# Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia

**Damien C. Tully and Charles Wood**

**Objective:** To reconstruct the onset date and evolutionary history of the HIV-1 subtype C epidemic in Ethiopia – one of the earliest recorded subtype C epidemics in the world.

**Design:** HIV-1 C *env* sequences with a known sampling year isolated from HIV-1 positive patients from Ethiopia between 1984 and 2003.

**Methods:** Evolutionary parameters including origin and demographic growth patterns were estimated using a Bayesian coalescent-based approach under either strict or relaxed molecular clock models.

**Results:** Bayesian evolutionary analysis indicated a most recent common ancestor date of 1965 with three distinct epidemic growth phases. Regression analysis of root-to-tip distances revealed a highly similar estimate for the origin of the clade. In addition, we reveal that the HIV-1C epidemic in Ethiopia has grown at a faster rate than the epidemic of subtype C in sub-Saharan Africa.

**Conclusion:** Reconstruction of the epidemic history in Ethiopia revealed that subtype C likely originated from either a single lineage or multiple descendents in the late 1960s or early 1970s where it grew exponentially throughout the mid-1970s and early 1980s, corresponding to a wave of urbanization and migration. In light of these findings, we suggest that subtype C strains were circulating at least a decade before previous estimates and the first recognition of symptomatic patients in Ethiopia. The timing of the Ethiopian epidemic is also in agreement with similar HIV-1 epidemics in sub-Saharan Africa. 

*AIDS* 2010, **24**:1577–1582

**Keywords: coalescent, Ethiopia, evolution, HIV-1, subtype C**

## Introduction

HIV-1 infections in Ethiopia were first documented in a young man and woman in 1984 in Addis Ababa when serum samples of 167 hospitalized patients with Bell's palsy were tested for anti-HIV-1 antibodies [1]. Subsequently, the first clinically overt case of AIDS was diagnosed in 1986 [2]. By the late 1980s a high prevalence of HIV-1 estimated at 17% was detected among commercial sex workers residing along the main trading roads and among long distance truck drivers [3,4]. The explosion of HIV into a major AIDS epidemic is almost solely attributed to subtype C with transmission largely sustained through heterosexual contact and to a lesser extent mother-to-child transmission. Phylogenetic analyses have also revealed that two genetically distinct viruses co-circulate in similar prevalence in all geographic regions and risk populations [5,6].

A timescale for the introduction of subtype C in Ethiopia has been proposed in the early 1980s (1980–1984) from the regression analysis of *env* gp120 V3 sequences [7,8].

Nebraska Center for Virology and School of Biological Sciences, University of Nebraska-Lincoln, Nebraska, USA.
Correspondence to Dr Damien C. Tully, Nebraska Center for Virology, School of Biological Sciences, University of Nebraska Lincoln, P.O. Box 830666, Lincoln, NE 68583-0900, USA.
Tel: +1 402 472 4559; fax: +1 402 472 3323; e-mail: dtully2@unl.edu
Received: 28 January 2010; revised: 1 March 2010; accepted: 10 March 2010.

DOI:10.1097/QAD.0b013e32833999e1

This implies a very quick fuse for the expansion of HIV-1 in Ethiopia, indicating that shortly after the introduction of the virus the first HIV/AIDS cases were registered. In contrast the HIV-1 epidemics in sub-Saharan Africa and the Americas are reminiscent of a slow fuse epidemic [9–11]. Using a combination of phylogenetic analyses and a Bayesian coalescent-based approach, we estimate with more precision the timescale of the epidemic and present new information on the epidemic growth patterns of HIV-1 subtype C. Our results suggest that subtype C was circulating within Ethiopia for a sustained period of time before its initial detection.

## Materials and methods

### Sequence collection and phylogenetic analyses

HIV-1 subtype C sequences belonging to Ethiopia with known collection dates were retrieved from the Los Alamos HIV Database (http://hiv.lanl.gov/content/index) [12]. In order to improve the accuracy of phylogenetic inference we excluded previously determined recombinant sequences, multiple sequences from single individuals and sequences containing frame-shift mutations. After the filtering of such data the only available gene suitable for coalescent analysis was the *env*. As a result, 119 *env* sequences with known collection dates between 1984 and 2003 were retrieved, all from Addis Ababa and its vicinity. The GenBank IDs of all sequences used in this study are provided as supplementary information (Table S1, http://links.lww.com/QAD/A22).

