HIV-1 subtype C in Ethiopia: genotypic and phenotypic variation
Abebe, A.

Citation for published version (APA):
CHAPTER V

Year of HIV-1 subtype C introduction into Ethiopia: estimated on the basis of genetic analysis of early samples and subsequent virus diversification

Almaz Abebe, Vladimir V. Lukashov, Georgios Pollakis, Aletta Kliphuis, Arnaud L. Fontanet, Jaap Goudsmit, and Tobias F. Rinke de Wit

Submitted for publication
Objective: To derive the year of introduction of HIV-1 subtype C into Ethiopia from sequence information on early samples and subsequent virus diversification in the course of the epidemic (1984-1997).

Design: A set of 474 serum samples obtained in Ethiopia in 1982-1985 was tested for HIV-1 antibodies. From positive samples, HIV-1env gp120 V3 and gag or pol regions were sequenced and analyzed together with sequences obtained in Ethiopia during later stages of the epidemic.

Results: None of 98 serum samples from 1982-1983 contained HIV-1 antibodies. One of 193 (0.5%) samples from 1984 as well as one of 183 (0.5%) samples from 1985 were HIV-1 positive. Phylogenetic analysis of virus sequences obtained from both samples revealed that they belonged to the Ethiopian C, - and not C', - cluster in all genomic regions analyzed. Analysis of 81 Ethiopian C V3 sequences from 1984-1997 revealed that the consensus sequence of the Ethiopian virus population is stable in the course of the epidemic. Both the 1984 and 1985 V3 sequences had no synonymous substitutions compared to the reconstructed common ancestor of the Ethiopian C virus population, in contrast to only 3 of 27 (11%) of the 1988 sequences and none of 51 (0%) of the 1992-1997 sequences. A highly significant correlation between the synonymous distances of the V3 sequences and their sampling years was demonstrated.

Conclusions: We confirmed a low prevalence of HIV-1 infection in Ethiopia in the early 1980's and the close distance of the 1984 and 1985 sequences to the reconstructed common ancestor of the Ethiopian C virus population. The increasing genetic heterogeneity together with stable consensus sequence of the Ethiopian HIV-1 C population demonstrate that evolution of virus population is characterized by an unbiased expansion around a stationary consensus sequence. Based on the rate of virus diversification within the Ethiopian population, we were able to estimate 1983 (95% CI: 1980-1984) as the year of HIV-1 C introduction into Ethiopia, which is in agreement with the seroepidemiological data.

KEYWORDS: HIV-1 epidemic, Ethiopia, HIV-1 genetic characterization, HIV-1 subtype C, molecular clock.

INTRODUCTION
As most countries of Sub-Saharan Africa, Ethiopia has been experiencing a severe HIV-1/AIDS epidemic during the last 15 years. Intensive retrospective seroepidemiological
studies, the results of which are summarized in Table I, have revealed the absence of HIV-1 in rural or urban Ethiopian populations prior to 1984. The first case of HIV-1 infection in Ethiopia has been registered in 1984 in the capital city Addis Ababa, when serum samples of 167 hospitalized patients have been tested for anti-HIV-1 antibodies. The first AIDS cases in Ethiopia have been registered in 1986 in hospitals in Addis Ababa. In 1986-1988, a high prevalence of HIV-1 has been demonstrated along the main trading roads of the country, varying from 13% among truck drivers to an average of 17% among commercial sex workers (CSW) in 1988. In 1994, data from eleven urban blood banks showed that the prevalence of HIV varied from 5% to 20%, being 6.6% in Addis Ababa (National Blood Transfusion Service, Ethiopian Red Cross Society, 1994, unpublished). Among antenatal care attendants in Addis Ababa, 11-13% of individuals were found to be HIV-1 seropositive in 1991, compared to 18% in 1996. A recent study among CSW of Addis Ababa demonstrated an HIV prevalence of 74% (Aklilu M. et al., submitted).

