introduction

One of the most epidemiologically and evolutionarily important characteristics of HIV-1 group M (HIV 1M) is its extremely rapid rate of evolution. The explosive diversification of HIV-1M over the past century since it first started infecting humans has been primarily fuelled by the ~0.3 mutation events (Rambaut et al. 2004) and ~5 recombination events (Jetzt et al. 2000; Rhodes, Wargo, and Hu 2003) that occur during every viral replication cycle.

Whereas many prospective cohort studies have suggested that there may be differences in rates of disease progression between subtypes (Taylor et al. 2008; Kiguoya et al. 2017), experimental efforts to detect differences in the fitness of viruses belonging to the different subtypes have primarily focused on quantifying individual gene functions. For example, whereas the *nef* genes of subtype B isolates seem to downregulate CD4 and class I HLA allele expression more effectively than do the *nef* genes from subtype A, C, and D isolates (Mann et al. 2013), the *vpu* genes from subtype C isolates tend to downregulate CD4 and tetherin more effectively than do the *vpu* genes of subtypes A, B, and A/D recombinant isolates (Rahimi et al. 2017). It is, however, very difficult to infer differences in the actual epidemiological potential (i.e. the overall fitness, reproductive rate, or pathogenesis) of viruses belonging to different HIV-1M subtypes based entirely on these types of directed fitness assays (Markle, Philip, and Brockman 2013). The relative fitness of different HIV-1M subtypes in an epidemiological sense should, at least in part, be reflected in their relative global prevalence and/or distributions over the course of the epidemic.

Subtype A was likely one of the most epidemiologically competent HIV-1M proto-subtypes circulating in the Congo Basin (CB) region during the 1950s (Worobey et al. 2008). The possibility that subtype A and its recombinant descendants (primarily including CRF02_AG) remain among the fittest HIV-1M lineages, is best supported by the fact that, unlike any of the other major HIV-1M subtypes, these viruses are still widely distributed and highly prevalent throughout equatorial west Africa; the region where HIV-1M initially emerged. Subtype A and its recombinant, CRF02_AG, account for ~43% of HIV-1M infections in the Democratic Republic of Congo (DRC), ~50% of HIV-1M infections in Cameroon, ~53.7% in Equatorial Guinea, 46.7% in Gabon, ~30% in the Republic of Congo, and ~22% in Angola (Niama et al. 2006; Bartolo et al. 2009; Djoko et al. 2010; Caron et al. 2012; Tongo et al. 2013; Rodgers et al. 2017b). While the possibility cannot be excluded that this distribution is just the result of subtype A or CRF02_AG having spread throughout this entire region by chance (i.e. without the need to invoke increased fitness), one should at least question how these two lineages have persisted as the predominant HIV lineages within the CB: a region where the potential for competition between these viruses and viruses belonging to almost all the other known HIV-1M subtypes is very high. Further, throughout the rest of the world, subtype A and its recombinant descendants account for ~25% of HIV-1M infections and are second only to subtype C in terms of global prevalence (Hemelaar et al. 2011). These viruses are present in east Africa, central and south-east Asia, western and Eastern Europe, and South America (Tebit and Arts 2011).

The viruses included within subtype A are in fact so diverse that they have been classified into six different sub-subtypes (referred to as A1 through A6) (Foley et al. 2016). As more evidence accumulates, additional sub-subtypes will likely need to be defined to keep track of the diversity within this subtype. For example, the existence of an A7 sub-subtype has been inferred through analyses of Angolan HIV-1M *env* and *pol* gene sequences (Bartolo et al. 2009) and an A8 sub-subtype has been inferred from the characterization of subtype A-derived sequences found within circulating recombinant forms (CRFs) 36_cpx and 37_cpx (Powell et al. 2007a,b).

Here, we gather for the first time a comprehensive dataset of subtype A full genome sequences together with fragments of subtype A-derived sequences from CRFs and unique recombinant forms (URFs) to phylogenetically infer when and where major recombination and dissemination events occurred during the evolutionary history of subtype A and subtypes A-derived recombinant viruses. We further assess whether there is evidence of subtle differences in selection pressures at individual codon sites within subtype A sequences sampled from different sub-epidemics.

2. Materials and methods

2.1 Selection of sequences

To study the evolutionary history of HIV subtype A lineages, we retrieved (1) 136 near full-length sequences classified as belonging to subtype A within the Los Alamos National Laboratory (LANL) HIV sequence database (LANL 2014) in June 2015; (2) 120 sequences from 21 CRFs containing A-attributed segments with a total length of at least 1,000 nucleotides (nt) according to the breakpoint locations defined by the LANL HIV sequence database; (3) two highly divergent URFs; and (4) ninety nine representatives of near full-length sequences from each of the other eight 'pure' HIV-1M subtypes that were available in the LANL database in June 2015 (LANL 2014). These representative sequences were specifically selected to include the broadest diversity of HIV-1M near full-length genomes previously identified as belonging to known HIV-1M subtypes (Tongo et al. 2015a). This selection procedure was also applied to CRFs 01_AE and 02_AG lineages because of the large numbers of sequences available for these two CRFs. This was achieved by constructing maximum likelihood (ML) trees from all available full-length sequences for each of the two CRFs using Fastree2 (Price, Dehal, and Arkin 2009) implemented in RDP4 (Martin et al. 2015), and selecting one sequence from each of the most basal lineages from the root of these CRFs (Tongo et al. 2015a). This yielded thirteen CRF01_AE and twelve CRF02_AG sequences that represent the entire breadth of genetic diversity within these clades (Supplementary Table S1).

2.2 Phylogenetic analysis

Full-length genome sequences were aligned using MUSCLE (Edgar 2004) with manual editing in IMPALE (Khoosal and Martin). Recombinant sequences were removed from the

resulting alignments and split into their constituent recombinationally derived fragments based on previously inferred breakpoint locations. Individual subtype A-derived fragments larger than 1,000 nt in length were then re-added to the initial alignment, with gap characters being added to the 3′ and 5′ ends of the fragments to ensure that they remained correctly aligned with the rest of the dataset. Genome segments corresponding to HXB2 nt positions 4267–5041 of CRF06_cpx, 4040–6046 of CRF11_cpx, and 8790–9442 of CRF18_cpx that were previously identified as being subtype A derived were not included in this alignment because the actual origins of these sequences are uncertain (Tongo, Dorfman, and Martin 2015b). A mid-point-rooted ML phylogenetic tree was constructed from the alignment with 1,000 full ML bootstrap replicates using RAxML version 8 (Stamatakis 2014) implemented in CIPRES (Miller, Pfeiffer, and Schwartz 2010). Although RAxML is limited to the use of GTR-based nucleotide substitution models (GTR-GAMMA in our case), it has been specifically designed to accurately infer phylogenies from alignments containing large amounts of missing data (Stamatakis and Alachiotis 2010; Stamatakis 2014). This ideally suited it to the analysis of our mostly inter-subtype recombination-free alignment.

2.3 Determining the times when, and locations where, ancestral sequences existed

Estimation of the time to the most recent common ancestors (tMRCAs) of various sub-lineages represented within our subtype A dataset was achieved using a Bayesian statistical approach implemented in the software, BEAST v1.8.4 (Drummond and Rambaut 2007). To ensure that the subtype A and subtype A-attributed genome fragment dataset contained sufficient phylogenetic signal to produce accurate and precise estimates of the dates of the nodes of a maximum clade credibility (MCC) tree determined using BEAST, we randomized dates of the sequences during the Markov chain Monte Carlo (MCMC) sampling procedure. This allowed the direct comparison of results obtained from the analyses of sequences with correct sampling times with those of these same sequences with randomized sampling times (Trovao et al. 2015). Significant temporal signal was inferred if no overlap existed between the 95% highest posterior density (HPD) credibility intervals for the nucleotide substitution rate estimates obtained from the real and randomized MCMC analyses. The best-fit nucleotide substitution model was estimated using J model test (Guindon and Gascuel 2003; Darriba et al. 2012) and the evolutionary model fit (clock model, coalescent demographic prior) was evaluated using marginal likelihood estimates (MLEs) obtained through the path sampling and stepping-stone sampling methods as implemented in BEAST (Baele et al. 2012). The analysed subtype A sequences were clustered into ten discrete location state categories (Supplementary Table S2).