Sequences were aligned using CLUSTAL X and manually edited for optimization [13]. All sequences were confirmed as being subtype C using the REGA subtyping tool version 2.0 [14]. To test for the presence of recombination, sequences were screened using the pairwise homoplasy index (PHI) test [15,16]. This powerful PHI method identifies the likelihood of recombination within a set of aligned sequences with a low-false positive rate [15]. A PHI score with a $P$-value $< 0.05$ shows with significance that recombination occurs in the data set. No significant recombination was observed ($P = 0.99$). Although unidentified intra-subtype recombination might increase the variance of dating estimates, it is unlikely to bias the dates in one direction or the other in an exponentially growing population [17].

### Bayesian Markov Chain Monte Carlo evolutionary analyses

Investigation of the evolutionary history [rate of nucleotide substitution, mode and rate of population growth, and time to the most recent common ancestor (tMRCA)] was estimated using a Bayesian Markov Chain Monte Carlo (MCMC) method implemented in the BEAST v1.4.8 program [18]. Four different coalescent priors were investigated: constant population size, exponential and logistic growth and the nonparametric Bayesian skyline plot (BSP) [19] assuming either a constant (strict) or a variable (relaxed) molecular clock [20]. In each case, we employed a HKY85 nucleotide substitution model with two partitions in the codon positions [21]. MCMC chains were run for sufficient time (50–100 million generations) to achieve convergence (assessed using the TRACER program; http://tree.bio.ed.ac.uk/software/tracer), with uncertainty in parameter estimates reflected in the 95% highest probability density (HPD). The results of multiple runs were then combined using the LogCombiner program (http://beast.bio.ed.ac.uk) after the removal of an appropriate burnin. The Maximum Clade Credibility (MCC) tree across all plausible trees was computed from the sampled posterior phylogenies obtained from BEAST using the TreeAnnotator program, with the first 10% trees removed as burnin. The Bayes Factor, which is the ratio of the marginal likelihoods (with respect to the prior) of the two models, was used to determine which clock and demographic model best fits the data [22,23].

To assess the reliability of our estimates for the substitution rate and the tMRCA and to determine the extent of temporal structure in the *env* sequence data, we performed a regression analysis of tree root-to-tip genetic distance against sampling date using the program Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen/ [24]) based on maximum likelihood trees estimated from PAUP* [25].

## Results

### Rates and dates of HIV-1 subtype C Ethiopian epidemic

To explore the origin and time course of the subtype C epidemic in Ethiopia, we conducted a detailed Bayesian MCMC phylogenetic analysis. Analyses were performed with a variety of different population models under both a strict and a relaxed clock model. For all model comparisons, the Bayes factor analysis favored the relaxed molecular clock over a strict clock (ln Bayes Factor $> 81$) and models allowing for population growth outperformed a constant population size model (data not shown; available on request). Inspection of the median estimates for the coefficient of variation parameter revealed significant lineage rate variation in the tree irrespective of the evolutionary model employed with values of 0.726, 0.476, 0.484 and 0.473 for the constant, exponential, logistic and Bayesian skyline using relaxed clock analyses respectively. Due to this high variation, the use of strict clock models with this data would be inappropriate and would probably yield misleadingly estimates with regard to both timing and substitution rates.

**Table 1. Bayesian estimates of population dynamics and evolutionary parameters for HIV-1 subtype C in Ethiopia.**

| Parameter | Estimates |
|---|---|
| Sample size | 119 |
| Sample date range | 1984–2003 |
| Best fit demographic model | Exponential growth (relaxed molecular clock) |
| MCMC chain length | 50 000 000 |
| Mean substitution rate[a] | $5.70 \times 10^{-3}$ ($4.24 \times 10^{-3}$–$7.33 \times 10^{-3}$) |
| MRCA (year) | 37.44 (29.13–43.09) |
| Actual time | 1965.56 (1959.91–1973.87) |
| Coefficient of variation | 0.479 (0.373–0.596) |
| Mean population growth rate (year$^{-1}$)[b] | 0.406 (0.289–0.536) |
| Mean epidemic doubling time (year) | 1.71 (1.29–2.40) |

95% HPD are indicated in parenthesis.
[a]Number of substitutions per site per year.
[b]Number of new infections per individual per year.

The estimated nucleotide substitution rates and tMRCA dates for the Ethiopian *env* sequences obtained under the four different evolutionary models were highly consistent (not shown). The mean rate of $5.70 \times 10^{-3}$ nucleotide substitutions per site per year produced an average estimate for the date of origin of the HIV-1C *env* sequences in the year 1965.56 (95% HPD, 1959.91–1973.87) (Table 1).

