Phylodynamics of HIV-1 Subtype C Epidemic in East Africa

Edson Oliveira Delatorre, Gonzalo Bello*

Laboratório de AIDS & Imunologia Molecular, Instituto Oswaldo Cruz, Rio de Janeiro, Brazil

Abstract

The HIV-1 subtype C accounts for an important fraction of HIV infections in east Africa, but little is known about the genetic characteristics and evolutionary history of this epidemic. Here we reconstruct the origin and spatiotemporal dynamics of the major HIV-1 subtype C clades circulating in east Africa. A large number ($n = 1,981$) of subtype C *pol* sequences were retrieved from public databases to explore relationships between strains from the east, southern and central African regions. Maximum-likelihood phylogenetic analysis of those sequences revealed that most (>70%) strains from east Africa segregated in a single regional-specific monophyletic group, here called C_{EA}. A second major Ethiopian subtype C lineage and a large collection of minor Kenyan and Tanzanian subtype C clades of southern African origin were also detected. A Bayesian coalescent-based method was then used to reconstruct evolutionary parameters and migration pathways of the C_{EA} African lineage. This analysis indicates that the C_{EA} clade most probably originated in Burundi around the early 1960s, and later spread to Ethiopia, Kenya, Tanzania and Uganda, giving rise to major country-specific monophyletic sub-clusters between the early 1970s and early 1980s. The results presented here demonstrate that a substantial proportion of subtype C infections in east Africa resulted from dissemination of a single HIV local variant, probably originated in Burundi during the 1960s. Burundi was the most important hub of dissemination of that subtype C clade in east Africa, fueling the origin of new local epidemics in Ethiopia, Kenya, Tanzania and Uganda. Subtype C lineages of southern African origin have also been introduced in east Africa, but seem to have had a much more restricted spread.

Citation: Delatorre EO, Bello G (2012) Phylodynamics of HIV-1 Subtype C Epidemic in East Africa. PLoS ONE 7(7): e41904. doi:10.1371/journal.pone.0041904

Editor: Marco Salemi, University of Florida, United States of America

Received April 23, 2012; **Accepted** June 27, 2012; **Published** July 27, 2012

Funding: This study was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro and Conselho Nacional de Desenvolvimento Científico e Tecnológico grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gbellobr@gmail.com

Introduction

Human immunodeficiency virus type 1 (HIV-1) sequences belonging to the pandemic group M are classified into nine subtypes (A–D, F–H, J, and K), six sub-subtypes (A1–A4, and F1–F2), and a variety of inter-subtype recombinant forms (Los Alamos HIV sequence database: http://hiv-web.lanl.gov/). Subtype C is the most prevalent variant, accounting for nearly half (48%) of all global infections [1]. This high prevalence is due to the predominance of subtype C in southern Africa, east Africa and India, with further infections in central Africa and Brazil.

Subtype C accounts for >95% of HIV infections in all southern African countries [1]. Several studies showed that subtype C sequences from neighboring southern African nations display a great degree of phylogenetic intermixing with no evidence of significant geographical clustering [2,3,4,5,6,7], indicating a largely unrestricted viral movement across the entire subcontinent. A more recent phylogenetic study revealed that after sequential pruning of ambiguously positioned taxa 10 strongly supported subtype C clusters becomes apparent in southern Africa, showing that the geographic subdivision of subtype C viruses circulating in this region is higher than expected by chance [8]. Most subtype C clusters identified, however, circulate in more than one southern African country and all four countries analyzed (Botswana, Malawi, South Africa and Zambia) comprise strains from multiple clusters. Thus, HIV epidemics in southern African countries are probably the result of the introduction and circulation of multiple subtype C strains with a variable level of local and regional dissemination.

In contrast to the southern African region, the prevalence of HIV-1 subtype C clade displays a great variation among eastern African countries. Subtype C reaches high prevalence in Burundi (>80%) [9,10], Djibouti (>70%) [11] and Ethiopia (>95%) [12,13,14,15], medium prevalence in Tanzania (20–40%) [16,17,18,19,20], and relatively low prevalence in Rwanda (14%) [21] and Uganda (<5%) [22,23,24,25,26,27,28]. Subtype C also accounts for a minor fraction (<15%) of HIV infections in western [29,30,31], coastal [28,32,33,34] and central [28,35,36,37] regions of Kenya; but displays a much higher frequency (25–50%) in some cities of the northern region that borders Ethiopia [38,39].

Little is known about the genetic characteristics of HIV-1 subtype C strains circulating in east Africa. Previous studies showed that two genetically different subtype C strains designated C and C′, have been co-circulating in roughly similar prevalence and among the same risk groups and geographical areas in Ethiopia [13,15,40]. A recent study of Thomson and Fernández-García [8] revealed that the Ethiopian-C clade corresponds to one subtype C cluster also found in other east African countries including Burundi, Djibouti, Kenya, and Uganda; while the Ethiopian-C′ clade was assigned to an independent cluster associated to southern Africa. Other studies performed in Kenya showed that subtype C samples from this country are not concentrated in a single cluster, but distributed in several independent lineages associated to sequences from both east and southern Africa [34,39]. Despite these previous studies, we still have an incomplete understanding of the number, onset date, and migration pattern of the distinct HIV-1 subtype C lineages circulating in the eastern African region.

To obtain a more comprehensive picture of the spatiotemporal dynamics of the HIV-1 subtype C epidemic in east Africa, we analyzed a large number of subtype C *pol* sequences sampled from the east (Burundi, Ethiopia, Kenya, Tanzania and Uganda), southern (Botswana, Malawi, Mozambique, South Africa, Zambia and Zimbabwe) and central (Angola and Democratic Republic of Congo) African regions over a time period of 25 years (1986–2010).

Materials and Methods

Sequence dataset

HIV-1 subtype C *pol* sequences from east, southern, and central African countries, that matched the selected genomic region (nt 2253–3272 relative to HXB2 clone) were retrieved from the Los Alamos HIV Database (http://hiv.lanl.gov). Countries were grouped in geographical regions according to the classification proposed by Hemelaar *et al* [1]. In order to improve the accuracy of phylogenetic inference only sequences from antiretroviral therapy naïve individuals were selected. The subtype assignment of all sequences was confirmed by the REGA HIV subtyping tool v.2 [41] and by maximum likelihood (ML) phylogenetic analysis (see below) with HIV-1 subtype reference sequences. Those sequences with incorrect subtype C classification, sequences containing frame-shift mutations or deletions, multiple sequences from the same individual and those sequences from countries poorly represented (<5 sequences) were removed. This resulted in a final dataset of 1,981 HIV-1 subtype C *pol* sequences sampled from 12 different African countries (Table 1). Sequences were aligned using the CLUSTAL X program [42] and alignment is available from the authors upon request.

Table 1. HIV-1 subtype C sequences.

African region	Country	*N*	Sampling date
Central	Angola	31	2001–2010
	Democratic Republic of Congo	22	2002–2007
Southern	Botswana	70	2001
	Malawi	46	2002
	Mozambique	101	2002–2004
	South Africa	1,031	1999–2009
	Zambia	150	1998–2008
	Zimbabwe	178	2007
East	Burundi	92	2002
	Ethiopia	102	1986–2003
	Kenya	39	1991–2007
	Tanzania	81	1997–2009
	Uganda	38	1990–2010

doi:10.1371/journal.pone.0041904.t001

Substitution saturation and likelihood mapping analyses

Substitution saturation was evaluated by plotting the estimated number of transitions and transversions against genetic distance for each pairwise comparison in our alignment of 1,981 HIV-1 subtype C *pol* sequences using DAMBE program [43]. The phylogenetic signal in the *pol* dataset was investigated with the likelihood mapping method [44] by analyzing 10,000 random quartets. Likelihood mapping was performed with TREE-PUZZLE program [45] using the online web platform Phylemon 2.0 [46].