To ensure convergence and sufficient mixing of the Markov chains for each of the evolutionary models, ten independent MCMC runs of 400 million generations with trees being sampled every 40,000 generations were performed until effective sample sizes for all model parameters were >200. When similar results were obtained from independent runs of the same model, log and tree files were combined using LogCombiner and the MCC trees constructed using TreeAnnotator (both part of the BEAST package). The resulting log and tree output files were visualised using Tracer v1.6 (Rambaut et al. 2014) and FigTree v1.4 (Rambaut and Drummond 2014).

We estimated the times when recombination events most likely occurred by first identifying, for each of the individual recombinationally acquired subtype A segments within each CRF, the node in the MCC tree that represented the most recent common ancestor (MRCA) of these segments. The inferred date and the upper bound of the 95% HPD for this node were respectively taken as being the upper bounds and upper 95% HPD bound of the date when the recombination event occurred. Similarly, the dates and lower bound of the 95% HPD for the node immediately ancestral to this MRCA node in the MCC tree were respectively taken as being the lower bound and lower 95% HPD bound of the date when the recombination event occurred (Tee et al. 2009). Similarly, the geographical location where the recombination event likely occurred was inferred from the average of the location state probabilities inferred for the two nodes bounding the branch along which the recombination event most likely occurred.

2.4 Analysis of natural selection patterns within codon alignments

We detected and compared signals of negative selection (natural selection disfavouring change) and positive selection (natural selection favouring change) within codon alignments of five sets of sequences drawn from the A1 sub-tree. These included: (1) a monophyletic clade of A1 strains circulating in Africa ($A1_{Afr}$), (2) a monophyletic clade of A1 strains circulating in Europe ($A1_{Eur}$), (3) a monophyletic clade of A1-derived sequences within CRF01_AE, (4) monophyletic A1-derived sequences within CRF02_AG, and (5) a monophyletic clade of A1-derived sequences within CRF22_01A1. These groups of sequences were selected because they display a degree of diversity that is sufficient to enable the quantification of relative synonymous (dS) and non-synonymous (dN) substitution rates at individual codon sites. For these analyses, all CRF 01_AE and 02_AG sequences currently available in the LANL HIV sequence database (LANL 2014) were retrieved (all of these sequences were not included in the original spatio-temporal sequence analysis). Genome fragments corresponding to the *gag*, *pol*, and *env* genes were codon aligned together with homologous sequences from $A1_{Eur}$, $A1_{Afr}$, and CRF22_01A1 using MUSCLE (Edgar 2004) and were manually edited in MEGA version 5 (Tamura et al. 2011). The alignments were segmented so that they only contained subtype A-derived genome fragments. The FUBAR method (Murrell et al. 2013) was used to estimate dN-dS scores for individual codons within these alignments, but only when these contained nucleotides that are most likely expressed in just one frame. The following pairwise comparisons and visualizations using SelectionMap (Stenzel et al. 2014) of selection patterns were made across genes drawn from the five different phylogenetically independent groups of viral sequences: $A1_{Afr}$ vs $A1_{Eur}$, $A1_{Afr}$ vs CRF01_AE, $A1_{Afr}$ vs CRF02_AG; and $A1_{Afr}$ vs CRF22_01 A.

3. Results

3.1 Phylogeny of HIV-1M subtype A lineages

We began our investigation into the evolutionary fitness of subtype A by investigating the evolutionary relationships between HIV-1M subtype A viruses and fragments of subtype A sequence found within recombinant HIV-1M genomes. We assembled a dataset containing 136 near full-length subtype A sequences and 120 contiguous subtype A-derived genome segments from

twenty-one CRF sequences and two URF sequences (each with a least one subtype A-derived genomic fragment greater than 1,000 nucleotides long). To phylogenetically contextualize the subtype A and subtype A-derived sequences, we included an additional ninety nine near full-length sequences representing the full spectrum of known HIV-1M diversity within subtypes B, C, F, G, H, J, and K.

The ML tree generated from this dataset revealed the presence of at least four major evolutionary sub-lineages within subtype A (A1, A2, A4, and A5), all of which form sub-trees with more than 90% bootstrap support (Fig. 1 and Supplementary Fig. S1). The first sub-lineage comprises almost all the subtype A near full-length sequences previously characterized as belonging to sub-subtype A1 (Fig. 1). This sub-lineage is highly diverse and contains a further three distinct groups. The first of these (cluster I) consists primarily of A1 viruses from Africa but also includes a small cluster of sequences sampled in Cyprus, which was likely founded by viruses disseminated directly from Africa. This lineage also includes subtype A-derived genome segments from CRFs 35_AD and 50_A1D (Fig. 1). The second major sequence lineage (cluster II) within A1 consists primarily of sequences sampled in Europe but also contains subtype A-attributed segments from CRFs 03_AB and 32_06A1 (Fig. 1). The third lineage includes two groups of highly divergent viruses (clusters IIIa and IIIb), most of which are the subtype A parental lineages of fourteen of the twenty-one CRFs analysed here (01_AE, 02_AG, 04_cpx, 06_cpx, 09_cpx, 11_cpx, 13_cpx, 18_cpx, 19_cpx, 22_01A1, 36_cpx, and 37_cpx, 45_cpx and 49_cpx). This cluster also includes viruses sampled in Senegal that have previously been identified as belonging to sub-subtype A3 (Fig. 1 and Supplementary Fig. S1).

It is likely that parental viruses of the A-attributed portions of the numerous CRFs were co-circulating in the CB and could have had the opportunity to recombine among themselves. To test this possibility, we generated two alignments corresponding to genome regions equivalent to the HXB2 genome coordinates 790–2155 and 3275–4174 of CRF02_AG; these are the only genome regions where both the 01_AE and 02_AG sequences (representing the two main lineages among the analysed CRFs) contained overlapping subtype A-attributed genome fragments. These alignments included not only the A-attributed fragments of these two CRFs, but also fragments from the 'pure' subtype A lineages. These two datasets were assessed for evidence of intra-subtype recombination events using RDP4 (Martin et al. 2015), which does not require prior identification of non-recombinant parental sequences and considers all sequences in the analysis as potential recombinant and/or parental sequences. These analyses failed to detect any convincing recombination signals within any of the analysed sequences (Supplementary data and rdp files are available on request).

Even if no evidence of intra-subtype recombination was detected in the subtype A-attributed genome fragments of CRFs 01_AE and 02_AG, there is still a possibility that A-attributed fragments from other CRFs included in our dataset were the result of intra-subtype A recombination. Therefore, to assess if this possibility could have influenced our phylogenetic tree, we constructed a new ML tree with a dataset that included only subtype A full-length sequences classified in the LANL database as A1–A4 and A6 plus representative sequences of the other eight 'pure' subtypes. The tree was constructed using RAxML (Stamatakis 2014) implemented in CIPRES (Miller, Pfeiffer, and Schwartz 2010). The distribution and classification of subtype A sequences in this tree was similar to that in Fig. 1, with sub-subtype A3 clustering within the A1 radiation (Supplementary Fig. S2).

The second major evolutionary sub-lineage within subtype A, A2, contains isolates previously described as sub-subtype A2 and includes subtype A-derived genome segments within CRFs 16_A2D and 21_A2D. Sub-lineage three, A4, contains all Congolese isolates previously identified as belonging to sub-subtype A4 and the final sub-lineage, A5, contains A-attributed genome segments from DRC CRF26_AU isolates that have been previously identified as being derived from a sub-subtype A5 parental virus (Fig. 1). While the A-attributed fragment in the URF, Z321, branches basal to sequences belonging to CRF13_cpx, the two A-segments from the 97CD1997.KMST91 genome have apparently been derived from divergent parental viruses that branch phylogenetically near the base of the subtype A sub-tree (Fig. 1), suggesting that subtype A might contain a variety of currently unsampled and/or undersampled divergent lineages; lineages which might even be still circulating at low frequencies.

3.2 Evidence of multiple different subtype A parents in some CRFs

It has been found that A-attributed segments of CRFs 02_AG appear to have been derived from different subtype A parental viruses (Zhang et al. 2010). To further investigate whether multiple subtype A parental viruses might have contributed to individual CRFs, we separately analysed each of the >1,000 nt long A-attributed segments from the CRFs in cluster III in our dataset. Contiguous subtype A-attributed fragments within each of these CRFs were separated into new sequences (i.e. the original CRF genomes containing multiple separated tracts of subtype A sequence were split into multiple different sequences) with gap characters being added to the 3′ and 5′ ends of the fragments to ensure that they remained correctly aligned with the remainder of the dataset. Fragments smaller than ~1,000 nt in length were removed from the alignment. In this regard, two fragments of CRF02_AG (HXB2 nt position 790–2155 and 6225–8311), two from CRF09_cpx (790–2155 and 3275–4175), two from CRF22_01A1 (2666–5452 and 6724–8470), three from CRF37_cpx (1126–2142, 2913–4663, and 6431–7743), and three from CRF45_cpx (790–2397, 4299–6041, and 6343–8236) were analysed separately. This yielded a dataset of 306 subtype A and subtype A-derived sequences. A MCC tree was then constructed with this dataset.

