Timing and Reconstruction of the Most Recent Common Ancestor of the Subtype C Clade of Human Immunodeficiency Virus Type 1

Simon A. A. Travers,1 Jonathan P. Clewley,2 Judith R. Glynn,3 Paul E. M. Fine,3 Amelia C. Crampin,3,4 Felix Sibande,4 Dominic Mulawa,4 James O. McInerney,1 and Grace P. McCormack1*

Biology Department, National University of Ireland, Maynooth, County Kildare, Ireland; Sexually Transmitted and Blood Borne Virus Laboratory, Health Protection Agency,2 and Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine,3 London, United Kingdom; and Karonga Prevention Study, Chilumba, Malawi4

Received 9 February 2004/Accepted 5 May 2004

Human immunodeficiency virus type 1 (HIV-1) subtype C is responsible for more than 55% of HIV-1 infections worldwide. When this subtype first emerged is unknown. We have analyzed all available gag (p17 and p24) and env (C2-V3) subtype C sequences with known sampling dates, which ranged from 1983 to 2000. The majority of these sequences come from the Karonga District in Malawi and include some of the earliest known subtype C sequences. Linear regression analyses of sequence divergence estimates (with four different approaches) were plotted against sample year to estimate the year in which there was zero divergence from the reconstructed ancestral sequence. Here we suggest that the most recent common ancestor of subtype C appeared in the mid- to late 1960s. Sensitivity analyses, by which possible biases due to oversampling from one district were explored, gave very similar estimates.

The diversity of human immunodeficiency virus type 1 (HIV-1) is expanding globally. Currently, there are nine recognized subtypes of HIV-1, 14 circulating recombinant forms, and many divergent strains (28, 29). While subtype B predominates in North America and Europe (23), it is estimated that more than 55% of worldwide HIV-1 infections are caused by HIV-1 subtype C (6). This high prevalence is due to the predominance of subtype C in southern and eastern Africa (1, 4, 21, 27, 38) and India (32) and its increasing prevalence in Brazil (33) and China (30).

It is not clear whether the predominance of subtype C is a reflection of a founder effect or whether it reflects a relatively high fitness of this virus subtype (22). Subtype C viruses display certain unique genetic characteristics that, by altering biological activity, may have contributed to the success of this subtype (19, 20). Examples include the presence of three or four NF-κB enhancer copies in the long terminal repeat region (23), Tat and Rev proteins that are prematurely truncated, and a 15-bp insertion of the 5’ end of the vpu reading frame (9). Subtype C also shows a preference for CCR5 co-receptor usage, and the variable V3 loop is relatively conserved in subtype C compared to subtype B (23, 32). Recent evidence suggests the emergence of CXC4-utilizing subtype C viruses in South Africa (39) and a rapidly growing subtype C epidemic in southern Brazil (33). Understanding the origin and subsequent diversification (the extent and rate of change) of this subtype is vitally important in efforts to design prevention and control therapies such as vaccines.

Estimating the date of the most recent common ancestor of viral strains, including dating the origin of HIV-1 and HIV-2, has proven to be possible with molecular data for well-characterized sequences with known sampling dates (12, 13, 31, 37). Thus, the HIV-1 pandemic is thought to have originated in the 1930s from a cross-species transmission of simian immunodeficiency virus (SIV) from chimpanzees (12, 31). Subsequently, this founder strain diversified into the various subtypes that we see worldwide today.