Phylogenetic analysis

ML phylogenetic trees were inferred under the GTR+I+Γ_4 nucleotide substitution model, selected using the jModeltest program [47]. ML tree was reconstructed with PhyML program [48] using an online web server [49]. Heuristic tree search was performed using the SPR branch-swapping algorithm and the reliability of the obtained topology was estimated with the approximate likelihood-ratio test (*aLRT*) [50] based on the Shimodaira-Hasegawa-like procedure. The ML trees were visualized using the FigTree v1.3.1 program [51].

Characterization of intrasubtype C/C′ recombinant sequences

Putative intrasubtype C/C′ recombinant sequences in Ethiopia were identified by Bootscanning using Simplot version 3.5.1 [52], following the same procedure described by Pollakis *et al* [40]. Bootstrap values supporting branching with reference sequences were determined in Neighbor-Joining (NJ) trees constructed using the K2-P nucleotide substitution model, based on 100 resamplings, with a 300 bp sliding window moving in steps of 10 bases.

Analysis of spatiotemporal dispersion pattern

The evolutionary rate (μ, units are nucleotide substitutions per site per year, subst./site/year), the age of the most recent common ancestor (T_{mrca}, years), and the spatial dynamics of major subtype C clades from east Africa were jointly estimated using the Bayesian Markov Chain Monte Carlo (MCMC) approach implemented in the BEAST software package v1.6.2 [53,54]. Analyses were performed using the GTR+I+Γ_4 nucleotide substitution model, an uncorrelated Lognormal relaxed molecular clock model [55], a Bayesian Skyline coalescent tree prior [56], and a discrete phylogeographic model in which all possible reversible exchange rates between locations were equally likely [57]. Two separate MCMC chains were run for 4×10^8 generations and adequate chain mixing was checked by calculating the effective sample size (ESS) after excluding an initial 10% for each run using program TRACER v1.4 [58]. MCMC runs converged to almost identical values and combined estimates showed ESS values >200. Maximum clade credibility (MCC) trees were summarized from the posterior distribution of trees with TreeAnnotator and visualized with FigTree v1.3.1. Migratory events were summarized using the cross-platform SPREAD application [59].

Results

Phylogenetic analysis

A large dataset of HIV-1 subtype C *pol* sequences ($n=1{,}981$) downloaded from the Los Alamos HIV Database (http://hiv.lanl.gov) was used to characterize the relationship between subtype C sequences sampled from east, central and southern African countries. The transition/transversion *vs* divergence graphics showed that both type of nucleotide substitutions increase linearly with the genetic distance, with transitions being higher than transversions (Figure S1a), thus indicating no substitution saturation in our alignment. While, the likelihood-mapping analysis showed that most (90%) of the randomly chosen quartets from the HIV-1 subtype C alignment were equally distributed in the three corners of the likelihood map (Figure S1b), indicating a strong tree-like phylogenetic signal in the data. Both analyses indicate that the HIV-1 subtype C *pol* dataset used in this study contains enough evolutionary information for reliable phylogenetic and molecular clock inferences.

The ML phylogenetic analysis revealed that most (73%) subtype C sequences from east Africa branched within a highly supported (*aLRT* = 0.93) monophyletic cluster, here called C_{EA}, that contains sequences from all five east African countries analyzed (Figure 1). Notably, the C_{EA} clade comprises a minor proportion (9%) of the 54 sequences from central Africa, but none of the 1,576 sequences from southern Africa here included. A minor fraction (11%) of subtype C sequences from east Africa branched in a second well supported (*aLRT* = 0.94) monophyletic cluster that comprises sequences from Ethiopia, and corresponds to the so called Ethiopian-C′ (C'_{ET}) clade (Figure 1). The remaining subtype C east African sequences (16%) were distributed in several independent lineages of small size ($n \leq 5$ sequences) that were intermixed among strains from southern African countries (Figure 1).

The analysis of sequence distribution among clades by country of origin revealed three different patterns within east Africa represented by Burundi/Uganda, Ethiopia and Kenya/Tanzania (Figure 2a). All or most subtype C strains circulating in Burundi and Uganda belong to the major clade C_{EA}. Subtype C strains from Ethiopia, by contrast, were mainly distributed into clades C_{EA} (61%) and C'_{ET} (37%). Finally, about 64% of subtype C sequences from Kenya and 49% from Tanzania branched within the major clade C_{EA}, while the remaining sequences were distributed in the multiple minor clades of southern African origin. Such geographical variation in the prevalence of different subtype C clades could be also observed at a more local scale in Tanzania (Fig. 2b). In the Kagera and Mwanza regions (north), most (>70%) subtype C strains belong to the C_{EA} clade. In the Kilimanjaro region (northeast), sequences from both the C_{EA} and "southern African" clades reach a roughly similar prevalence. In the Mbeya region (southwest), only "southern African" clades were detected.

Migration pattern of HIV-1 C_{EA} clade

A closer inspection of the HIV-1 C_{EA} clade showed that sequences from Burundi occupies the most basal position in the clade (Figure S2), thus suggesting that Burundi was the most probable epicenter of dissemination of this subtype C lineage. The migration pattern of the C_{EA} lineage was reconstructed using a Bayesian statistical framework that allows ancestral reconstruction of the locations at the interior nodes of Bayesian tree while accommodating phylogenetic uncertainty. Sequences with no information about sampling date ($n = 2$), sequences with unexpectedly long branches in the phylogenetic analysis ($n = 10$), and Ethiopian sequences with evidence of intra-subtype recombination ($n = 8$, see below) were excluded from this analysis. This resulted in a final dataset of 236 sequences (Burundi = 92, Ethiopia = 47, Kenya = 24, Tanzania = 40, and Uganda = 33) sampled between 1990 and 2010.

The Bayesian MCC tree supports the hypothesis that the C_{EA} clade originated in Burundi (*PP* = 1) and was later exported to the other east African countries where it further spread, establishing new local epidemics (Figures 3 and 4). Estimation of viral movement among countries, obtained by counting the state changes along the tree nodes, points to the role of Burundi as the most important hub of dissemination of this subtype C lineage in east Africa, followed by Tanzania (Table 2). Several migration events of the lineage C_{EA} from Burundi to Ethiopia ($n = 4$), Kenya ($n = 5$), Tanzania ($n = 8$) and Uganda ($n = 8$) were detected, as well as from Tanzania to both Kenya ($n = 3$) and Uganda ($n = 7$). Importation of the C_{EA} lineage into Burundi from other east African countries, and viral exchanges between Ethiopia, Kenya and Uganda were seldom detected in our dataset.

The Bayesian analysis also supports an important phylogeographic subdivision within the C_{EA} lineage. Consistent with the ML topology (Figure S2), most subtype C sequences from Ethiopia, Kenya, Tanzania and Uganda branched in country-specific monophyletic sub-clusters that most probably ($PP \geq 0.93$) had a Burundian origin (Fig. 3). The C_{ET1} and C_{ET2} lineages, that correspond to the so called Ethiopian-C clade, comprise 44% of all Ethiopian sequences here included and were almost exclusively composed by sequences from this country. The C_{KE} and C_{UG} lineages comprise 33% and 37% of all sequences from Kenya and Uganda, respectively, and their circulation seems to be mainly restricted to those countries. Finally, the C_{TZ} lineage comprises 39% of all Tanzanian sequences analyzed and has also been disseminated to Kenya and Uganda. Both ML and Bayesian analyses further suggest that the C_{EA} clade branched in two major sub-clades: one composed by sequences from Burundi and lineages C_{ET1}, C_{ET2} and $C_{UG;}$ the other one composed by sequences from Burundi and lineages C_{KE} and C_{TZ}. The statistical support of such major sub-clades in Bayesian analysis, however, was not significant ($PP < 0.50$) and this observation should be interpreted with caution.