Earlier, we and others have reported that the HIV-1 epidemic in Ethiopia is caused by subtype C viruses. Among the Ethiopian env gp120 V3 sequences, two phylogenetic clusters can be distinguished and are designated C and C'. It has been shown, that both C and C' viruses have been circulating over the last decade among the same risk groups and geographical areas.

In this study, we report the identification of two HIV-1 positive serum samples from Addis Ababa, which have been obtained from HIV-1 seropositive individuals sampled in 1984 and 1985. To study whether these two viruses are genetically close to the founder virus which has been originally introduced into Ethiopia, we obtained and analyzed genetic information for these two earliest Ethiopian viruses in conjunction with sequence data from later years, obtained in the same geographic area.

MATERIALS AND METHODS

Clinical samples

Four hundred and seventy four serum samples, obtained in the period of 1982-1985 from hepatitis B virus (HBV) infected patients who visited the former National Research Institute of Health (NRIH) Virology laboratory for routine diagnosis of HBV infection, were used for our study. These 474 samples were tested for the presence of HIV-1 antibodies by ELISA (Vironostika Uniform II, Organon Teknika, The Netherlands) and Western Blot (HIV Blot version 2.2, Genelabs Diagnostic Biotechnology, USA) according to the
The samples selected for the linear regression analysis \((n=81)\) belong to the main C group [Abebe A. et al in press] and originated from the period of 1984-1997. Sera were collected from: HBV and HIV-1 positive samples \((n=2)\), Addis Ababa, in 1984 and 1985, an AIDS patient, \((n=1)\), Addis Ababa 1986; commercial sex workers, pregnant women, and blood donors \((n=26)\), Addis Ababa, over 1992-1995; and commercial sex workers and blood donors of seven different Ethiopian towns including Addis Ababa \((n=52)\), in 1988 and 1996/7.

### Sequencing

For HIV-1 positive samples, RT-PCR of three genomic regions \((env, gp120 V3, gag p17/p24, and pol protease/partial RT)\) was performed. Briefly, RNA was extracted from 100\(\mu\)l of serum by the silica-based method \(^{20}\). After washing and elution from the silica with 100\(\mu\)l of sterile water, 10\(\mu\)l of the eluate was used in a reverse transcription reaction with avian myeloblastosis virus reverse transcriptase. A region of 287 bp of the HIV-1 envelope V3 region was amplified with C2-V3 primers and further amplified with a nested PCR. The outer primers were 3'-V3-NOT (5'-GCAGCCGCCCCTCCTCTACAATTTAAACTGTG-3') and 5'-V3-NOT (5'-GCAGCCGCCCCTCCTCTACAATTTAAACTGTG-3'); the inner primers were 5'-V3-SP6 (5'-GATTTAGGTGACATATAG-3') and 3'-V3-T7 (5'-TAATACGACTCAGATATACCC-3'), as described earlier \(^{21-23}\). For amplification of a 743 bp fragment of the gag gene (complete \(p17\) and partial \(p24\)), outer primers 3'-SK39 (5'-GCATTCTGGACATAAGAACAGACCACAA-3') and 5'-GAG-1 (5'-GCAGAGGCCGTCAC-3') and inner primers 5'-GAG-2-SP6 (5'-GGGAAAATTCGGTAAAGGG-3') and 3'-GAG-AE3-T7 (5'-TAGGACCCTTAATTTATTTATACCA-3'), were used as described earlier \(^{24,25}\). For the amplification of the pol region (1178 bp including the protease and 297 amino acids of the RT gene), the first PCR reaction was performed with the outer primers 5'-protFM (5'-CAAGGGAGCGCCAGGGAAATTT-3', HxB2 positions 2111-2130) and 3'-halfRT (5'-TGACCAGATACCTTAATTACGAGAAATA, HxB2 positions 3505-3535). Subsequently, DNA was amplified in three nested PCR reactions (fragments A, B and C), covering the region. The inner primers used were: for fragment A, 5'-SP6-prot/RT (5'-CTTTAAGCTTCCTCCTGATGACCACT-3', HxB2 positions 2242-2263) and 3'-T7prot (5'-CCTATTGAAACTGTACCAGTA-3', HxB2 positions 2558-2578), for fragment B, 5'-SP6p66/out (5'-GACCTACCTGCTGCAACATA-3', HxB2 positions 2484-2505) and 3'-endprotT7 (5'-TGGAAAGGATCACCAGCAATATT-3', HxB2
positions 3005-3027); for fragment C, 5'-SP6p66 (5'-AGATATCAGTACAATGTGTT-3', HxB2 positions 2975-2994) and 3'-halfpol (5'-AAGCAGAGCTAGAACTGGCAGA-3', HxB2 positions 3441-3462) using the conditions described earlier.