The two A-attributed fragments of CRF09_cpx cluster together within the same region of the phylogenetic tree indicating that it is plausible that they could have originated from the same parental virus (Fig. 2 and Supplementary Fig. S3). Similarly, the A-attributed fragments of CRF22_01A1 also cluster in the same sub-tree, also indicating a same origin. In contrast, the two analysed segments of CRF02_AG do not cluster within the same sub-tree and, as has been suggested elsewhere (Zhang et al. 2010), could plausibly have originated from two different parental sources. Similarly, the three A-attributed fragments of CRF37_cpx and CRF45_cpx that were analysed, possibly also originated from two different sources (Fig. 2 and Supplementary Fig. S3).

It is therefore apparent that some of the analysed CRFs likely have multiple different subtype A parental viruses, suggesting that multiple recombination events, possibly occurring in multiple different infected individuals, were involved in the assembly of these genomes.

3.3 Identifying when and where recombinants arose

To estimate the approximate dates when subtype A-derived sequences were incorporated within the recombinant genomes,

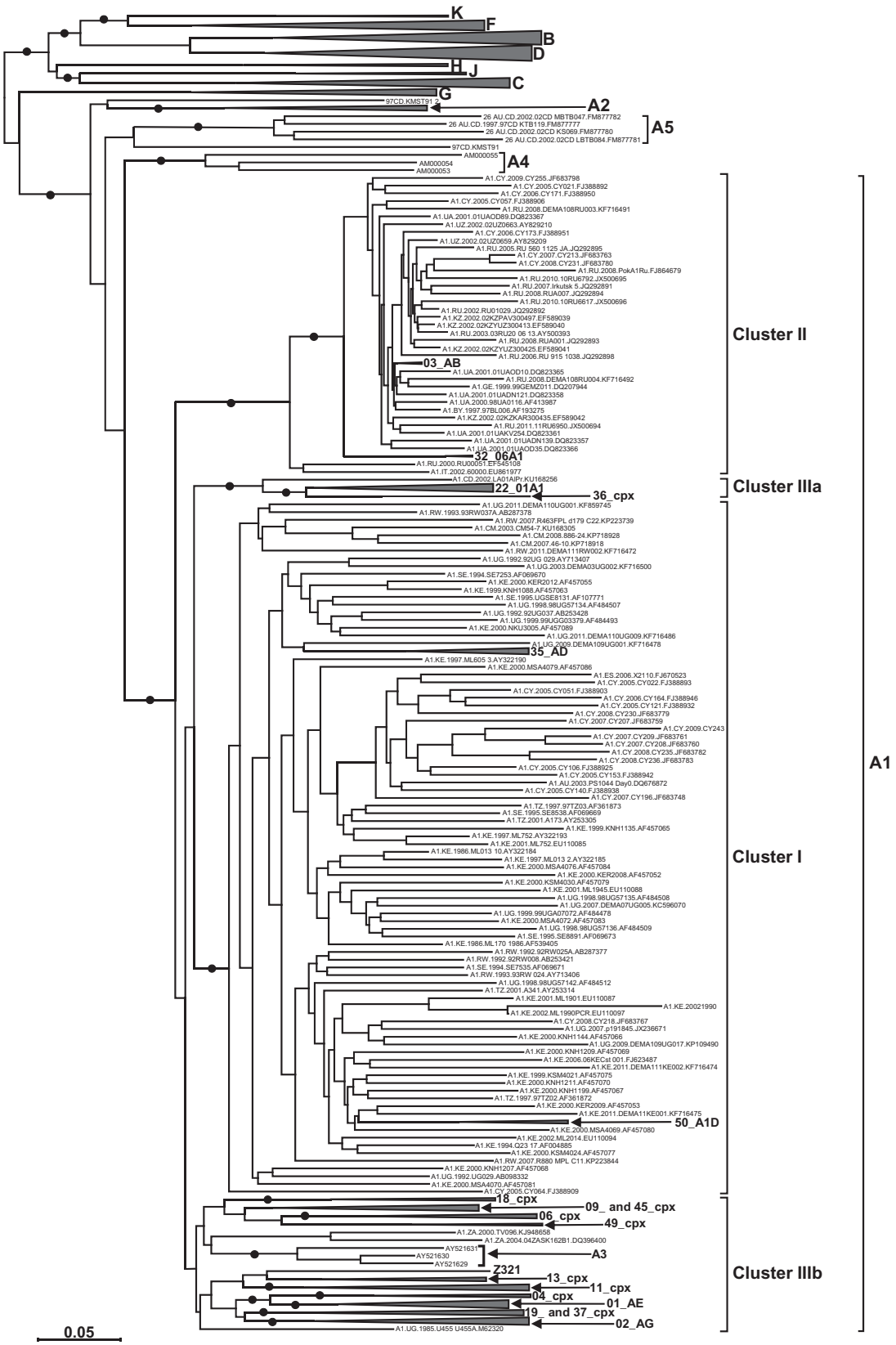

Figure 1. ML tree indicating the phylogenetic relationships between 357 HIV-1M genome sequences. These represent all published near full-length subtype A sequences that were available in the LANL database in June 2015 (LANL 2014), contiguous subtype A-derived genome segments from twenty-one CRFs with a total length of at least 1,000 nt, three highly divergent subtype A-like sequence fragments from two URFs and ninety nine near full-length sequences representing the full spectrum of known diversity within subtypes B, C, F, G, H, J, and K. The tree is mid-point rooted and was constructed with 1,000 full ML bootstrap replicates using RAxML (Stamatakis 2014). Nodes with bootstrap ≥70% are indicated with black dots. Some clades have been condensed for the sake of clarity.