The most recent common ancestor of subtype B is thought to have appeared between 1960 and 1976 (12, 16, 31). Abebe et al. (2, 3) estimated that subtype C was introduced into Ethiopia in the early 1980s (1980 to 1984), and it was first recorded in Karonga District, northern Malawi, in 1983. The earliest subtype C DNA sequences available from other parts of Africa are from the mid- to late 1980s in South Africa (36, 42), Zambia (9, 18, 34), Somalia (9, 14, 15, 34), Tanzania (43), Gabon (5), and Rwanda (11). To date there have been no published attempts to estimate the year in which the most recent common ancestor of this subtype may have appeared. Population-based studies in Karonga District from the early 1980s to the present day (8, 17, 24, 25; Glynn et al., XIth Int. Conf. AIDS STDs Africa, 1999) have yielded HIV-1 subtype C gag and env gene sequence fragments with known sampling dates ranging from 1983 to 2000 from a large number of HIV-1-positive individuals. With these sequences along with all other available subtype C sequences with known sampling dates from the HIV-1 sequence database (http://hiv-web.lanl.gov) and a number of different phylogenetic approaches, we estimated the date of the most recent common ancestor of subtype C and of the epidemic in Karonga District, Malawi.

MATERIALS AND METHODS

Sequence information. Non-intersubtype-recombinant subtype C sequences spanning approximately 620 bp of the gag p17-p24 region and approximately 420 bp of the env C2-V3 of subtype C sequences with known sample dates from the HIV-1 sequence database (http://hiv-web.lanl.gov).
bp of the env C2-V3 region with known sampling dates were available from molecular epidemiological studies of HIV-1 in the Karonga District, Malawi (17). Only one member of each linked spouse pair was included. All available sequences covering the same gene regions with known sampling dates were retrieved from the Los Alamos HIV database (http://hiv-web.lanl.gov), excluding recombinants and sequences from multiple clones. Alignments were produced that contained only subtype C sequences from the Karonga District, Malawi, and that contained all available subtype C sequences for each gene region (Table 1).

Ancestral sequence and phylogeny reconstruction. The most recent common ancestor of the subtype C clade and of the subtype C epidemic in the Karonga District, Malawi, were reconstructed for both gene regions. Reference sequences from each HIV-1 group M subtype were obtained from the Los Alamos HIV database and used to root the subtype C phylogeny in eight separate ancestral sequence reconstructions with MrBayes 2 (10). A consensus sequence of the eight resulting ancestral sequences was taken as the most recent common ancestor and used as the outgroup in subsequent phylogeny reconstructions. Modeltest (26) was used to select the optimal model of substitution for each data set, and model parameters were further optimized in PAUP* (35). Phylogeny reconstruction was carried out with both distance and Bayesian methods. With distance methods, a heuristic search strategy was used, with starting trees obtained by random stepwise addition of taxa (10 replicates) and branch swapping, as implemented in PAUP* (35). Bayesian phylogeny reconstruction was performed with MrBayes (10) for 4 million generations, sampling every 100 generations, and model parameters were updated at each tree being constructed from generations 2.5 to 4 million to ensure settling of the likelihood values. All tree files and alignments are available from the authors on request.

Dating methods. Four independent estimates of the genetic divergence of individual sequences from the subtype C ancestral sequence were calculated from each data set: synonymous distances (dS), calculated with the codeml program from the PAML package (41); genetic distances (with the optimal substitution model) calculated with PAUP* (35); branch lengths from individual sequences to the ancestral node on phylogenies reconstructed from distance data (optimized with maximum likelihood and the appropriate substitution model in PAUP* (35) and extracted from the .tre files with p4 (3); P. Foster, www.nhm.ac.uk/zooloogy/external/p4.htm); and branch lengths from individual sequences to the ancestral node of the Bayesian phylogeny.

Correlation between the above divergence values and the sampling year were examined with linear regression analysis. The year at which the regression line crossed the x axis was taken as an indication of the year of origin, i.e., when sequence divergence from the ancestral sequence was zero. Both individual sequence values and mean values per year were plotted against sampling year for each data set. For each analysis, the significance of the correlation was calculated with one-tailed t tests, and 95% confidence intervals for the year in which the most recent common ancestor occurred were calculated (confidence interval for the x intercept) with the Mintab statistical software.