Both strands of the nested-PCR fragments were directly sequenced by using the SP6 and T7 primers. Sequencing was performed with Taq dye primers (Applied Biosystems, Foster City, Calif.) and the ThermoSequenase fluorescence-labeled primer cycle-sequencing kit (Amersham International, Little Chalfont, England). The sequence products were analyzed on an automatic DNA sequencer (model 373A Applied Biosystems, Foster City, CA).

Sequence analysis

Nucleotide sequences obtained in this study were aligned manually together with 79 sequences obtained from Ethiopia in earlier studies. Positions containing an alignment gap were excluded from pairwise sequence comparisons. The most recent common ancestor for the Ethiopian subtype C epidemic was reconstructed as the common node of 81 Ethiopian C sequences using several phylogenetic methods. Phylogenetic trees were constructed by using the neighbor-joining (NJ) and maximum-likelihood algorithms (ML) as implemented in PHYLIP package (NEIGHBOR and DNAML, respectively) (http://evolution.genetics.washington.edu/phylip.html). DNAML method was based on empirically found base frequencies and transition/transversion ratios, considering different rates of evolution at different positions. The neighbor-joining method was based on gamma distances for the Jukes-Cantor method. For both methods, reference sequences of HIV-1 subtypes other than C, provided by the Los Alamos database (http://hiv-web.lanl.gov), were used to root the trees. Subtype C sequences from other countries were also included.

Subsequently, neighbor-joining phylogenetic trees based on synonymous (D$_s$, Nei-Gojobori method with Jukes-Cantor correction) evolutionary distances were built by using the MEGA package. For all methods, evolutionary distances of individual sequences to the most recent common ancestor were calculated and analyzed by using several statistical approaches.

Statistical methods

All statistical calculations were done by using the SPSS/PC+ software (version 5.0, SPSS Inc., Chicago, Illinois, USA). The relationship (correlation) between sampling years of individual HIV-1 sequences and their synonymous distances to the most recent common ancestor was examined by using linear regression analysis. Each sequence was considered...
to be statistically independent. The distance data were studied by using linear regression analysis in two ways. In the first method, the distances of all individual sequences to the common node were analyzed in relation to their sampling year (all sequences equally contributed to the analysis). In the second approach, the mean distances calculated per year were used (all years equally contributed to the analysis). For each analysis, 95% confidence intervals for the time of virus introduction into Ethiopia, has been calculated (CI for X-intercept).

RESULTS

A total of 474 serum samples, collected from 1982 to 1985 from Ethiopian individuals, were tested by ELISA and Western Blot for HIV-1 antibodies. Among samples obtained in 1982 (n=44) and 1983 (n=54), none were found to be HIV-1 positive. One of 193 (0.5%) samples from 1984 as well as 1 of 183 (0.5%) samples from 1985 were found to contain anti-HIV-1 antibodies (Table 1). These samples were obtained from a 26- and a 41-years old male, respectively.