Similar estimates for the tMRCA were observed using simple root-to-tip regression (Fig. 1c) where the $\times$ intercept represents the time to common ancestry whereas the slope of the regression is an estimate of the rate of evolution. Results for the *env* gene indicate a root age of 1966 with an evolutionary rate corresponding to $9.20 \times 10^{-3}$ nucleotide substitutions per site per year.

To determine the rate of increase in the effective number of infections with time since the initial introduction of the virus the exponential growth model was used to determine the growth rate. The mean parametric estimates of $r$, where $r$ is the population growth rate estimated in BEAST, was 0.406 new infections per individual per year (95% HPD, 0.29–0.54 infections/year) (Table 1). This equates to a mean epidemic doubling time of 1.71 years (1.29–2.40) using the relation $\lambda = \ln(2)/r$. This mean estimated growth rate is in comparison slightly lower than previously obtained for this subtype in the Brazil epidemic (~0.6–0.8/year) [26,27] but higher than the previously estimated doubling time of 2.4 years for the subtype C epidemic in sub-Saharan Africa [28]. The use of the exponential model is in correspondence with the best parametric model selected by the Bayes Factor analysis. However, a discrepancy exists between this result and the BSP, which suggests logistic growth (see below). One explanation for this is that the model with fewer parameters is preferred to avoid overparameterization. Nevertheless, similar growth rates were observed using a logistic model of growth (not shown).

### Population growth dynamics

To track the estimation of effective population size through time directly from sequence data the BSP was utilized. Reconstruction of the demographic history from the BSP identified a low number of effective infections during the initial epidemic and an exponential growth phase during the late 1970s/early 1980s followed by a plateau during the mid-1980s/early 1990s onwards (Fig. 1b). This period of exponential growth coincides with the birth and rapid diversification of lineages. The Bayesian genealogies for *env* showed two well supported monophyletic clades within subtype C for Ethiopia previously designated as main C group and subcluster C′ ($P = 0.82$ and 0.91 for each clade respectively) (Fig. 1a).

## Discussion

To investigate the timescale and epidemic history of subtype C in Ethiopia we utilized established evolutionary techniques that significantly improved the dating approach previously used by Abebe *et al.* [7,8]. Our estimated date for the tMRCA of 1965 (1959–1973) is based on the assumption that this date corresponds to the introduction of HIV into Ethiopia. It should be noted that this date also corresponds to the time of the ancestry for regionally circulating subtype C viruses including those that gave rise to the Ethiopian epidemic. Two different scenarios exist for the introduction of the virus into Ethiopia. Either a single lineage was introduced and then subsequently split within the country or the tMRCA was actually still in another country and several