Time-scale of the HIV-1 C_{EA} clade

The median estimated evolutionary rate for the *pol* region of the C_{EA} clade was 1.8×10^{-3} (95% highest posterior density [HPD]: 1.1×10^{-3}–2.4×10^{-3}) subst/site/year, similar to that previously estimated for HIV-1 subtype C lineages circulating in South America [60] and southern Africa [6]. Importantly, the coefficient of rate variation was higher than zero (0.26 [95% HPD: 0.21–0.31]), thus demonstrating a significant variation of substitution rate among branches in the C_{EA} clade and supporting the use of a relaxed molecular clock model to reconstruct the time-scale of this lineage. According to this analysis the C_{EA} clade started to spread in Burundi at 1962 (95% HPD: 1942–1975), while major sub-clades C_{ET1}/C_{ET2}, C_{KE}, C_{TZ} and C_{UG} began to expand in Ethiopia, Kenya, Tanzania and Uganda, respectively, by the early 1970s (Figure 3).

Time-scale of the HIV-1 subtype C Ethiopian clades

The time-scale of the two major Ethiopian clades (C_{ET} and C'_{ET}) was also estimated by combining all sequences from this country in a single dataset and incorporating the posterior distribution of the substitution rate previously estimated for the C_{EA} lineage as an informative prior. This analysis resulted in a Bayesian MCC tree in which clades C_{ET} and C'_{ET} were poorly supported ($PP < 0.5$) and several strains branched outside those major clades (Figure S3). A careful exploration of Ethiopian sequences, revealed that some strains initially classified within clades C_{ET} ($n = 8$) or C'_{ET} ($n = 10$) and those strains that branched outside major Ethiopian clades ($n = 2$) are putative C/C′ intrasubtype recombinant viruses (Figure S3). When those viruses were excluded, the clades C_{ET} and C'_{ET} segregate in two highly supported ($PP > 0.9$) reciprocally monophyletic groups (Figure 5). According to this new Bayesian MCC tree, the median T_{mrca} was

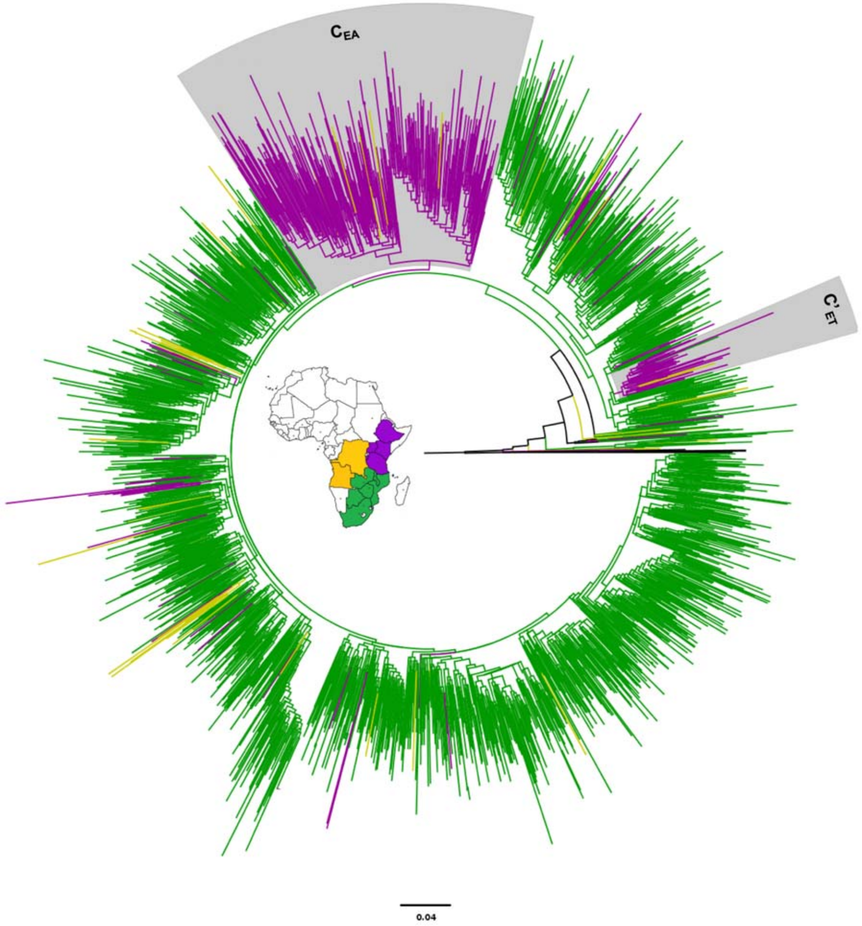

Figure 1. Maximum likelihood phylogenetic tree based on 1,981 HIV-1 subtype C *pol* (~1,000 pb) sequences. Sequences were sampled at different countries from the east ($n = 352$), central ($n = 53$) and southern ($n = 1{,}576$) African regions shown in Table 1. The color of branches represents the geographic region from where the subtype C sequences originated, according to the map given in the figure. The boxes highlight the position of the major east African subtype C lineages. The tree was rooted using HIV-1 subtype A1 and D reference sequences (black branches). Horizontal branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site.
doi:10.1371/journal.pone.0041904.g001

estimated at 1978 for clade C_{ET}, 1981 for sub-clade C_{ET1}, 1984 for sub-clade C_{ET2}, and 1981 for clade C'_{ET} (Figure 5).

Discussion

This study demonstrates a significant phylogeographic subdivision of HIV-1 subtype C strains circulating in the east respect to those circulating in the central and southern African regions, consistent with a recent study [8]. Most (73%) subtype C sequences from east Africa analyzed in this study branched within a highly supported monophyletic clade, here called C_{EA}, that comprise 100% of subtype C sequences from Burundi, 97% from Uganda, 64% from Kenya, 61% from Ethiopia, and 49% from Tanzania. This major east African clade also comprises a minor proportion (<10%) of sequences from central Africa, but no sequence from southern Africa, thus indicating that its circulation is mainly restricted to the east African region. Of note, the genealogies previously inferred for HIV-1 subtypes A and D also support a model of limited introduction of each subtype into east Africa, followed by a subsequent local expansion [61].

Our phylogeographic study suggests that the C_{EA} clade most probably originated in Burundi and after a period of local expansion, this viral lineage was disseminated at multiple times to Ethiopia, Kenya, Tanzania and Uganda, where it generated new local epidemics. Several introductions of the C_{EA} lineage from Tanzania into both Kenya and Uganda were also detected, while viral exchanges between Ethiopia, Kenya and Uganda were less frequent. Five major country-specific monophyletic sub-clusters were detected within the C_{EA} clade that comprise 44%, 33%, 37%, and 39% of all sequences from Ethiopia, Kenya, Uganda and Tanzania here included, respectively. Thus, despite frequent viral movement among east African countries, a significant proportion of subtype C infections in Ethiopia, Kenya, Tanzania and Uganda most likely resulted from the expansion of a few ancestral C_{EA} strains.