Method validation. The most recent common ancestor of a large clade within the gag subtype C phylogeny, restricted primarily to sequences originating in Malawi and possibly introduced by a single individual shortly before 1983 (17), was reconstructed, and its date of origin was estimated in one approach to test the ability of our methods to produce a sensible date for an ancestral sequence. To ensure that tree space was well searched, 500 heuristic searches, from a random starting point with total branch reconstruction branch swapping, were carried out on both Karonga env and gag sequence alignments. Branch lengths for each tree were plotted against sampling date in linear regression analyses as described above. Further, 500 bootstrap replicates were performed on the same datasets in PAUP (35), and maximum-likelihood optimized branch lengths of each replicate were similarly plotted.

All sequences available for analysis dating from 1983 and 1984 had been collected in Karonga District, and to investigate their possible effect on the estimated date of the most recent common ancestor, alignments were produced that had all sequences dating from 1983 and 1984 removed and that had every possible combination of these sequences included for both gene regions. For each alignment, maximum-likelihood optimized branch lengths of heuristic trees were plotted against sampling date. Finally, to attempt to account for possible bias due to the disproportionate number of sequences from Karonga District present in the entire subtype C alignments (55.4 and 42.6% of sequences in the entire gag and env data sets, respectively), representative data sets containing a more balanced geographic representation of sequences were assembled and analyzed for each gene region.

## RESULTS

### Origin of the subtype C epidemic

When all available gag p17-p24 subtype C sequences were included for analysis with each sequence contributing equally, linear regression analysis indicated a date between 1966 and 1969 for the timing of the origin of subtype C; the 95% confidence intervals ranged from 1959 to 1973 (entire gag, Table 2). There was a significant correlation between genetic divergence and time ($P < 0.005$), consistent with progressive evolution over time, i.e., clocklike behavior. Linear regression analysis of genetic distances and

<table>
<thead>
<tr>
<th>Data set</th>
<th>Description</th>
<th>Length (bp)</th>
<th>Size (no. of taxa)</th>
<th>Years represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire gag</td>
<td>All subtype C gag sequences with known sampling date and with desired fragment</td>
<td>660</td>
<td>376</td>
<td>1983–1984, 1986–2000</td>
</tr>
<tr>
<td>Representative entire gag</td>
<td>Selected from entire gag to give a balanced representation of countries for each sampling year (as far as possible)</td>
<td>639</td>
<td>59</td>
<td>1983–1984, 1986–2000</td>
</tr>
<tr>
<td>Entire env</td>
<td>All subtype C env sequences with known sampling date and with desired fragment</td>
<td>435</td>
<td>299</td>
<td>1984, 1986–2000</td>
</tr>
<tr>
<td>Karonga env</td>
<td>All Karonga region subtype C env sequences (including only one from each linked spouse pair)</td>
<td>417</td>
<td>125</td>
<td>1984, 1987–1989, 1997–2000</td>
</tr>
<tr>
<td>Representative entire env</td>
<td>Selected from entire env set to give a balanced representation of countries for each sampling year (as far as possible)</td>
<td>435</td>
<td>59</td>
<td>1984, 1986–2000</td>
</tr>
</tbody>
</table>

* gag sequences used are fragments of the p17–p24 region, and env fragment are from the C2–V3 region.
tree branch length data from the env alignments indicated a date between 1962 and 1965 (1955 to 1972) (entire env, Table 2) for the time at zero divergence, while the ds values suggested 1972 (95% confidence interval, 1966 to 1976). All correlations were significant (P < 0.005). Estimates produced by allowing each year to contribute equally to the analysis (plotting through the mean value for each year) were less consistent (data not shown).

**Origin of the subtype C epidemic in the Karonga District, Malawi.** Linear regression analysis of all gag sequences from Karonga indicated a date between 1966 and 1969 for the ancestor of the Karonga subtype C epidemic when genetic distances and branch length data were used (95% confidence interval, 1960 to 1973, Karonga gag; Table 2). However, linear regression analysis of ds values calculated from the Karonga gag sequences indicated 1960 to 1961 as a likely date for the ancestor of the Karonga epidemic (P < 0.005; confidence interval, 1948 to 1966). Analysis of the Karonga env data set indicated that the origin of the epidemic was dated at 1966 to 1973 with genetic distance and branch length data and 1975 for ds data (Karonga env, Table 2).