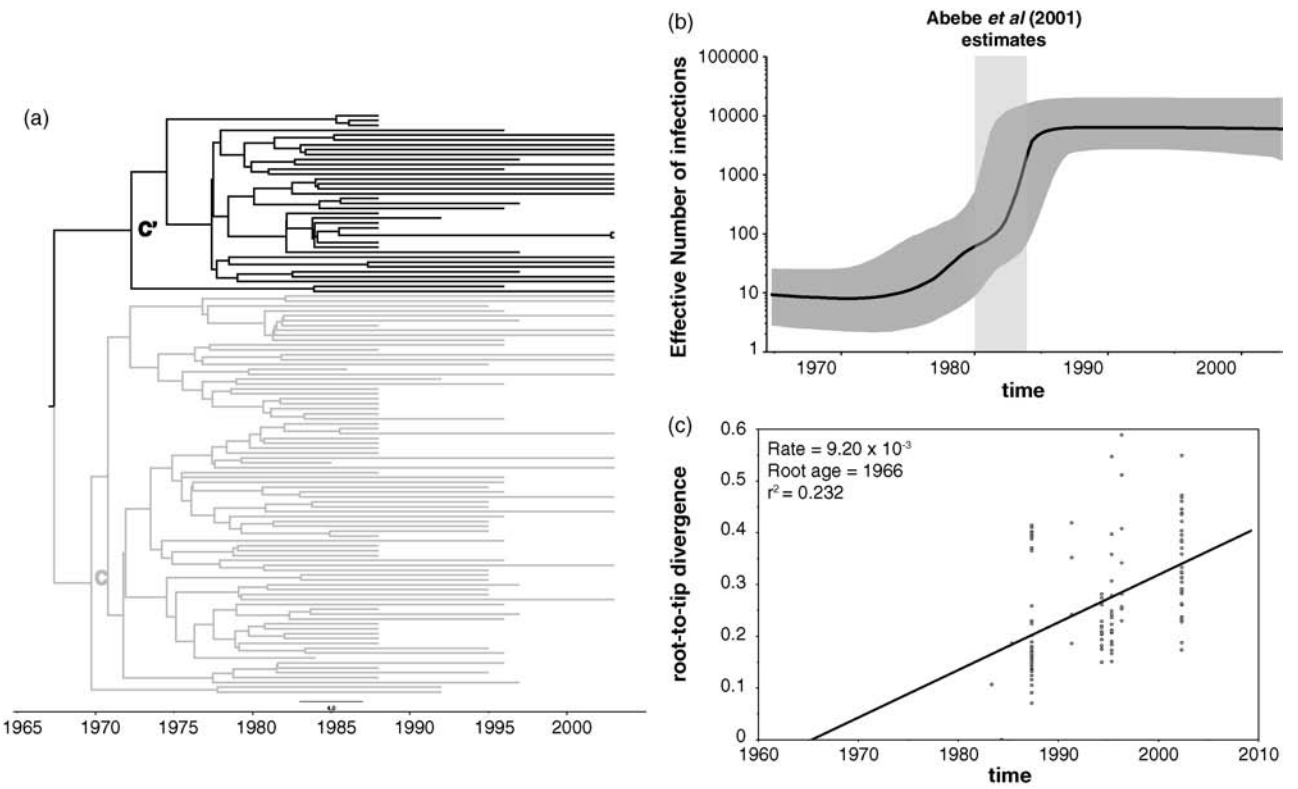


**Fig. 1. (a) Bayesian maximum clade-credibility tree for HIV-1 subtype C *env* in Ethiopia. Trees were inferred using BEAST v1.4.8 from a posterior distribution of 10 000 trees employing constant population size,exponential and logistic growth and the Bayesian skyline plot coalescent tree prior. Branch lengths are depicted in units of time (years). Two subclusters are designated as the C or 'main' group (branches colored in grey) and the C' cluster (branches colored in black). (b) Bayesian skyline plot estimated under a relaxed clock model for the Ethiopian HIV-1C epidemic. Bayesian skyline plot representing estimates of effective number of infections through time for the HIV-1 subtype C *env* variants. The plot begins at the mean posterior tMRCA. The solid bold line represents the inferred median effective population size over time with the 95% upper and lower HPD estimates shaded in blue. (c) Regression of root-to-tip distances versus sampling date for *env*. The inferred rate of nucleotide substitution is given by the slope with time of the most recent common ancestor designated by the X intercept. The correlation coefficient is also shown.**

of its descendents entered Ethiopia independently at approximately the same time (Fig. 1a). The analysis presented here provides no evidence in favor of either of these alternatives, thus both are equally possible. Despite this uncertainty, the results still indicate a significant lag time between the estimated appearance of the MRCA and the first recognition of symptomatic patients in Ethiopia in the 1980s. Such an observation is similar to the conclusions of Gilbert *et al.* [10] where subtype B was circulating cryptically in the United States a decade before AIDS was recognized. Interestingly our tMRCA date coincides with another HIV epidemic in central east Africa where subtype C predominates [29]. Similar to Ethiopia, the first evidence of HIV-1 in Malawi dates back to 1982 and yet the timing of this epidemic dates as far back as the mid-1960s [11]. Our estimates also coincide with the introduction of subtypes A and D into east Africa [30].

Another piece of evidence that is difficult to reconcile with the original dates proposed by Abebe *et al.* [7,8] is the time discrepancy between the initial origin and first AIDS cases. Their initial estimates for the origin of subtype C in Ethiopia are 1982/1983 respectively for each cluster. In stark contrast the first AIDS cases in Ethiopia were recorded in 1986. Therefore, this would represent a very short time span for typical progression to AIDS given that the mean time from infection with HIV to the development of AIDS-related symptoms is approximately 10–12 years [31]. Even if those first cases were indeed rapid progressors they would constitute only a minority of several hundred people infected, although it has been suggested from a few studies done in the African setting that the rate of disease progression is faster among resource-poor patient populations [32–35]. However, due to the small numbers associated with these studies our knowledge about the pattern of disease progression in HIV infection in developing countries is limited [36]. Whether the same trend of rapid disease progression may hold true for Ethiopian patients needs to be further determined [37].