It has been suggested that interconnectivity between population centers was a critical factor in the spread of HIV-1 subtypes A and D across Africa [61]. The restricted circulation of the C_{EA} lineage in southern African countries is consistent with this model, considering the relative inaccessibility between the principal population centers of eastern and southern African regions. This model, however, is not consistent with the proposed role of Burundi as the main hub of dissemination of the C_{EA} clade in the region, as this small country is poorly interconnected to other east African countries. Previous studies have also shown a strongly supported phylogenetic relationship between subtype C sequences

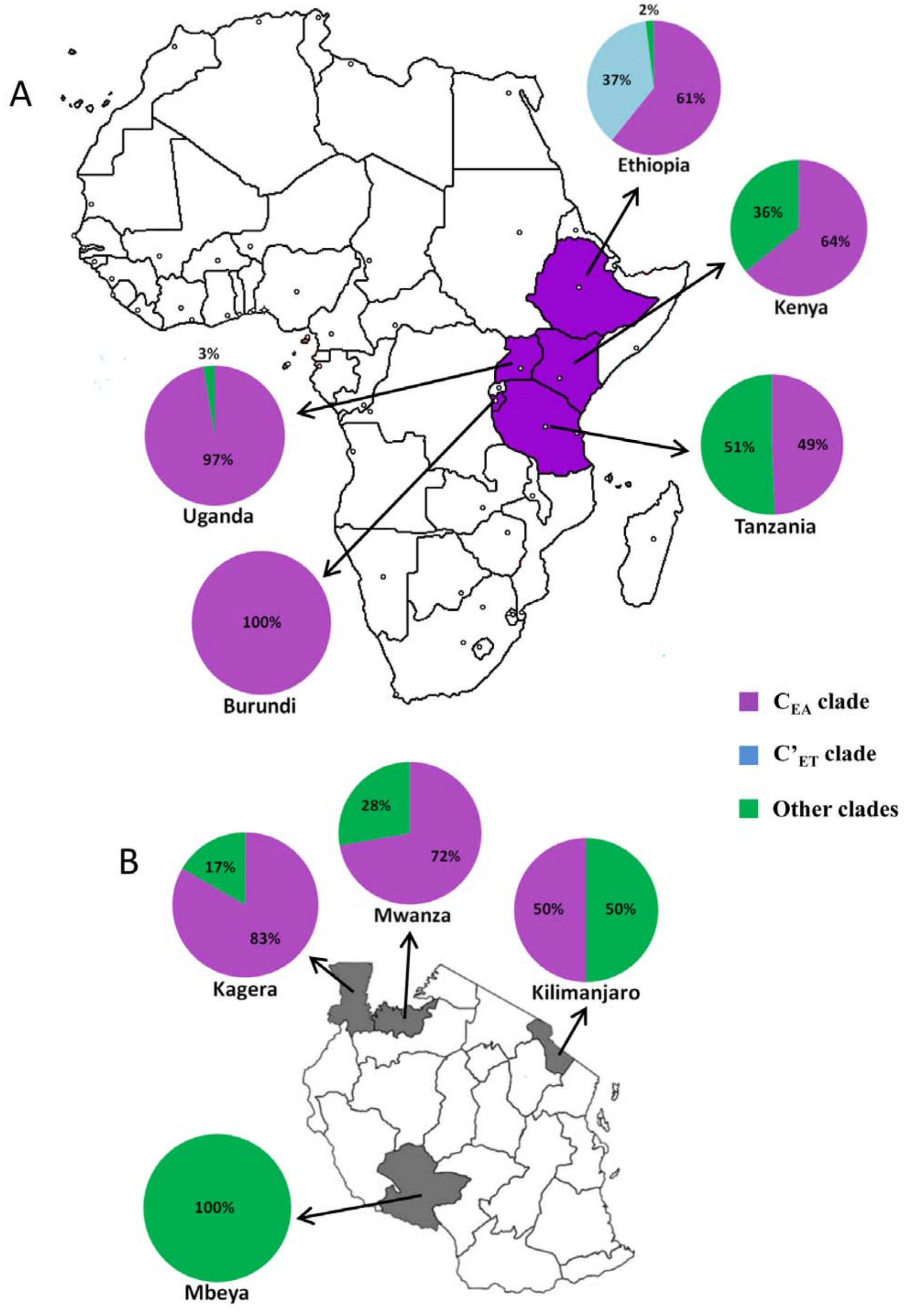

Figure 2. Geographic distribution of HIV-1 subtype C clades in east Africa. a) Map of Africa showing the frequency of distinct HIV-1 subtype C clades across the five countries from the east region here studied (Burundi, Ethiopia, Kenya, Uganda and Tanzania). b) Map of Tanzania showing the frequency of distinct HIV-1 subtype C clades across different country regions where patients included in the present study resided (Kagera, Mwanza, Kilimanjaro and Mbeya). The legend for the colors on graphics is shown on the right.
doi:10.1371/journal.pone.0041904.g002

from Brazil, the UK, Burundi and Kenya; thus indicating that the C_{EA} clade has also been disseminated to South America and Europe [60,62,63]. These evidences suggest that factors other than accessibility may have shaped the dissemination of the C_{EA} clade at both local and global scale.

Burundi has known many violent ethnic conflicts mainly since the 1960s that resulted in large migration flows. Two major civil conflicts that took place in Burundi in 1972 and 1993 generated especially large human movements with the former producing around 300,000 refugees and the latter producing about 687,000 [64]. Most refugees initially crossed the border of their country in the east, fleeing to neighboring Tanzania, followed by movement into other neighboring African countries and later to Europe and North America. It has been estimated that there are about 200,000 Burundians currently living in Tanzania, 18,000 in the Democratic Republic of the Congo, 4,000 in Uganda, 10,000 in the