<table>
<thead>
<tr>
<th>year</th>
<th>risk group</th>
<th>location</th>
<th>n</th>
<th>HIV+</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>hepatitis B</td>
<td>AA</td>
<td>44</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>1982-83 (ref.2)</td>
<td>hospital patients</td>
<td>AA, AS, HA, YI, NE</td>
<td>500</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>1983</td>
<td>hepatitis B</td>
<td>AA</td>
<td>54</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>1983 (ref.2)</td>
<td>blood donors</td>
<td>AA</td>
<td>459</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>1984</td>
<td>hepatitis B</td>
<td>AA</td>
<td>193</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>1984 (ref.2)</td>
<td>Bell’s palsy + controls</td>
<td>Ethiopia</td>
<td>267</td>
<td>2 (0.7%)</td>
</tr>
<tr>
<td>1985</td>
<td>hepatitis B</td>
<td>AA</td>
<td>183</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>1985-87 (ref.2)</td>
<td>liver diseases</td>
<td>AA</td>
<td>528</td>
<td>13 (2.5%)</td>
</tr>
</tbody>
</table>

Table 1. Early HIV-1 seroprevalence in Ethiopia (1982-1987). The table summarizes earlier published data (2) and those from the present study. AA – Addis Ababa, AS – Asmara, HA – Harar, YI – Yirgalem, NE – Nekemte, n – number of serum samples tested.

From these two samples, virus-specific RNA was isolated, and RT-PCR was performed for the HIV-1 env, pol, and gag genetic regions. While amplification of the gp120 V3 region was possible for both samples, the gag region was successfully amplified for the 1984 sample, and the pol region was amplified for the 1985 sample only. Since both gag and pol regions are generally more conserved compared to the env V3 region, it is likely that negative results of amplifications are related to longer size of these fragments. The predicted amino acid sequences of the V3 regions of the 1984 and 1985 samples are shown in Figure 1.
Figure 1: Predicted amino acid sequences of the env gp120 V3 region of the Ethiopian 1984 and 1985 samples. The sequences are shown in comparison with the Ethiopian consensus sequences. Dots indicate amino acid identities, dashes indicate deletions. Question marks represent cases when two or more amino acids were the most often ones at that positions, should be included. The consensus sequences for several periods of the Ethiopian epidemic are also shown.

Phylogenetic analysis revealed that both the 1984 and 1985 viruses belong to the C cluster of Ethiopian subtype C viruses in the gp120 V3 region (Figure 2). The gag region of the 1984 and the pol region of the 1985 samples also clustered with the C sequences, and not with the C' sequences (data not shown). To analyze whether the consensus sequence of the V3 region of the subtype C epidemic in Ethiopia is changing in the course of the epidemic, we calculated consensus sequences for six periods of the epidemic based on 81 sequences obtained in this and earlier studies: consensus 1984-1986 (based on 3 sequences sampled between 1984 and 1986), 1988 (27 sequences sampled in 1988), 1992 (6 sequences), 1995 (20 sequences), 1996 (19 sequences), and 1997 (6 sequences) (Figure 1).

For each sequence position, the most often amino acid was included in the consensus, even if it was not present in the majority of sequences (question marks represent cases when two or more amino acids were the most often ones at this position). Our analysis revealed that, during the HIV-1 epidemic in Ethiopia, its consensus sequence has remained stationary. We observed that, while there were differences among consensus sequences of different periods of the epidemic, at neither position an amino acid that was absent or in minority among early samples, became consistently dominant among late samples. Differences between the consensus sequences of the six periods of the epidemic were observed at fourteen positions, and all but one different amino acids were seen in one consensus sequence only. At the only sequence position (position 36), an amino acid difference was seen in the consensus sequences of two periods of the epidemic. At this position, the early consensus sequences (1984-1986 and 1988) had an E (glutamic acid), while in the consensus sequences of 1992 and 1995 a K (lysine) was the most prevalent amino acid. Yet, the latest consensus sequences (1996 and 1997) again contained an E at this position.
Earlier obtained Ethiopian C and C' sequences, as well as sequences of HIV-1 subtypes other than C and subtype C sequences from India and Brazil, are also included. All reference sequences are labeled as in the original studies. The 1984 and 1985 sequences are boxed. The reconstructed most recent common ancestor of the Ethiopian C epidemic is marked by a closed circle. The C and C' clusters are indicated. Bootstrap values are shown (% of 100 replications). In the gag and pol regions, the sequences of the 1984 and 1985 samples also belonged to the C cluster (not shown).