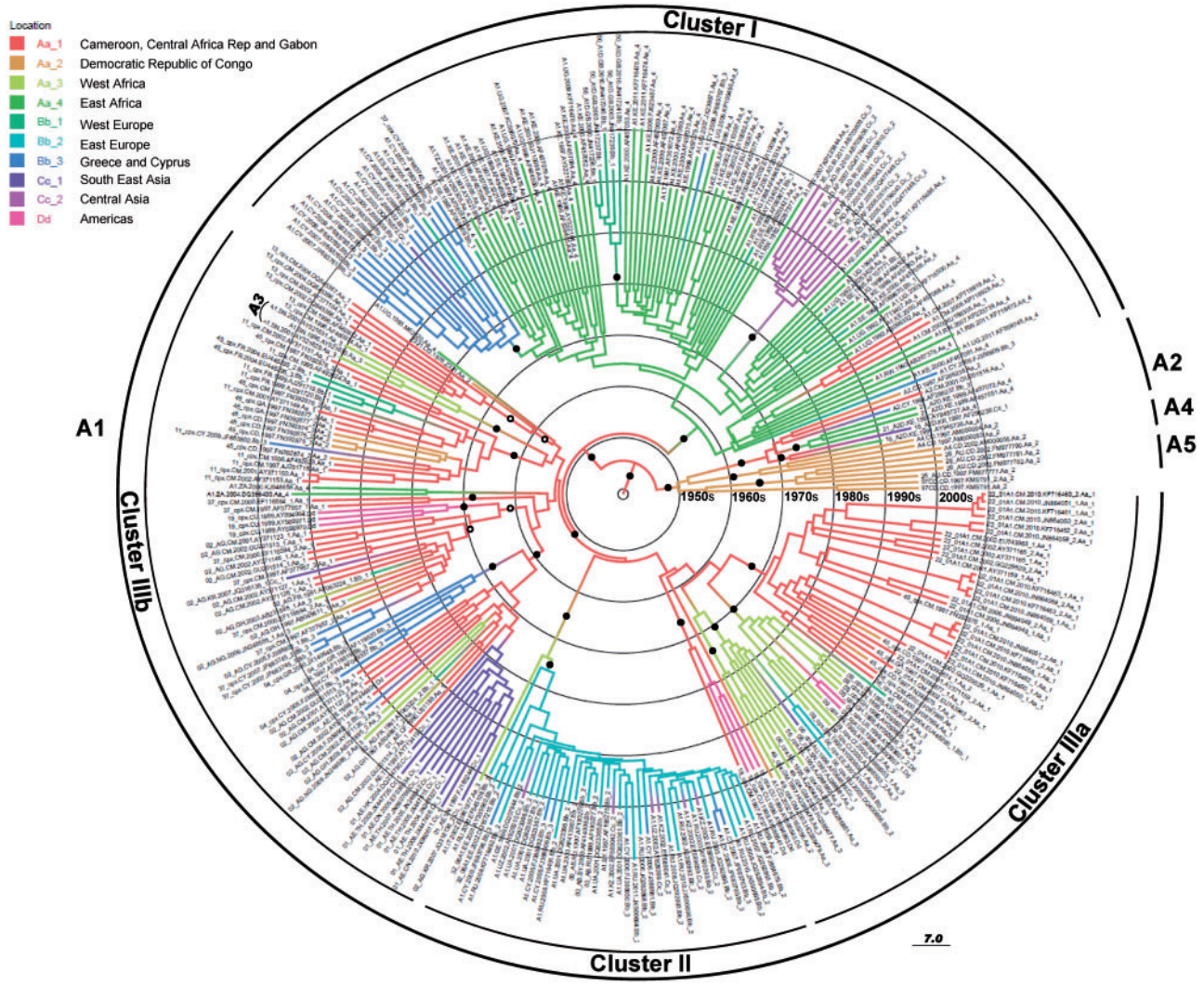

Figure 2. A time-scaled MCC tree indicating the most probable geographical locations of sequences that are ancestral to 306 subtype A sequences. These subtype A sequences include full-length subtype A genomes and contiguous A-attributed fragments within CRF and URF genomes. The colours of the branches correspond to the probable geographic locations of the ancestral viruses indicated by the branches, as indicated in the legend. The tree is temporally scaled such that distances between the concentric grey circles represent 10 years of evolution. Nodes with posterior probabilities >0.95 and >0.80 are indicated with black and open dots, respectively.

we analysed the 306 subtype A and subtype A-attributed genome fragment dataset using a Bayesian phylogeographic approach (Lemey et al. 2009) implemented in BEAST v1.8.4 (Drummond and Rambaut 2007). The best-fit nucleotide substitution model identified using J model test (Darriba et al. 2012; Guindon and Gascuel 2003) was the generalized time-reversible model with four gamma rate categories and a proportion of invariant sites (GTR + G4 + I). The best-fit molecular clock and demographic model combination identified using MLE by both path and stepping stone methods was the uncorrelated lognormal relaxed clock, with the Bayesian Gaussian Markov random field (GMRF) skygrid coalescent tree prior (Gill et al. 2013; Minin, Bloomquist, and Suchard 2008) (Supplementary Table S3). No overlap existed between the distribution of the nucleotide substitution rate estimates (substitutions per site, per year) from the BEAST runs using the correct sampling dates (mean rate $= 4.10 \times 10^{-3}$, 95% HPD $= 3.65–6.25 \times 10^{-3}$) and those with the date-randomization procedure applied (mean rate $= 5.35 \times 10^{-4}$, 95% HPD $= 4.35–6.25 \times 10^{-4}$), providing evidence for significant temporal signal in the time-stamped dataset (Duchene et al. 2015).

The MCC tree yielded by these analyses (Fig. 2) is broadly consistent with the ML tree produced from these sequences (Fig. 1); four sub-lineages within subtype A were also identified and the A1 sub-lineage also contains the same three distinct clusters.

Analysis of the MCC tree produced from the posterior distribution of trees yielded by the BEAST analysis revealed that the MRCA of all the analysed subtype A sequences likely existed in ~1946 (95% HPD: 1939–53; Table 1). This is ~10 years earlier than prior estimates for subtype A based upon analyses of partial *gag* sequences (1956 ± 1) (Abidi et al. 2014), but similar to the proposed date of the ancestor of subtypes B and C (~1944 and 1940, respectively) (Faria et al. 2014). Further, the MRCAs of A1, A2, A4, and A5 are all likely to have existed at some time between the early 1950s and 1970s (Table 1).

These analyses also indicated that, of all the assessed locations, Cameroon, Gabon, or the Central African Republic (CAR) were the most probable ($P = 0.53$) origin of the subtype A MRCA, with the DRC being the next most probable ($P = 0.47$). Collectively, there is very strong support ($P = 1.0$) for the CB as a whole being the region from which subtype A emerged (Fig. 2 and Table 1).

Table 1. Calculated support for the time and origin of the lineages within subtype A.

Group	tMRCA	95% HPD	Posterior prob.	Location	Location prob.	Time of the recombination event	95% HPD	Location of the recombination event	Location prob.
A	1946	1939–53	1	CM/GA/CF	0.53				
A1	1951	1946–57	1	CM/GA/CF	1				
Cluster I ($A1_{Afr}$)	1959	1953–64	1	East Africa	0.59				
35_AD	1989	1986–92	1	Central Asia	1	1968–89	1965–92	East Africa or Central Asia	0.5 each
50_A1D	1988	1985–92	1	West Europe	0.99	1976–88	1973–92	West Europe	0.99
Cluster II ($A1_{Eur}$)	1977	1972–81	0.97	East Europe	0.93				
03_AB	1993	1992–5	0.99	East Europe	1	1992–3	1991–5	East Europe	1
32_06A1	1997	1995–9	1	East Europe	1	1984–97	1980–99	East Europe	0.99
Cluster III (divergent lineages)									
04_cpx	1973	1968–78	1	GE/CY	0.99	1962–73	1956–78	CM/GA/CF-GE/CY	0.5 each
06_cpx	1974	1970–9	1	West Africa	0.93	1967–74	1961–79	West Africa	0.82
09_cpx	1974	1969–78	1	West Africa	0.95	1963–74	1959–78	West Africa	0.61
02_AG_2 and 01_AE	1969	1964–73	0.99	CM/GA/CF	1	1962–69	1956–73	CM/GA/CF	1
02_AG_1 and 37_cpx_2&3	1970	1966–74	0.81	CM/GA/CF	1	1963–70	1959–74	CM/GA/CF	1
13_cpx	1972	1967–77	0.87	CM/GA/CF	1	1961–72	1956–77	CM/GA/CF	0.99
18_cpx	1972	1966–79	1	CM/GA/CF	0.92	1957–72	1951–79	CM/GA/CF	0.94
19_cpx	1974	1968–79	1	Americas	0.98	1965–74	1957–79	CM/GA/CF-CUBA	0.51 and 0.49
22_01A1 and 45_cpx_1 and 36_cpx	1971	1966–76	0.95	CM/GA/CF	1	1963–71	1959–76	CM/GA/CF	0.87
37_cpx_1	1977	1971–84	0.83	CM/GA/CF	1	1957–77	1952–84	CM/GA/CF	1
45_cpx_2&3 and 11_cpx	1966	1961–71	1	CM/GA/CF	1	1956–66	1951–71	CM/GA/CF	1
49_cpx	1980	1974–86	1	West Africa	0.99	1967–80	1961–86	West Africa	0.95
A3	1975	1969–81	1	West Africa	0.97				
A2	1968	1964–73	1	CM/GA/CF	0.98				
16_A2D	1973	1968–77	1	CM/GA/CF	0.96	1968–73	1964–77	CM/GA/CF	0.97
21_A2D	1977	1973–81	1	East Africa	0.99	1970–7	1966–81	CM/GA/CF-East Africa	0.5 each
A4	1966	1958–74	1	DRC	0.99				
A5	1973	1968–78	1	DRC	1	1956–73	1948–78	DRC	0.97

CM, Cameroon; GA, Gabon; CF, Central Africa Republic; GE, Greece; CY, Cyprus; DRC, Democratic Republic of Congo.
02_AG_1 = fragment 1 (HXB2 position 790–2155); 02_AG_2 = fragment 2 (HXB2 position 6225–8311); 37_cpx_1 = fragment 1 (HXB2 position 1126–2142); 31_cpx = fragment 2 (HXB2 position 2913–4663); 37_cpx_3 = fragment 3 (HXB2 position 6431–7743); 45_cpx_1 = fragment 1 (HXB2 position 790–2397); 45_cpx_cpx_2 = fragment 2 (HXB2 position 4299–6041); 45_cpx_3 = fragment 3 (HXB2 position 6343–8236).