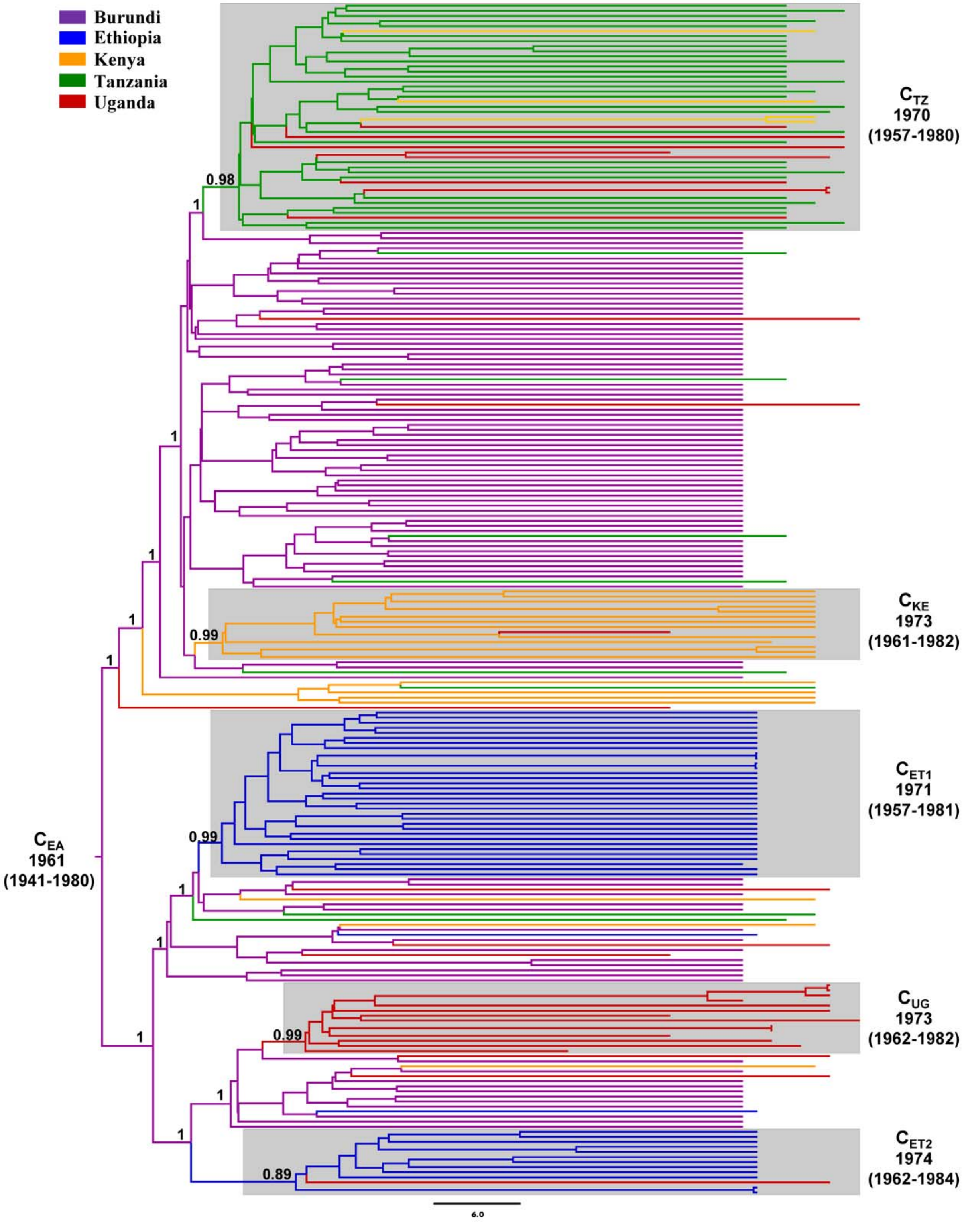

Figure 3. Time-scaled Bayesian MCC tree of the HIV-1 C_{EA} lineage. Branches are colored according to the most probable location state of their descendent nodes. The legend for the colors is shown on the left. The state posterior probability is indicated only at key nodes. The boxes highlight the position of the major country-specific sub-clades detected in our study. The median age (with 95% HPD interval in parentheses) of those country-specific sub-clades is shown. Horizontal branch lengths are drawn to scale with the bar at the bottom indicating years. The tree was automatically rooted under the assumption of a relaxed molecular clock.

doi:10.1371/journal.pone.0041904.g003

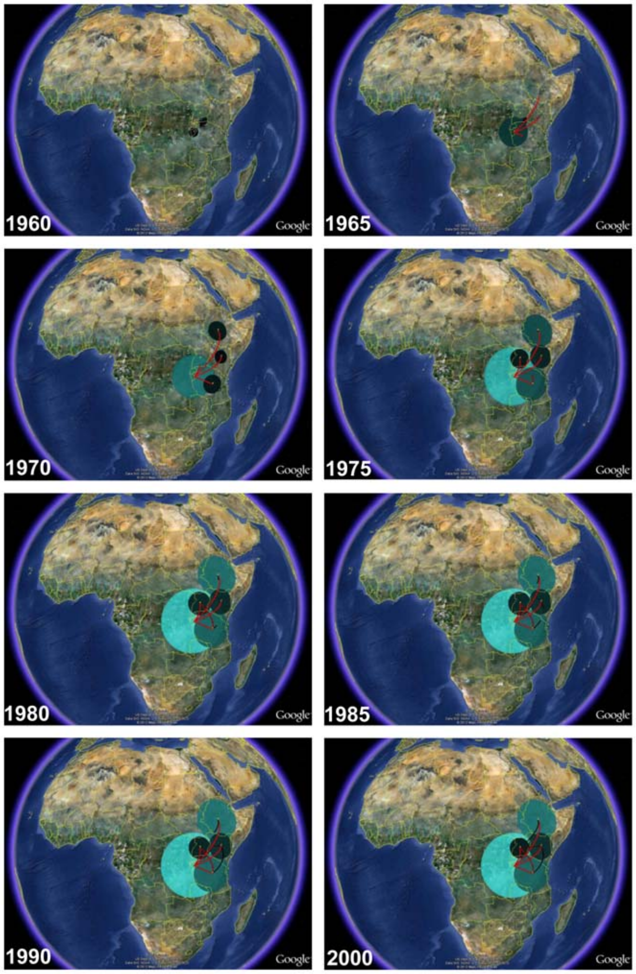

Figure 4. Spatiotemporal dynamic of HIV-1 C_{EA} clade dissemination in east Africa. We provide snapshots of the dispersal pattern for the years 1960, 1965, 1970, 1975, 1980, 1985, 1990 and 2000. Lines between locations represent branches in the Bayesian MCC tree along which location transition occurs. Location circle diameters are proportional to square root of the number of Bayesian MCC branches maintaining a particular location state at each time-point. The white-green color gradient informs the relative age of the transitions (older-recent). The maps are based on satellite pictures made available in Google™ Earth (http://earth.google.com).
doi:10.1371/journal.pone.0041904.g004

Table 2. Estimated number of migration events of HIV-1 C_{EA} clade among east African countries.

From/To	Burundi	Ethiopia	Kenya	Tanzania	Uganda
Burundi	-	4	5	8	8
Ethiopia	0	-	0	0	1
Kenya	0	0	-	1	1
Tanzania	0	0	3	-	7
Uganda	0	0	0	0	-

doi:10.1371/journal.pone.0041904.t002

European Union, and about 3,000 in the USA and Canada [64]. The molecular clock analysis clade traced the origin of the C_{EA} lineage in Burundi to the early 1960s, while the onset date of the major sub-clades circulating in Ethiopia, Kenya, Tanzania and Uganda was estimated at around the early 1970s, coinciding with the first large Burundian migration flow. These analyses support the notion that the Burundian migration flow occurring in 1972 may have played a fundamental role in the regional and international dissemination of the C_{EA} clade.

While subtype C epidemic in Burundi and Uganda is largely dominated by the C_{EA} clade, a second major subtype C lineage is also circulating in Ethiopia. Our results showed that the two Ethiopian lineages previously designated C and C' [13], resulted from independent founder strains originated in the eastern and southern African regions, respectively, and further confirmed the circulation of intra-subtype C/C' recombinants in Ethiopia [40]. The prevalence of C/C' recombinant viruses estimated in our dataset (20%) was equal to the percentage found in the general Ethiopian population [40]. The onset date of Ethiopian clades C and C' was dated to between the early 1970s and the early 1980s; consistent with previous estimations [65,66,67].

A large collection of minor subtype C lineages of southern African origin were detected in Kenya and Tanzania, which together represent 36% and 51% of sequences from those countries here analyzed, respectively. These lineages seem to have a more restricted expansion than the C_{EA} clade, although they were particularly prevalent (100%) in southwest Tanzania (Mbeya region), close to Zambia and Malawi. The co-circulation of subtype C sequences from both east and southern African origin in Tanzania is consistent with its intermediate geographical position between eastern and southern countries. It is unclear whether subtype C clades of southern African origin detected in Kenya were introduced from Tanzania and/or directly from southern Africa.

It is also unclear the relevance of these findings for HIV-1 vaccine design. Possible correlations of distinct HIV-1 subtype C clades with differential susceptibility to neutralizing antibody and/or cellular immune responses should be explored to justify the selection of vaccines incorporating one or multiple immunogens derived from major African subtype C clades [8]. It is also uncertain whether distinct subtype C lineages may possess different biological properties that affect disease progression and viral transmission. A recent study conducted in Ethiopia showed that infection with clade C_{ET} is associated with initially lower HIV-1 RNA plasma loads but more rapid onset of disease than infections with clade C'_{ET} [68]. The authors proposed that the clade C_{ET} may be less efficiently transmitted than clade C'_{ET}, which is consistent with epidemiological evidence that show that the strain C'_{ET} has gained ground and surpassed the clade C_{ET} over time [40,68]. New studies are necessary to determine if subtype C lineages of east African origin are less transmissible than those originated in southern Africa.