Our subsequent analysis of the amino acid distribution at this position among individual sequences revealed that both early and late sequences had various amino acids at this position (1988 – 25 E, 2 K; 1992 – 1 E, 4 K, 1 R; 1995 – 7 E, 10 K, 3 R; 1996 – 12 E, 5 K, 1 R, 1 N, 1997 – 3 E, 1 K, 2 R). To test whether the 1984 and 1985 V3 sequences are close
to the founder C virus which had originally been introduced into Ethiopia, we reconstructed
the most recent common ancestor of the Ethiopian C epidemic as the common node (Figure
2) of 81 Ethiopian V3 sequences with known sampling years, which were obtained over a
period of 1984-1997 and belonged to the C cluster. This genomic region was used because
intensive longitudinal sequence data for the V3 region are available from Ethiopia.
Subsequent analysis of the evolutionary distances of the 1984 and 1985 sequences to the
reconstructed common node of the Ethiopian C epidemic revealed that these sequences are
closer to the founder virus, compared to later sequences. The difference in distances of the
1984-1985 and later samples to the reconstructed founder virus was especially manifest
when only synonymous substitutions were analyzed, with both the 1984 and 1985 samples
(100%) having a zero synonymous distance to the node. Among later samples (1986-1997),
only 3 of 79 sequences (4%) had zero synonymous distances to the node. All three later
samples with zero distances to the node were obtained in 1988 (3 of 27 sequences, 11%),
while none of 51 sequences (0%) from 1992-1997 had a zero distance to the reconstructed
founder virus. The mean synonymous distances (±SE) of later samples to the node were:
1988 – 0.040±0.007, 1992 – 0.063±0.013, 1995 – 0.094±0.012, 1996 – 0.094±0.013, and
1997 – 0.098±0.030. The mean nucleotide distance of the 1984-1985 sequences to the node
calculated by using the NJ method was also lower compared to those of samples taken
during any of later periods: 1984-1985 – 0.040±0.024, 1986 – 0.049 (one sample, SE cannot
be computed), 1988 – 0.063±0.005, 1992 – 0.088±0.018, 1995 – 0.112±0.009, 1996 –
0.093±0.007, 1997 – 0.115±0.011 (1986-1997 – 0.090±0.043). Similar results were also
obtained by using the ML method (data not shown).
Regression analysis of evolutionary distances of 81 Ethiopian V3 sequences revealed that,
there is a highly significant (p<0.0000001) positive correlation between sampling years of
individual sequences and their synonymous distances to the reconstructed common ancestor
of the Ethiopian C epidemic (Figure 3). The extrapolation of the regression line of
synonymous distances back to the date when no synonymous heterogeneity was present in
the Ethiopian HIV-1 C population allowed to estimate 1983 (95% CI: 1980-1984) as the
year of HIV-1 C introduction into Ethiopia. Similar results were obtained when regression
analysis was based on the mean synonymous distances of all sequences sampled in each year
(Figure 3). This approach resulted in a marked increase of correlation (r=0.99, r²=0.97)
between the mean synonymous distances of sequences and their sampling years.
DISCUSSION