Consistent with previous results (Abecasis, Vandamme, and Lemey 2009), the MRCA of the most sampled group, A1, could be traced back to 1951 (95% HPD 1946–57) and was most probably also located in Cameroon/Gabon/CAR ($P = 1.0$). As is shown in Table 1, eleven out of sixteen recombination events involving A1 lineages (mostly from cluster III) happened between 1957 and 1977, before the discovery of the global HIV epidemic in the 1980s. Seven of these likely occurred in Cameroon/Gabon/CAR (with a probability ranging from 0.87 to 1), two either in Cameroon/Gabon/CAR ($P = 0.5$) or in other regions of the world ($P = 0.5$) and two in West Africa ($P = 0.61$ and 0.82). None of the detected recombination events involving A1 parental viruses are inferred to have occurred in the CB region after the 1980s.

The MRCA of the A4 sub-lineage was estimated to have been circulating in ~1966 (1958–74) in the DRC ($P = 0.99$), around the same time as the MRCA of the A2 lineage (~1968; 1964–73), which was most probably circulating in Cameroon/Gabon/CAR ($P = 0.98$). The recombination events involving the A fragments of the A2-derived CRFs also likely happened before the global epidemic was first detected in the 1980s, with the recombination event in CRF16_A2D happening between ~1968 and 1973 (1964–77) in Cameroon/Gabon/CAR ($P = 0.97$) and that in CRF 21_A2D happening between 1970 and 1977 (1966–81) either in Cameroon/Gabon/CAR ($P = 0.5$) or East Africa ($P = 0.5$) (Fig. 2 and Table 1). The MRCA of the A5 lineage likely existed in ~1973 (1968–78) in the DRC ($P = 1$).

These analyses also indicated that A1 was already present in East Africa by 1960 and in West Africa by 1975. In addition, the MRCAs of A-attributed fragments of CRF04_cpx, CRF19_cpx, CRF06_cpx, and CRF09_cpx were already circulating outside the CB before the 1980s (Table 1). The fact that there are so many detectable long-distance movements of A1 and A1-derived recombinants prior to the 1980s is certainly consistent with A1 viruses being predisposed to starting new epidemics.

A sequence too short to be included in our dataset and derived from a virus sampled in the DRC in 1960 (DRC60) was previously found to fall basal to the A4 viruses (Worobey et al. 2008). To work out where this sequence would have been located in our MCC tree, we manually aligned the 508 fragmentary nucleotides of this sequence to the rest of our dataset and constructed a ML tree with RAxML. This tree indicated that, as expected, DRC60 would have likely been located, near the base of the A4 cluster in our MCC tree (Supplementary Fig. S4). The MRCA of the A4 cluster was estimated to have existed in ~1959 (Worobey et al. 2008). This date falls within the 95% HPD of our estimate of the A4 MRCA (1958–74). Further, our inference that the A4 MRCA was located within the DRC is also consistent with the analysis of Worobey et al. (2008).

3.4 Comparison of signals of selection between different subtype A sequence lineages

To uncover evidence of substantial evolutionary innovations within different sub-subtype A1 lineages during the era after these viruses began dispersing throughout the world, we compared signals of codon evolution between various monophyletic A1 sequence lineages. Specifically, we compared signals of selection acting at individual codon sites within the *gag*, *pol*, and *env* genes of A1 sequences circulating in Africa ($A1_{Afr}$) with those of four independent monophyletic clades within the A1 radiation, namely (1) A1 viruses circulating in Europe ($A1_{Eur}$), (2) CRF 01_AE, (3) CRF02_AG, and (4) CRF22_01A1. The alignments were segmented such that they contained only subtype A-derived genome fragments in all the analysed genes. Specifically, we tested whether signals of negative selection evident within A1 lineage sequences close to the location where A1 originated were preserved in A1 sequences that have either left Africa (in the case of $A1_{Eur}$), or have left the context of 'pure' A1 genomes (in the case of CRFs 01_AE, 02_AG, and 22_01A1). If A1 sequences are in general genetically predisposed either to become the founders of foreign epidemics, or to increase the fitness of the genomes into which they are transferred, then we anticipated that patterns of negative selection (i.e. selection favouring the maintenance of adaptively favourable amino acid sequences) should be broadly conserved between these different groups of viruses. Alternatively, if the A1 sequences that emigrated from Africa were not particularly well adapted to their new host populations then we would expect them to display increased evidence relative to $A1_{Afr}$ of codon sites evolving under either diversifying positive selection (i.e. where changes away from amino acid sites found in $A1_{Afr}$ are selectively favoured), or directional positive selection (i.e. where changes toward specific amino acid sequences are favoured). Similarly, if the sub-subtype A1 sequences that had been transferred into foreign genomic backgrounds were not particularly well adapted to their new genomic contexts then we would also expect them to display increased evidence relative to $A1_{Afr}$ of codon sites evolving under either diversifying positive selection or directional positive selection.

Overall, the distributions of negatively selected sites were strongly conserved between the $A1_{Afr}$ genes and those of the $A1_{Eur}$, CRFs 01_AE, 02_AG, and 22_01A1 datasets (Fig. 3 and Supplementary Fig. S2) in *gag* and *pol*. This broad conservation of negative selection patterns is consistent with sub-subtype A1 *gag*, and *pol* sequences being pre-adapted to functioning in host and viral genomic environments different to those within which they evolved.

It is noteworthy, however, that there were some codon sites within the A1-attributed *gag*, *pol*, and *env* segments of CRFs 01_AE, 02_AG, 22_01A1 and in the $A1_{Eur}$ sequences, which were detectably evolving under negative selection favouring the maintenance of different amino acids than those which have been selectively favoured in the $A1_{Afr}$ viruses (sites indicated in green in Fig. 3 and in Table 2). For example, whereas in the 01_AE *gag* sequences, negative selection is detectably favouring the maintenance of asparagine at positions 165 (from the beginning of *gag* and with gaps deleted) and lysine at position 384, in $A1_{Afr}$, Serine and Asparagine are selectively favoured at positions 165 and 384, respectively (Table 2). Similarly, Isoleucine at position 267 and Arginine at position 406 were replaced by Valine and Lysine, respectively in 02_AG (Table 2). At all these positions in the 01_AE and 02_AG sequences, there were a minority of viruses that had the same amino acid at these sites as viruses in the $A1_{Afr}$ lineage (Table 2). This suggests that directional selection favouring the non-$A1_{Afr}$ states has likely operated only after the founding of the 01_AE and 02_AG clades.

Another example of evidence of differential selection pressures operating on these different groups of viruses is seen at position 559 of *pol*. Whereas an aspartic acid at this position is selectively favoured in the $A1_{Afr}$ clade, a glutamic acid is predominantly found at the site in all four of the other groups (Supplementary Fig. S2 and Table 2).

Taken together, these results imply that during the divergence of A1 sequences within the $A1_{Eur}$, CRFs 01_AE, 02_AG and 22_01A clade from their $A1_{Afr}$ relatives, only low numbers of sites have been evolving under directional selection away from the $A1_{Afr}$ states, presumably to adapt these sequences to their new host or genomic environments. The overwhelming majority of negatively selected sites in $A1_{Afr}$ sequences have been consistently evolving to maintain the same encoded amino