The notable disparity between our estimates and the earlier estimates probably arise from a number of weak

assumptions and inferences associated with the linear regression method. Such limitations include the inability to deal with certain types of evolutionary idiosyncrasies such as different evolutionary rates among lineages and the nonindependence of sampled sequences. Furthermore, the criteria for linear regression may not have been fulfilled in the analysis of synonymous distance versus time in the earlier analysis.

A final striking point of support is that the origin of this epidemic can also be traced back to the first wave of urbanization when Addis Ababa began its rise to a major city between 1967–1975 when rural to urban migration was at its peak [38]. This was followed by a second growth wave between 1975 and 1987 where the population of Addis Ababa skyrocketed, which is in good agreement with the exponential growth phase from our BSP.

In summary, our demographic and evolutionary reconstruction of the Ethiopian epidemic suggests that either a single lineage entered the country and then split within the country to form two distinct clades or several descendents of closely related subtype C viruses with a common ancestor originating in the mid-1960s entered the country. In any case, this was followed by explosive growth in the late 1970s to early 1980s. This is also consistent with an increase of urban agglomerations, which may have facilitated the initial establishment and diffusion of nascent HIV-1 lineages – a trait previously implicated in the birth of the HIV-1/AIDS epidemic [9]. Nevertheless our perspective on the evolutionary history of HIV-1 is hampered by the lack of sequences recovered from the 1960s to 1980s, which could yield additional insights into the dynamics of how and where these epidemics emerged.

## Acknowledgement

We thank the Bioinformatics Core Research Facility at UNL and the Holland Computing Center for computational support. This study was supported in part by PHS grant CA75903, P01AI48240 and NCRR COBRE grant RR15635 to C.W.

D.C.T. designed the study, performed experiments, analyzed data and wrote the manuscript. C.W. contributed to discussions on study design, interpretation and provided critical input on the manuscript.

## References

1. Tsega E, Mengesha B, Nordenfelt E, Hansson BG, Lindberg J. **Serological survey of human immunodeficiency virus infection in Ethiopia.** *Ethiop Med J* 1988; **26**:179–184.
2. Lester FT, Ayehunie S, Zewdie D. **Acquired immunodeficiency syndrome: seven cases in an Addis Ababa hospital.** *Ethiop Med J* 1988; **26**:139–145.
3. Mehret M, Khodakevich L, Zewdie D, Gizaw G, Seyoum A, Shanko B, *et al*. **HIV-1 infection and related risk factors among female sex workers in urban areas of Ethiopia**. *Ethiop J Health Dev* 1990; 4: 163–170.
4. Mehret M, Khodakevich L, Zewdie D, Gizaw G, Seyoum A, Shanko B, *et al*. **HIV-1 infection among employees of the Ethiopian Freight Transport Corporation.** *Ethiop J Health Dev* 1990; **4**:177–182.
5. Abebe A, Pollakis G, Fontanet AL, Fisseha B, Tegbaru B, Kliphuis A, *et al*. **Identification of a genetic subcluster of HIV type 1 subtype C (C') widespread in Ethiopia.** *AIDS Res Hum Retroviruses* 2000; **16**:1909–1914.
6. Abebe A, Kuiken CL, Goudsmit J, Valk M, Messele T, Sahlu T, *et al*. **HIV type 1 subtype C in Addis Ababa, Ethiopia.** *AIDS Res Hum Retroviruses* 1997; **13**:1071–1075.
7. Abebe A, Lukashov VV, Pollakis G, Kliphuis A, Fontanet AL, Goudsmit J, de Wit TF. **Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification.** *AIDS* 2001; **15**:1555–1561.
8. Abebe A, Lukashov VV, Rinke De Wit TF, Fisseha B, Tegbaru B, Kliphuis A, *et al*. **Timing of the introduction into Ethiopia of subcluster C' of HIV type 1 subtype C.** *AIDS Res Hum Retroviruses* 2001; **17**:657–661.
9. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, *et al*. **Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960.** *Nature* 2008; **455**:661–664.
10. Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. **The emergence of HIV/AIDS in the Americas and beyond.** *Proc Natl Acad Sci U S A* 2007; **104**:18566–18570.
11. Travers SA, Clewley JP, Glynn JR, Fine PE, Crampin AC, Sibande F, *et al*. **Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1.** *J Virol* 2004; **78**:10501–10506.
12. Kuiken C, Korber B, Shafer RW. **HIV sequence databases.** *AIDS Rev* 2003; **5**:52–61.
13. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997; **25**:4876–4882.
14. Abecasis AB, Wang Y, Libin P, Imbrechts S, de Oliveira T, Camacho RJ, Vandamme AM. **Comparative performance of the REGA subtyping tool version 2 versus version 1.** *Infect Genet Evol* 2010; 10: 380–385.
15. Bruen TC, Philippe H, Bryant D. **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006; **172**:2665–2681.
16. Huson DH, Bryant D. **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006; **23**:254–267.
17. Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, *et al*. **The molecular population genetics of HIV-1 group O.** *Genetics* 2004; **167**:1059–1068.
18. Drummond AJ, Rambaut A. **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007; **7**:214.
19. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol* 2005; **22**:1185–1192.
20. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. **Relaxed phylogenetics and dating with confidence.** *PLoS Biol* 2006; **4**:e88.
21. Shapiro B, Rambaut A, Drummond AJ. **Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences.** *Mol Biol Evol* 2006; **23**:7–9.
22. Suchard MA, Weiss RE, Sinsheimer JS. **Bayesian selection of continuous-time Markov chain evolutionary models.** *Mol Biol Evol* 2001; **18**:1001–1013.
23. Kass RaR AE. **Bayes Factor.** *J Am Stat Assoc* 1995; **90**:773–795.
24. Rambaut A. **Estimating the rate of molecular evolution: incorporating noncontemporaneous sequences into maximum likelihood phylogenies.** *Bioinformatics* 2000; **16**:395–399.
25. Swofford D. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). 4 edition: Sinauer Associates, Sunderland, Massachusetts; 2003.
26. Salemi M, de Oliveira T, Soares MA, Pybus O, Dumans AT, Vandamme AM, *et al*. **Different epidemic potentials of the HIV-1B and C subtypes.** *J Mol Evol* 2005; **60**:598–605.