In conclusion, the results presented here point to the existence of a HIV-1 subtype C lineage characteristic of east Africa, which accounts for $>70\%$ of subtype C infections in this African region. This lineage probably emerged in Burundi in the 1960s and about 10 years later spread to Ethiopia, Kenya, Uganda and Tanzania, where it disseminated establishing new local epidemics. The subtype C epidemics in Ethiopia, Kenya and Tanzania also resulted from the introduction and dissemination of additional lineages of southern African origin. The explanation for the pattern of spread of the HIV-1 subtype C epidemic in east Africa is probably multifactorial and includes founder effects, massive migration between countries as a consequence of ethnic conflicts and geographical proximity.

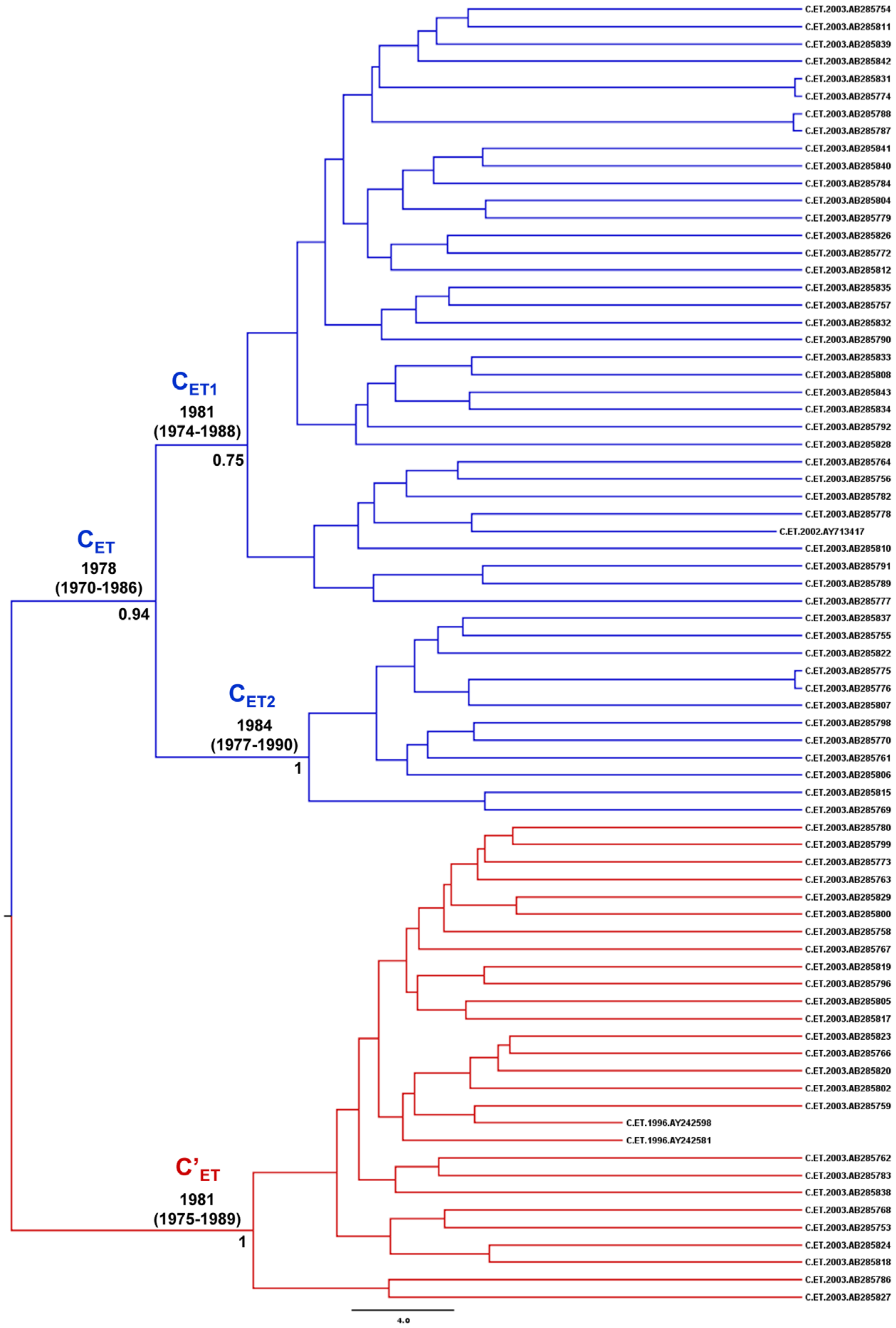
C_{ET1}
1981
(1974-1988)
0.75
C_{ET}
1978
(1970-1986)
0.94
C_{ET2}
1984
(1977-1990)
1
C'_{ET}
1981
(1975-1989)
1
C.ET.2003.AB285754
C.ET.2003.AB285811
C.ET.2003.AB285839
C.ET.2003.AB285842
C.ET.2003.AB285831
C.ET.2003.AB285774
C.ET.2003.AB285788
C.ET.2003.AB285787
C.ET.2003.AB285841
C.ET.2003.AB285840
C.ET.2003.AB285784
C.ET.2003.AB285804
C.ET.2003.AB285779
C.ET.2003.AB285826
C.ET.2003.AB285772
C.ET.2003.AB285812
C.ET.2003.AB285835
C.ET.2003.AB285757
C.ET.2003.AB285832
C.ET.2003.AB285790
C.ET.2003.AB285833
C.ET.2003.AB285808
C.ET.2003.AB285843
C.ET.2003.AB285834
C.ET.2003.AB285792
C.ET.2003.AB285828
C.ET.2003.AB285764
C.ET.2003.AB285756
C.ET.2003.AB285782
C.ET.2003.AB285778
C.ET.2002.AY713417
C.ET.2003.AB285810
C.ET.2003.AB285791
C.ET.2003.AB285789
C.ET.2003.AB285777
C.ET.2003.AB285837
C.ET.2003.AB285755
C.ET.2003.AB285822
C.ET.2003.AB285775
C.ET.2003.AB285776
C.ET.2003.AB285807
C.ET.2003.AB285798
C.ET.2003.AB285770
C.ET.2003.AB285761
C.ET.2003.AB285806
C.ET.2003.AB285815
C.ET.2003.AB285769
C.ET.2003.AB285780
C.ET.2003.AB285799
C.ET.2003.AB285773
C.ET.2003.AB285763
C.ET.2003.AB285829
C.ET.2003.AB285800
C.ET.2003.AB285758
C.ET.2003.AB285767
C.ET.2003.AB285819
C.ET.2003.AB285796
C.ET.2003.AB285805
C.ET.2003.AB285817
C.ET.2003.AB285823
C.ET.2003.AB285766
C.ET.2003.AB285820
C.ET.2003.AB285802
C.ET.2003.AB285759
C.ET.1996.AY242598
C.ET.1996.AY242581
C.ET.2003.AB285762
C.ET.2003.AB285783
C.ET.2003.AB285838
C.ET.2003.AB285768
C.ET.2003.AB285753
C.ET.2003.AB285824
C.ET.2003.AB285818
C.ET.2003.AB285786
C.ET.2003.AB285827
4.0