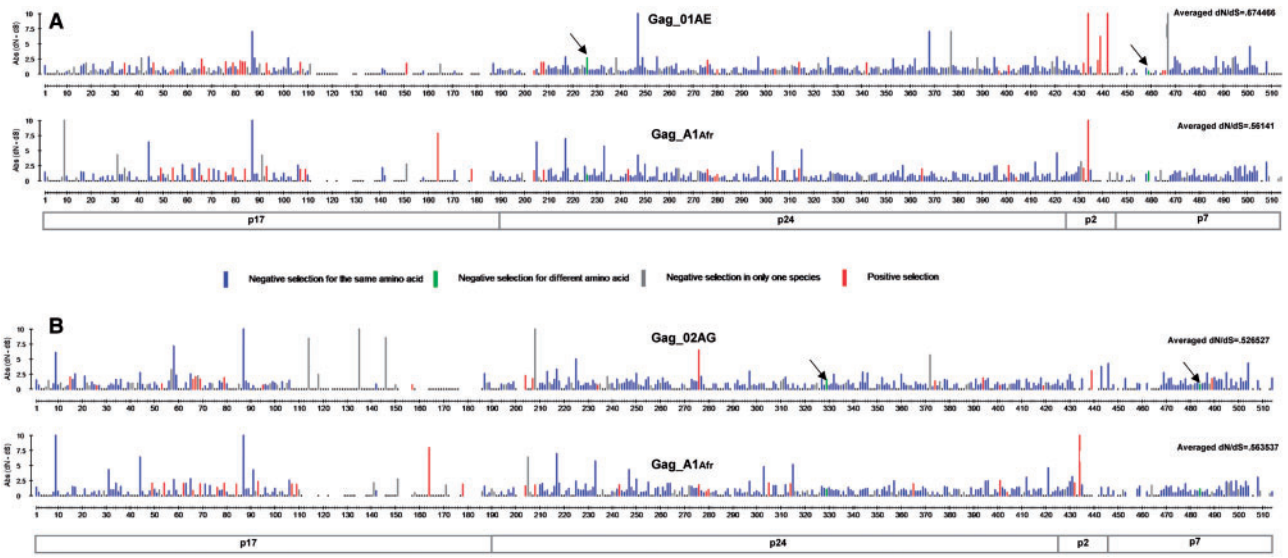

Figure 3. Patterns of natural selection acting at sub-subtype A1-derived *gag* codon sites in CRFs 01_AE (A) and 02_AG (B) compared to those acting on homologous codon sites in sub-subtype A1 genomes that were sampled in Africa ($A1_{Afr}$). Codon site locations that are indicated are specific for the alignment used to analyse all these sequences. At each site, absolute values of the inferred non-synonymous substitution rate minus the synonymous substitution rate (dN − dS) are plotted (as determined by the FUBAR method). Significantly positive dN − dS values which are indicative of positive selection are indicated in red, whereas significantly negative dN − dS values which are indicative of negative selection are plotted in blue if negative selection for the same amino acid state is found in the two compared lineages. They are plotted in green if negative selection is detected in both compared lineages, but different amino acids are being selected, and in grey if negative selection is only detectable in one of the compared lineages. Overall dN/dS ratios <1 indicate that, as expected, the *gag* gene in all three clades is evolving under predominantly negative selection. Black arrows indicate clusters of codon sites under negative selection for different amino acids in the two compared lineages.

Table 2. Amino acid sites variation between $A1_{Afr}$, $A1_{Eur}$, CRFs 01_AE, 02_AG, and 22_01A1.

Site[a]	$A1_{Afr}$	$A1_{Eur}$	01_AE	02_AG	22_01A1
Gag-165	**Ser**(Asn)[b]	**Ser**(Gly)	**Asn**(Ser, Lys)	**Ser**(Asn)	
Gag-267	**Ile**	**Ile**	Ile(Val)	**Val/Ile**[c]	
Gag-384	**Arg**(Lys)	**Arg**(Gly)	**Lys**(Thr, Asn, Arg)	**Arg**(Lys)	
Gag-406	**Arg**(Lys)	**Arg**(Lys)	**Arg**	**Lys**(Arg)	
Pol-484	**Ile**(Val, Leu)	**Ile**	**Val**(Ile, Leu)	**Ile**(Val, Leu)	**Ile**(Val)
Pol-559	**Asp**(Glu)	**Glu**	**Glu**	**Glu**(Asp)	**Glu**
Pol-828	**Val**(Ile)	**Val**(Ile)	**Ile**(Val)		**Ile**(Val)
Env-514	**Val**(Met, Ala, Phe, Leu)	**Ala**(Val)		**Val**(Phe, Ala, Leu, Ile)	**Val**(Phe, Leu)

[a]Sites numbered from the beginning of each *gene* and according to the reference strain HIV-1/HXB2.
[b]Amino acid found in the majority (>80%) of analysed sequences in bold, while minor amino acids are in brackets.
[c]Amino acids found at approximately the same proportion of analysed sequences in bold. Grey shaded boxes indicate site under negative selection.

acids at these sites across all the A1 sequence lineages, presumably to maintain the fitness advantages provided by these amino acids.

4. Discussion

Subtype A remains a major circulating lineage within the CB region where HIV-1M originated, and is both the most widely distributed and most genetically diverse of the known HIV-1M subtypes. Here, for the first time, we have analysed when and where the main subtype A lineages and subtype A-derived recombinant lineages arose using a comprehensive set of published subtype A, near full-length sequences together with subtype A-derived genome fragments from CRFs and URFs. Our analysis indicates the origins of most of the subtype A-derived CRFs were likely within the CB. We also present evidence supporting the occurrence of multiple independent epidemics being seeded by subtype A viruses and viruses containing subtype A-derived sequences throughout the world prior to the discovery of the HIV epidemic in the 1980s. We further examined subtype A *gag*, *pol*, and *env* genes for patterns of evolution that might be expected within an HIV lineage that was genetically predisposed to founding new epidemics.

We inferred that, starting in the late 1950s within the CB, the best sampled of the subtype A sub-subtypes, A1, experienced a burst of diversification that involved frequent inter-subtype recombination. This suggests both that A1 was circulating at high frequencies within the CB at that time and that these high frequencies could have contributed to the global spread of A1 and A1-derived sequences out of this region during the 1960s and 1970s.

Although both subtypes B (and its sister clade D) and C originated in the CB at approximately the same time as subtype A (Faria et al. 2014), and today collectively account for approximately 60% of infections worldwide (Hemelaar et al. 2011), they have experienced only limited spread within the CB region; subtype C represents around 10% of viruses circulating in Angola

(Bartolo et al. 2009), its prevalence is less than 5% in the DRC (Rodgers et al. 2017b) and has only rarely been found in other countries of the CB. Subtype D, accounts for approximately 15% of HIV infections in the Republic of Congo (Niama et al. 2006) and 10% in the DRC (Rodgers et al. 2017b), but has also only been very rarely found in other CB countries. Further, despite the broad spectrum of HIV-1M genetic diversity within the CB today, there remains a predominance of subtype A and subtype A-derived recombinant forms in the DRC, Republic of Congo and Angola (where A1 dominates [Niama et al. 2006; Bartolo et al. 2009; Rodgers et al. 2017b]), Cameroon, Gabon, and Equatorial Guinea (where CRF02 AG dominates [Pandrea et al. 2002; Djoko et al. 2010; Tongo et al. 2013]) and the CAR (where CRF11_cpx predominates [Marechal et al. 2006]). It is possible that at the onset of the global epidemic, the A1 lineage was circulating within the CB at higher frequencies than subtypes B/D, C, and G. Although this could just have been a consequence of A1 viruses simply being present at the right places and times for dispersal, it remains plausible that within this large and diverse A1 population there were genetic variants with increased reproductive fitness that, upon dispersal, were poised to 'colonize' both the entire CB region and the rest of the world.

The global dissemination of A1 and A1-derived recombinants contrasts starkly with the lower prevalence and more geographically constrained ranges of the other main subtype A lineages. Just as there appear to be subtype-specific differences in rates of disease progression and pathogenesis (Mann et al. 2013; Brandenberg et al. 2015; Venner et al. 2016), it is possible that different subtype A lineages might also vary with respect to their epidemiologically relevant biological characteristics. The absence of non-disseminating subtype A lineages within or around hubs of human migration within the DRC during the 1940s, 1950s, and 1960s might explain the absence of these viruses in other parts of the world. It cannot, however, explain why A2, A3, A4, and A5 viruses have not spread throughout the CB as extensively as A1 viruses: especially given that viruses that were most closely related to the present A4 lineage were likely circulating within populous cities such as Kinshasa in the 1950s and 1960s (Worobey et al. 2008). It therefore remains plausible that viruses belonging to the rarer subtype A lineages could have simply been less fit than those belonging to the A1 lineage and, therefore, that in the 1950s, A1 was evolutionarily better poised for successful global dissemination. In addition, we identified fifteen subtype A-derived CRFs that most likely arose somewhere in the CB prior to the 1980s (Table 1). Twelve of these have an A1 parent, two an A2 parent, and one an A5 parent. This is certainly consistent with the hypothesis that members of the A1 lineage were already fitter and/or more prevalent than members of the other A sub-subtypes throughout the 1960s and 1970s, when these recombination events likely occurred.