27. Bello G, Guimaraes ML, Passaes CP, Almeida SE, Veloso VG, Morgado MG. **Short Communication: evidences of recent decline in the expansion rate of the HIV type 1 subtype C and CRF31_BC epidemics in Southern Brazil**. *AIDS Res Hum Retroviruses* 2009; 25: 1065–1069.
28. Walker PR, Pybus OG, Rambaut A, Holmes EC. **Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth.** *Infect Genet Evol* 2005; **5**:199–208.
29. McCormack GP, Glynn JR, Crampin AC, Sibande F, Mulawa D, Bliss L, *et al*. **Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi.** *J Virol* 2002; **76**:12890–12899.
30. Gray RR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, Serwadda D, *et al*. **Spatial phylodynamics of HIV-1 epidemic emergence in east Africa.** *AIDS* 2009; **23**:F9–F17.
31. NIAID: how HIV causes AIDS. National Institutes of Health, USA; 2004.
32. Morgan D, Malamba SS, Orem J, Mayanja B, Okongo M, Whitworth JA. **Survival by AIDS defining condition in rural Uganda.** *Sex Transm Infect* 2000; **76**:193–197.
33. Morgan D, Mahe C, Mayanja B, Whitworth JA. **Progression to symptomatic disease in people infected with HIV-1 in rural Uganda: prospective cohort study.** *BMJ* 2002; **324**:193–196.
34. Anzala OA, Nagelkerke NJ, Bwayo JJ, Holton D, Moses S, Ngugi EN, *et al*. **Rapid progression to disease in African sex workers with human immunodeficiency virus type 1 infection.** *J Infect Dis* 1995; **171**:686–689.
35. Vasan A, Renjifo B, Hertzmark E, Chaplin B, Msamanga G, Essex M, *et al*. **Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype.** *Clin Infect Dis* 2006; **42**:843–852.
36. Jaffar S, Grant AD, Whitworth J, Smith PG, Whittle H. **The natural history of HIV-1 and HIV-2 infections in adults in Africa: a literature review.** *Bull World Health Organ* 2004; **82**:462–469.
37. Rinke de Wit TF, Tsegaye A, Wolday D, Hailu B, Aklilu M, Sanders E, *et al*. **Primary HIV-1 subtype C infection in Ethiopia.** *J Acquir Immune Defic Syndr* 2002; **30**:463–470.
38. Tesfaghiorghis H. **The growth of urbanization in Ethiopia, 1966–1984.** *East Afr Econ Rev* 1986; **2**:157–167.