Figure 5. Time-scaled Bayesian MCC tree of major Ethiopian HIV-1 subtype C lineages. MCC tree was obtained after exclusion of putative C/C′ intrasubtype recombinant sequences. Branches are colored according to the initial clade assignment of each sequence based on ML analysis: C_{ET} (blue) and C'_{ET} (red). The *PP* support and the median age (with 95% HPD interval in parentheses) are indicated only at key nodes. Horizontal branch lengths are drawn to scale with the scale at the bottom indicating years. The tree was automatically rooted under the assumption of a relaxed molecular clock.
doi:10.1371/journal.pone.0041904.g005

Supporting Information

Figure S1 Substitution saturation and likelihood mapping analyses. a) Transition (blue line) and transversion (green line) versus divergence plot for the HIV-1 subtype C *pol* dataset. b) Likelihood mapping of 10,000 random quarters selected from the HIV-1 subtype C *pol* dataset. Distribution (left triangle) and percentage (right triangle) of dots plotted in each region of the map. Each dot represents the likelihoods of the three possible tree topologies for a set of four sequences (quartets) selected randomly from the dataset. The dots localized on the vertices, in the centre and on the laterals represent the tree-like, the star-like and the network-like phylogenetic signals, respectively.
(PPT)

Figure S2 Close view of the HIV-1 C_{EA} lineage despited in Figure 1. The color of branches represents the country from where the sequence originated, according to the legend shown on the left. The boxes highlight the position of the major country-specific sub-clades detected in our study. The aLRT support values are indicated only at key nodes.
(PPTX)

Figure S3 Time-scaled Bayesian MCC tree of HIV-1 subtype C *pol* sequences from Ethiopia. Branches are colored according to the initial clade assignment of each sequence based on ML analysis: C_{ET} (blue), C'_{ET} (red), other clades (green). The *PP* support is indicated only at key nodes. Positions of the putative interclade C/C′ recombinant sequences are marked with asterisks. Horizontal branch lengths are drawn to scale with the scale at the bottom indicating years. The tree was automatically rooted under the assumption of a relaxed molecular clock. Representative bootscanning plots of some putative C/C′ intrasubtype recombinant sequences are depicted on the right. Query sequences were compared to reference sequences of HIV-1 clades A1 (AB253429), D (AY371157), C_{ET} (AY242589), and C'_{ET} (AY242581).
(PPTX)

Acknowledgments

We wish to thank Dr. Vera Bongertz for critical review of the manuscript and anonymous reviewers for comments.

Author Contributions

Conceived and designed the experiments: GB. Performed the experiments: GB EOD. Analyzed the data: GB EOD. Contributed reagents/materials/analysis tools: GB EOD. Wrote the paper: GB EOD.

References

1. Hemelaar J, Gouws E, Ghys PD, Osmanov S (2011) Global trends in molecular epidemiology of HIV-1 during 2000–2007. Aids 25: 679–689.
2. Parreira R, Piedade J, Domingues A, Lobao D, Santos M, et al. (2006) Genetic characterization of human immunodeficiency virus type 1 from Beira, Mozambique. Microbes Infect 8: 2442–2451.
3. Bredell H, Martin DP, Van Harmelen J, Varsani A, Sheppard HW, et al. (2007) HIV type 1 subtype C gag and nef diversity in Southern Africa. AIDS Res Hum Retroviruses 23: 477–481.
4. Deho L, Walwema R, Cappelletti A, Sukati H, Sibandze D, et al. (2008) Subtype assignment and phylogenetic analysis of HIV type 1 strains in patients from Swaziland. AIDS Res Hum Retroviruses 24: 323–325.
5. Lahuerta M, Aparicio E, Bardaji A, Marco S, Sacarlal J, et al. (2008) Rapid spread and genetic diversification of HIV type 1 subtype C in a rural area of southern Mozambique. AIDS Res Hum Retroviruses 24: 327–335.
6. Dalai SC, de Oliveira T, Harkins GW, Kassaye SG, Lint J, et al. (2009) Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. Aids 23: 2523–2532.
7. Novitsky V, Wang R, Lagakos S, Essex M (2010) HIV-1 Subtype C Phylodynamics in the Global Epidemic. Viruses 2: 33–54.
8. Thomson MM, Fernandez-Garcia A (2011) Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning. Virology 415: 30–38.
9. Koch N, Ndihokubwayo JB, Yahi N, Tourres C, Fantini J, et al. (2001) Genetic analysis of hiv type 1 strains in Bujumbura (burundi): predominance of subtype c variant. AIDS Res Hum Retroviruses 17: 269–273.
10. Vidal N, Niyongabo T, Nduwimana J, Butel C, Ndayiragije A, et al. (2007) HIV type 1 diversity and antiretroviral drug resistance mutations in Burundi. AIDS Res Hum Retroviruses 23: 175–180.
11. Maslin J, Rogier C, Berger F, Khamil MA, Mattera D, et al. (2005) Epidemiology and genetic characterization of HIV-1 isolates in the general population of Djibouti (Horn of Africa). J Acquir Immune Defic Syndr 39: 129–132.
12. Abebe A, Kuiken CL, Goudsmit J, Valk M, Messele T, et al. (1997) HIV type 1 subtype C in Addis Ababa, Ethiopia. AIDS Res Hum Retroviruses 13: 1071–1075.
13. Abebe A, Pollakis G, Fontanet AL, Fisseha B, Tegbaru B, et al. (2000) Identification of a genetic subcluster of HIV type 1 subtype C (C′) widespread in Ethiopia. AIDS Res Hum Retroviruses 16: 1909–1914.
14. Hussein M, Abebe A, Pollakis G, Brouwer M, Petros B, et al. (2000) HIV-1 subtype C in commerical sex workers in Addis Ababa, Ethiopia. J Acquir Immune Defic Syndr 23: 120–127.
15. Kassu A, Fujino M, Matsuda M, Nishizawa M, Ota F, et al. (2007) Molecular epidemiology of HIV type 1 in treatment-naive patients in north Ethiopia. AIDS Res Hum Retroviruses 23: 564–568.
16. Renjifo B, Chaplin B, Mwakagile D, Shah P, Vannberg F, et al. (1998) Epidemic expansion of HIV type 1 subtype C and recombinant genotypes in Tanzania. AIDS Res Hum Retroviruses 14: 635–638.
17. Kiwelu IE, Renjifo B, Chaplin B, Sam N, Nkya WM, et al. (2003) HIV type 1 subtypes among bar and hotel workers in Moshi, Tanzania. AIDS Res Hum Retroviruses 19: 57–64.
18. Herbinger KH, Gerhardt M, Piyasirisilp S, Mloka D, Arroyo MA, et al. (2006) Frequency of HIV type 1 dual infection and HIV diversity: analysis of low- and high-risk populations in Mbeya Region, Tanzania. AIDS Res Hum Retroviruses 22: 599–606.
19. Nyombi BM, Kristiansen KI, Bjune G, Muller F, Holm-Hansen C (2008) Diversity of human immunodeficiency virus type 1 subtypes in Kagera and Kilimanjaro regions, Tanzania. AIDS Res Hum Retroviruses 24: 761–769.
20. Mosha F, Urassa W, Aboud S, Lyamuya E, Sandstrom E, et al. (2011) Prevalence of genotypic resistance to antiretroviral drugs in treatment-naive youths infected with diverse HIV type 1 subtypes and recombinant forms in Dar es Salaam, Tanzania. AIDS Res Hum Retroviruses 27: 377–382.
21. Servais J, Lambert C, Karita E, Vanhove D, Fischer A, et al. (2004) HIV type 1 pol gene diversity and archived nevirapine resistance mutation in pregnant women in Rwanda. AIDS Res Hum Retroviruses 20: 279–283.
22. Brennan CA, Lund JK, Golden A, Yamaguchi J, Vallari AS, et al. (1997) Serologic and phylogenetic characterization of HIV-1 subtypes in Uganda. Aids 11: 1823–1832.
23. Rayfield MA, Downing RG, Baggs J, Hu DJ, Pieniazek D, et al. (1998) A molecular epidemiologic survey of HIV in Uganda. HIV Variant Working Group. Aids 12: 521–527.
24. Hu DJ, Baggs J, Downing RG, Pieniazek D, Dorn J, et al. (2000) Predominance of HIV-1 subtype A and D infections in Uganda. Emerg Infect Dis 6: 609–615.
25. Gale CV, Yirrell DL, Campbell E, Van der Paal L, Grosskurth H, et al. (2006) Genotypic variation in the pol gene of HIV type 1 in an antiretroviral treatment-naive population in rural southwestern Uganda. AIDS Res Hum Retroviruses 22: 985–992.
26. Herbeck JT, Lyagoba F, Moore SW, Shindo N, Biryahwaho B, et al. (2007) Prevalence and genetic diversity of HIV type 1 subtypes A and D in women attending antenatal clinics in Uganda. AIDS Res Hum Retroviruses 23: 755–760.
27. Ssemwanga D, Ndembi N, Lyagoba F, Bukenya J, Seeley J, et al. (2011) HIV Type 1 Subtype Distribution, Multiple Infections, Sexual Networks, and