It remains unclear why, before the almost simultaneous exodus of subtype A, B, D, and C viruses from the CB, subtype A viruses appear to have participated more extensively in the generation of the known circulating recombinant viruses than those in these other subtypes. The geographical origins of these viruses may explain this enigma. It is presently believed that, as with subtypes B and C (Faria et al. 2014), A1 most likely originated in the CB. Our analyses, however, indicate that the MRCA of all the analysed A1 genomes most probably existed in a region that included Cameroon, Gabon, and the CAR. Furthermore, 64% of the recombination events involving transfers of A1 sequences that were inferred to have happened prior to the 1980s, were also predicted in our analyses to have most probably occurred within this same region (Fig. 2 and Table 1). Consistent with this possibility is the fact that eleven out of the eighteen CRFs that contain A1-derived genome fragments have been sampled in Cameroon and four of these eleven (13, 36, 37, and 22) have been found only in either Cameroon or in Cameroonian individuals (whenever these have been found outside of Cameroon).

It is difficult to explain how so much recombination took place during the early HIV epidemic in Cameroon. Cameroon presently has an HIV prevalence of only 5%, which is very low by African standards, and presumably also has a correspondingly low prevalence of dual infections. It is, however, possible that isolated high HIV prevalence hotspots may have persisted for decades within the country. For example, in some remote Cameroonian villages, up to 20% of individuals are HIV positive (Zhong et al. 2002) and in others, an unparalleled diversity of HIV variants with A-attributed fragments, including numerous complex URFs, have been discovered (Konings et al. 2006; Carr et al. 2010; Rodgers et al. 2017a). It therefore follows that high HIV-1M prevalence in some of these remote Cameroonian villages might had provided the opportunity for A1 to extensively recombine. A factor that may, at least in part, account for these pockets of high prevalence, could be that HIV-1M variants within these regions of Cameroon have likely been adapting to a genetically consistent human population for far longer than HIV variants found in other, more cosmopolitan, parts of the world. If these Cameroonian viruses are better adapted to infecting humans than are HIV variants found elsewhere, this could be reflected in the average duration of HIV-1M infections and/or the transmissibility of viruses being substantially greater in Cameroon than they are in other parts of the world.

In our study, we found that the sub-subtype A3 is actually a sublinage inside the sub-subtype A1 radiation, and remains so in a tree constructed with only 'pure' subtype A sequences. This result contradicts the classification proposed by the Los Alamos HIV sequence database (LANL 2014). The apparent disagreement possibly stems from the fact that at the time when the A3 sequences were classified there were fewer and/or less diverse A1 whole genome sequences that were publically available for comparison. Further, the fact that the A3 clade connects with the A1 clade via a deep branch within the A1 radiation emphasizes the uncertainty of whether this clade should be considered a distinct sub-subtype or simply a divergent lineage of A1.

Although numerous studies have described signals of positive and negative selection acting on codon sites within HIV-1M genomes (Ngandu et al. 2008; Mayrose et al. 2013), here we looked for evidence consistent with homologous codons from African A1 genomes, European A1 genomes, and CRF genomes being under subtly different selection pressures. We identified multiple codon sites within the A-attributed gene segments of some recombinants at which different encoded amino acids were selectively favoured relative to members of the parental African A1 lineage. The few exceptional sites that display this pattern of evolution have potentially undergone post-recombination or post-movement changes to better adapt fragments of subtype A sequence to their new genomic or environmental contexts. One example of this might come from the subtype A1 epidemic in Russia were a subtype A variant with a non-nucleoside reverse transcriptase inhibitor resistance mutation is highly prevalent (Kolomeets et al. 2014). However, our results also indicate that the vast majority of codon sites within the subtype A-derived portions of CRFs 01_AE, 02_AG, and 22_01A1 have not obviously undergone extensive adaptation to their new genomic contexts; that is that patterns of negative selection at these sites have

remained broadly unchanged between these different groups of viruses and their parental A1 lineage. This suggests that, even before A1 sequences emigrated from Africa, or A1 genomic fragments were transferred into foreign genomic backgrounds, these sequences may have already been highly adapted to the geographical and/or genomic environments into which they were transferred. It is also likely, however, that, regardless of their geographical and/or genomic environments, similarities between the selection signals observable in the sequences of these different A1 lineages are a consequence of these lineages being subjected to similar selective forces following their diversification. In this regard, while similarities between these selection signals are consistent with pre-adaptation, they are not by themselves proof of pre-adaptation.

We have presented circumstantial evidence that is consistent with the hypothesis that, at the onset of the global HIV epidemic, sub-subtype A1 viruses circulating in the CB were genetically predisposed to successfully found HIV epidemics in other parts of the world. This evidence includes: (1) the high number of instances where sub-subtype A1 viruses were inferred to have been transferred to, and founded new subepidemics in, different parts of the world prior to the 1980s; (2) the genesis within the CB during the 1960s and 1970s of large numbers of CRFs containing A1-derived genome fragments and the ensuing high prevalence of A1 attributable genome fragments within recombinant HIV genomes; and (3) evidence of consistently pervasive negative selection favouring the maintenance of the same amino acid states within A1-derived coding regions both between independent groups of A1 viruses sampled in Africa and Europe, and between parental A1 lineages and A1-derived genome fragments within CRFs. If sub-subtype A1 sequences are indeed predisposed to successfully seed new epidemics, it would be of utmost importance to identify the precise genetic underpinnings of this predisposition as knowing these could tremendously enhance our capacity to predict future HIV-1M emergence and dissemination events.

Acknowledgements

Computational analyses were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing Facility (http://www.hpc.uct.ac.za) and the South African National Bioinformatics Institute.

Funding

This work was supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, or the UK government. This research was also supported by the Poliomyelitis Research Foundation (PRF) of South Africa. M.T. is a SANTHE Postdoctoral Fellow. D.P.M. is supported by the National Research Foundation (NRF) of South Africa. T. d.O is supported by a South African Medical Research Council (SA MRC) Flagship grant (MRC-RFA-UFSP-01- 2013/UKZN HIVEPI) as well by an UK Royal Society Newton Advanced Fellowship. The authors declare that they have no competing interests.
Disclaimer: The opinions expressed herein are those of the authors and should not be construed as official or representing the views of the U.S. Department of Defence or Department of the Army. Mention of trade names, commercial products, or organizations does not imply endorsement by the US government.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Abecasis, A. B. et al. (2009) 'Quantifying Differences in the Tempo of Human Immunodeficiency Virus Type 1 Subtype Evolution', *Journal of Virology*, 83: 12917–24.

Abidi, S. H. et al. (2014) 'HIV-1 Subtype A *gag* Variability and Epitope Evolution', *PLoS One*, 9: e93415.

Baele, G. et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29: 2157–67.

Bartolo, I. et al. (2009) 'Highly Divergent Subtypes and New Recombinant Forms Prevail in the HIV/AIDS Epidemic in Angola: New Insights into the Origins of the AIDS Pandemic', *Infection, Genetics and Evolution*, 9: 672–82.

Brandenberg, O. F. et al. (2015) 'Different Infectivity of HIV-1 Strains Is Linked to Number of Envelope Trimers Required for Entry', *PLoS Pathogens*, 11: e1004595.

Caron, M. et al. (2012) 'Prevalence, Genetic Diversity and Antiretroviral Drugs Resistance-Associated Mutations among Untreated HIV-1-Infected Pregnant Women in Gabon, Central Africa', *BMC Infectious Diseases*, 12: 64.

Carr, J. K. et al. (2010) 'HIV-1 Recombinants with Multiple Parental Strains in Low-Prevalence, Remote Regions of Cameroon: Evolutionary Relics?', *Retrovirology*, 7: 39.

Darriba, D. et al. (2012) 'jModelTest 2: More Models, New Heuristics and Parallel Computing', *Nature Methods*, 9: 772.

Djoko, C. F. et al. (2010) 'HIV Type 1 Pol Gene Diversity and Genotypic Antiretroviral Drug Resistance Mutations in Malabo, Equatorial Guinea', *AIDS Research and Human Retroviruses*, 26: 1027–31.

Drummond, A. J., and Rambaut, A. (2007) 'BEAST: Bayesian Evolutionary Analysis by Sampling Trees', *BMC Evolutionary Biology*, 7: 214.

Duchene, D. A. et al. (2015) 'Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations', *Molecular Biology and Evolution*, 32: 2986–95.

Edgar, R. C. (2004) 'MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity', *BMC Bioinformatics*, 5: 113.

Faria, N. R. et al. (2014) 'HIV Epidemiology. The Early Spread and Epidemic Ignition of HIV-1 in Human Populations', *Science*, 346: 56–61.