In the present study, we were able to demonstrate and further confirm a very low prevalence of HIV-1 in Addis Ababa, Ethiopia, in the early 1980's. Among the 474 samples from 1982-1985 tested in our study, only two were found to be positive for anti-HIV-1 antibodies. None of the 98 samples from 1982 and 1983, that were obtained from hepatitis B patients in Addis Ababa, were HIV-1 positive. Only in 1984, the first HIV-1 positive samples were identified, resulting in a low prevalence of 0.5% (1 of 193 samples). This prevalence has remained unchanged in 1985, when one of 183 samples (0.5%) were found to be HIV-1 positive. Our data are in agreement with previously published data on the prevalence of HIV-1 in Ethiopia in the early 1980’s.\(^1,2\) When taken together with HIV-1 prevalence of up to 38% among CSW in 23 cities in Ethiopia in 1988,\(^5\) the above data indicate an extremely rapid spread of HIV-1 infection through the country between 1984 and 1988. From both the 1984 and 1985 HIV-1-positive samples, we were able to obtain genetic information for the env gp120 V3 and pol or gag regions of the HIV-1 genome. To the best of our knowledge, these two samples represent the earliest sequences available for
the Ethiopian epidemic. Phylogenetic analysis revealed that both samples belonged to the Ethiopian C, and not to C', cluster.

To study virus diversification within the Ethiopian C epidemic, we analyzed a set of 81 V3 sequences obtained in the period of 1984-1997 for this and earlier studies. We were able to demonstrate that the amino acid consensus sequence of the Ethiopian HIV-1 C population has remained stationary in the course of the epidemic. While we did observe amino acid variations between consensus sequences of different periods of the epidemic, none of these differences was consistent (Figure 1). In other words, none of the amino acids, that were absent or in minority among early Ethiopian viruses, has been subsequently selected for during the course of the epidemic. This observation is similar to earlier findings for the subtype B epidemic [25-29]. Together with increasing genetic heterogeneity among individual sequences over time, which we showed in this study, this observation demonstrates that evolution of HIV-1 C virus population in Ethiopia is characterized by an unbiased expansion around a stationary consensus sequence.

To test whether the 1984 and 1985 V3 sequences are close to the virus which had originally been introduced into Ethiopia, we used several phylogenetic methods to reconstructed the most recent common ancestor of the Ethiopian C epidemic. Our comparison of evolutionary distances of the 1984 and 1985 sequences with those of sequences obtained later in the epidemic revealed that the 1984-1985 sequences are closer to the ancestor. While being evident for nucleotide distances, this difference was especially pronounced for synonymous distances, with both the 1984 and 1985 sequences, compared to 11% of the 1988 and 0% of the 1992-1997 sequences, having zero synonymous distances to the ancestor. This observation agrees with previous data for other epidemics which indicate that (synonymous) genetic heterogeneity of virus populations is increasing in the course of the epidemic [28-32].

By analyzing the synonymous distances of the Ethiopian sequences to the ancestor in the course of the epidemic (1984-1997), like it has been done earlier for the US and Dutch HIV-1 epidemics (Lukashov, V.V and Goudsmit, J., in preparation), we were able to demonstrate their highly significant correlation with the sequence sampling years (Figure 3). The extrapolation of the regression line back to the date when no synonymous heterogeneity was present in the Ethiopian HIV-1 C population allowed us to estimate 1983 (95% CI: 1980-1984) as a year of HIV-1 C introduction into Ethiopia (Figure 3), which is in agreement with seroepidemiological data (Table 1).
REFERENCES


27. Kumar, S., Tamura, K., and Nei, M. Molecular evolutionary genetics analysis (MEGA). Version 1.01. Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania, USA. 1993.


ACKNOWLEDGEMENTS

This study is part of the Ethio-Netherlands AIDS Research Programme (ENARP), a collaborative effort of the Ethiopian Health and Nutrition Research Institute (EHNRI), the Amsterdam Municipal Health Service (GG&GD), the Central Laboratory of the Netherlands Red Cross Blood Transfusion Service (CLB), and the Academic Medical Center of the University of Amsterdam (AMC). ENARP is financially supported by the Netherlands Ministry of Foreign Affairs and the Ethiopian Ministry of Health (MOH) as a bilateral project.