Partnership Histories in Female Sex Workers in Kampala, Uganda. AIDS Res Hum Retroviruses.

28. Hamers RL, Wallis CL, Kityo C, Siwale M, Mandaliya K, et al. (2011) HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. Lancet Infect Dis 11: 750–759.
29. Yang C, Li M, Shi YP, Winter J, van Eijk AM, et al. (2004) Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, western Kenya. AIDS Res Hum Retroviruses 20: 565–574.
30. Oyaro M, Mbithi J, Oyugi F, Laten A, Anzala O, et al. (2011) Molecular characterization of HIV type 1 among HIV-infected respondents in a cohort being prepared for HIV Phase III vaccine clinical trials, Western Kenya. AIDS Res Hum Retroviruses 27: 257–264.
31. Nyagaka B, Kiptoo MK, Lihana RW, Khamadi SA, Makokha EP, et al. (2011) HIV Type 1 gag Genetic Diversity Among Antenatal Clinic Attendees in North Rift Valley, Kenya. AIDS Res Hum Retroviruses.
32. Rainwater S, DeVange S, Sagar M, Ndinya-Achola J, Mandaliya K, et al. (2005) No evidence for rapid subtype C spread within an epidemic in which multiple subtypes and intersubtype recombinants circulate. AIDS Res Hum Retroviruses 21: 1060–1065.
33. Khamadi SA, Lihana RW, Osman S, Mwangi J, Muriuki J, et al. (2009) Genetic diversity of HIV type 1 along the coastal strip of Kenya. AIDS Res Hum Retroviruses 25: 919–923.
34. Hue S, Hassan AS, Nabwera H, Sanders EJ, Pillay D, et al. (2012) HIV Type 1 in a Rural Coastal Town in Kenya Shows Multiple Introductions with Many Subtypes and Much Recombination. AIDS Res Hum Retroviruses 28: 220–224.
35. Neilson JR, John GC, Carr JK, Lewis P, Kreiss JK, et al. (1999) Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. J Virol 73: 4393–4403.
36. Lihana RW, Khamadi SA, Kiptoo MK, Kinyua JG, Lagat N, et al. (2006) HIV type 1 subtypes among STI patients in Nairobi: a genotypic study based on partial pol gene sequencing. AIDS Res Hum Retroviruses 22: 1172–1177.
37. Kageha S, Lihana RW, Okoth V, Mwau M, Okoth FA, et al. (2012) HIV Type 1 Subtype Surveillance in Central Kenya. AIDS Res Hum Retroviruses 28: 228–231.
38. Khamadi SA, Ochieng W, Lihana RW, Kinyua J, Muriuki J, et al. (2005) HIV type 1 subtypes in circulation in northern Kenya. AIDS Res Hum Retroviruses 21: 810–814.
39. Khamadi SA, Lihana RW, Mwaniki DL, Kinyua J, Lagat N, et al. (2008) HIV type 1 genetic diversity in Moyale, Mandera, and Turkana based on env-C2-V3 sequences. AIDS Res Hum Retroviruses 24: 1561–1564.
40. Pollakis G, Abebe A, Kliphuis A, De Wit TF, Fisseha B, et al. (2003) Recombination of HIV type 1C (C′/C″) in Ethiopia: possible link of EthHIV-1C′ to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. AIDS Res Hum Retroviruses 19: 999–1008.
41. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, et al. (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics 21: 3797–3800.
42. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.
43. Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92: 371–373.
44. Strimmer K, von Haeseler A (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A 94: 6815–6819.
45. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
46. Sanchez R, Serra F, Tarraga J, Medina I, Carbonell J, et al. (2011) Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Res 39: W470–474.
47. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25: 1253–1256.
48. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
49. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33: W557–559.
50. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol 55: 539–552.
51. Rambaut A (2009) FigTree v1.3.1: Tree Figure Drawing Tool. Available from http://treebioedacuk/software/figtree/. Accessed 2012 April 20.
52. Ray S Simplot v2.5.0. Available from: http://sray.med.som.jhmi.edu/SCRoftware/simplot/. Accessed 2012 April 20.
53. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161: 1307–1320.
54. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.
55. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4: e88.
56. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–1192.
57. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5: e1000520.
58. Rambaut A, Drummond A (2007) Tracer v1.4. Available from http://beastbioedacuk/Tracer. Accessed 2012 April 20.
59. Bielejec F, Rambaut A, Suchard MA, Lemey P (2011) SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. Bioinformatics 27: 2910–2912.
60. Bello G, Passaes CP, Guimaraes ML, Lorete RS, Matos Almeida SE, et al. (2008) Origin and evolutionary history of HIV-1 subtype C in Brazil. AIDS 22: 1993–2000.
61. Gray RR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, et al. (2009) Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. Aids 23: F9–F17.
62. de Oliveira T, Pillay D, Gifford RJ (2010) The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. PLoS ONE 5: e9311.
63. Fontella R, Soares MA, Schrago CG (2008) On the origin of HIV-1 subtype C in South America. AIDS 22: 2001–2011.
64. Fransen S, Ong'ayo A (2010) Migration in Burundi: History, Current Trends, and Future Prospects. Paper Series: Migration and Development Country Profiles Maastricht: Maastricht Graduate School of Governance Available from http://mgsog.merit.unu.edu/ISacademie/docs/CR_burundi.pdf. Accessed 2012 April 20.
65. Abebe A, Lukashov VV, Pollakis G, Kliphuis A, Fontanet AL, et al. (2001) Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. AIDS 15: 1555–1561.
66. Abebe A, Lukashov VV, Rinke De Wit TF, Fisseha B, Tegbaru B, et al. (2001) Timing of the introduction into Ethiopia of subcluster C′ of HIV type 1 subtype C. AIDS Res Hum Retroviruses 17: 657–661.
67. Tully DC, Wood C (2010) Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. Aids 24: 1577–1582.
68. Ayele W, Mekonnen Y, Messele T, Mengistu Y, Tsegaye A, et al. (2010) Differences in HIV type 1 RNA plasma load profile of closely related cocirculating Ethiopian subtype C strains: C and C′. AIDS Res Hum Retroviruses 26: 805–813.