Foley, B. T. et al. (2016) 'Primate Immunodeficiency Virus Classification and Nomenclature: Review', *Infection, Genetics and Evolution*, 46: 150–8.

Gill, M. S. et al. (2013) 'Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci', *Molecular Biology and Evolution*, 30: 713–24.

Guindon, S., and Gascuel, O. (2003) 'A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood', *Systematic Biology*, 52: 696–704.

Hemelaar, J. et al. (2011) 'Global Trends in Molecular Epidemiology of HIV-1 during 2000-2007', *Aids (London, England)*, 25: 679–89.

Jetzt, A. E. et al. (2000) 'High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome', *Journal of Virology*, 74: 1234–40.

Khoosal, A., and Martin, D. P., www.cbio.uct.ac.za/~arjun/.

Kiguoya, M. W. et al. (2017) 'Subtype-Specific Differences in Gag-Protease-Driven Replication Capacity Are Consistent with Intersubtype Differences in HIV-1 Disease Progression', *Journal of Virology*, 91: e00253-17.

Kolomeets, A. N. et al. (2014) 'A Uniquely Prevalent Nonnucleoside Reverse Transcriptase Inhibitor Resistance Mutation in Russian Subtype a HIV-1 Viruses', *Aids*, 28: F1–8.

Konings, F. A. et al. (2006) 'Genetic Analysis of HIV-1 Strains in Rural Eastern Cameroon Indicates the Evolution of Second-Generation Recombinants to Circulating Recombinant Forms', *Journal of Acquired Immune Deficiency Syndromes*, 42: 331–41.

LANL. 2014. 'Los Alamos National Laboratory HIV-1 database' <http://hiv-web.lanl.gov/content/hiv-db> accessed June 2016.

Lemey, P. et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520.

Mann, J. K. et al. (2013) 'Ability of HIV-1 Nef to Downregulate CD4 and HLA Class I Differs among Viral Subtypes', *Retrovirology*, 10: 100.

Marechal, V. et al. (2006) 'Increasing HIV Type 1 Polymorphic Diversity but No Resistance to Antiretroviral Drugs in Untreated Patients from Central African Republic: A 2005 Study', *AIDS Research and Human Retroviruses*, 22: 1036–44.

Markle, T. J. et al. (2013) 'HIV-1 Nef and T-Cell Activation: A History of Contradictions', *Future Virology*, 8

Martin, D. P. et al. (2015) 'RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes', *Virus Evolution*, 1: 391–404.

Mayrose, I. et al. (2013) 'Synonymous Site Conservation in the HIV-1 Genome', *BMC Evolutionary Biology*, 13: 164.

Miller, M. A. et al. (2010) 'Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees', in *Proceedings of the Gateway Computing Environments Workshop* (GCE), pp 1–8, New Orleans, LA.

Minin, V. N. et al. (2008) 'Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics', *Molecular Biology and Evolution*, 25: 1459–71.

Murrell, B. et al. (2013) 'FUBAR: A Fast, Unconstrained Bayesian Approximation for Inferring Selection', *Molecular Biology and Evolution*, 30: 1196–205.

Ngandu, N. K. et al. (2008) 'Extensive Purifying Selection Acting on Synonymous Sites in HIV-1 Group M Sequences', *Virology Journal*, 5: 160.

Niama, F. R. et al. (2006) 'HIV-1 Subtypes and Recombinants in the Republic of Congo', *Infection, Genetics and Evolution*, 6: 337–43.

Pandrea, I. et al. (2002) 'Analysis of Partial Pol and Env Sequences Indicates a High Prevalence of HIV Type 1 Recombinant Strains Circulating in Gabon', *AIDS Research and Human Retroviruses*, 18: 1103–16.

Powell, R. L. et al. (2007a) 'Circulating Recombinant Form (CRF) 37_Cpx: An Old Strain in Cameroon Composed of Diverse, Genetically Distant Lineages of Subtypes a and G', *AIDS Research and Human Retroviruses*, 23: 923–33.

—— et al. (2007b) 'Identification of a Novel Circulating Recombinant Form (CRF) 36_cpx in Cameroon That Combines Two CRFs (01_AE and 02_AG) with Ancestral Lineages of Subtypes a and G', *AIDS Research and Human Retroviruses*, 23: 1008–19.

Price, M. N. et al. (2009) 'FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix', *Molecular Biology and Evolution*, 26: 1641–50.

Rahimi, A. et al. (2017) 'In Vitro Functional Assessment of Natural HIV-1 Group M *vpu* Sequences Using a Universal Priming Approach', *Journal of Virological Methods*, 240: 32–41.

Rambaut, A., and and Drummond, A. J. (2014). 'Figtree v1.4 <http://tree.bio.ed.ac.uk/software/figtree/> accessed 2017.

—— et al. (2004) 'The Causes and Consequences of HIV Evolution', *Nature Reviews Genetics*, 5: 52–61.

—— et al. (2014) 'Tracer v1. 6 <http://beast.bio.ed.ac.uk/Tracer> accessed 2017.

Rhodes, T. et al. (2003) 'High Rates of Human Immunodeficiency Virus Type 1 Recombination: Near-Random Segregation of Markers One Kilobase Apart in One round of Viral Replication', *Journal of Virology*, 77: 11193–200.

Rodgers, M. A. et al. (2017a) 'Identification of Rare HIV-1 Group N, HBV AE, and HTLV-3 Strains in Rural South Cameroon', *Virology*, 504: 141–51.

—— et al. (2017b) 'Sensitive Next-Generation Sequencing Method Reveals Deep Genetic Diversity of HIV-1 in the Democratic Republic of the Congo', *Journal of Virology*, 91: e01841-16.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics (Oxford, England)*, 30: 1312–3.

——, and Alachiotis, N. (2010) 'Time and Memory Efficient Likelihood-Based Tree Searches on Phylogenomic Alignments with Missing Data', *Bioinformatics*, 26: i132–9.

Stenzel, T. et al. (2014) 'Pigeon Circoviruses Display Patterns of Recombination, Genomic Secondary Structure and Selection Similar to Those of Beak and Feather Disease Viruses', *Journal of General Virology*, 95: 1338–51.

Tamura, K. et al. (2011) 'MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods', *Molecular Biology and Evolution*, 28: 2731–9.

Taylor, B. S. et al. (2008) 'The Challenge of HIV-1 Subtype Diversity', *The New England Journal of Medicine*, 358: 1590–602.

Tebit, D. M., and Arts, E. J. (2011) 'Tracking a Century of Global Expansion and Evolution of HIV to Drive Understanding and to Combat Disease', *The Lancet Infectious Diseases*, 11: 45–56.

Tee, K. K. et al. (2009) 'Estimating the Date of Origin of an HIV-1 Circulating Recombinant Form', *Virology*, 387: 229–34.

Tongo, M. et al. (2015a) 'Near Full-Length HIV Type 1M Genomic Sequences from Cameroon: Evidence of Early Diverging

under-Sampled Lineages in the Country', *Evolution, Medicine, and Public Health*, 2015: 254–65.

—— et al. (2015b) 'High Degree of HIV-1 Group M (HIV-1M) Genetic Diversity within Circulating Recombinant Forms: Insight into the Early Events of HIV-1M Evolution', *Journal of Virology*, 90: 2221–9.

—— et al. (2013) 'Characterization of HIV-1 *gag* and *nef* in Cameroon: Further Evidence of Extreme Diversity at the Origin of the HIV-1 Group M Epidemic', *Virology Journal*, 10: 29.

Trovao, N. S. et al. (2015) 'Host Ecology Determines the Dispersal Patterns of a Plant Virus', *Virus Evolution*, 1: vev016.

Venner, C. M. et al. (2016) 'Infecting HIV-1 Subtype Predicts Disease Progression in Women of Sub-Saharan Africa', *EBioMedicine*, 13: 305–14.

Worobey, M. et al. (2008) 'Direct Evidence of Extensive Diversity of HIV-1 in Kinshasa by 1960', *Nature*, 455: 661–4.

Zhang, M. et al. (2010) 'The Role of Recombination in the Emergence of a Complex and Dynamic HIV Epidemic', *Retrovirology*, 7: 25.

Zhong, P. et al. (2002) 'HIV Type 1 Group M Clades Infecting Subjects from Rural Villages in Equatorial Rain Forests of Cameroon', *Journal of Acquired Immune Deficiency Syndromes*, 31: 495–505.