**Method validation.** The regression analysis carried out on sequences belonging to a Malawi clade evident within the subtype C phylogeny indicated 1979 to 1981 as the date of the most recent common ancestor of this clade. Maximum-likelihood optimized branch lengths of 500 heuristic trees from both Karonga gag and env alignments plotted against the sampling year of each sequence estimated the likely range for the date of the most recent common ancestor of the Karonga epidemic (Fig. 1) as either 1964 to 1972 (gag) or 1956 to 1970 (env), with the 500 bootstrap replicate estimates being 1964 to 1974 and 1954 to 1974 for gag and env, respectively.

The influence of the early subtype C sequences collected in Karonga District in 1983 and 1984 on the estimated dates of the appearance of the most recent common ancestor of the subtype C epidemic itself and of the subtype C epidemic in Karonga is shown in Fig. 1. Regression analysis of branch lengths from heuristic trees drawn from gag sequences suggest a date between 1967 and 1972 for the Karonga epidemic (Fig. 1A) and 1965 to 1970 for the entire subtype C epidemic (Fig. 1B), while similar analysis on the env sequences suggested 1942 to 1970 for the Karonga epidemic (Fig. 1C) and 1961 to 1967 for the entire subtype C epidemic (Fig. 1D). Estimated dates of the subtype C most recent common ancestor calculated from the representative alignments ranged from 1968 to 1974 (95% confidence interval, 1958 to 1978) for gag and 1947 to 1971 (95% confidence interval, 1875 to 1978) for env (see Table 2).

**DISCUSSION**

In one of the most intensive studies of its kind, we have estimated the date of origin of the most prevalent subtype of HIV-1, subtype C. Our work suggests that the most recent common ancestor of subtype C appeared in the late 1960s and that the most recent common ancestor of the subtype C epidemic in Karonga District, Malawi, also dates back to this period. This is consistent with the origin of HIV-1 group M (1930s) (12, 31) and the first evidence of subtype C (1983) in Malawi (17) as well as with estimates of the origins of subtype B (some time between 1960 and 1976) (12, 16, 31). A number of sensitivity analyses were used to account for known biases in the data, and these supported the main result.

We applied our methods to the estimation of the most recent common ancestor of a well-described (17) clade within the subtype C phylogeny with gag sequences. Regression analysis of four divergence estimates versus time suggested that the ancestor of this clade appeared between 1979 and 1981. This date is consistent with our previous conclusion, from molecular and epidemiological evidence, that a single individual may have been responsible for the introduction of this cluster shortly before 1983 (17). The good agreement between the methods and the plausible date produced give us a certain degree of confidence in the ability of these methods to estimate the most recent common ancestor of the larger data sets. For the most part, the correlation between genetic divergence and time was significant, consistent with clocklike behavior. Furthermore, estimates of the date of the most recent common ancestor resulting from plotting 500 heuristic trees were consistent with dates produced with the four different methods of estimating divergence (1964 to 1972 for gag, 1956 to 1970 for

---

**TABLE 2. Estimated date of origin of subtype C and of the subtype C epidemic in the Karonga District, Malawi, from gag and env gene sequences**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Method</th>
<th>Date of origin&lt;sup&gt;a&lt;/sup&gt; (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MrBayes BLens</td>
<td>1965 (1958–1972) 1973 (1968–1977) 1947 (NA) (P &lt; 0.05)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> ds, synonymous distances; ML Opt Blens, maximum-likelihood optimized heuristic distance tree branch lengths; MrBayes BLens, branch lengths obtained from phylogeny inferred with MrBayes. NA: confidence intervals incalculable.

<sup>b</sup> Data significant at P < 0.005 except as noted. Four estimates of genetic divergence of individual sequences to the subtype C ancestral node were used: (i) synonymous distances, ds (codeml program, PAML package); (ii) genetic distances with the optimal substitution model (PAUP*); (iii) branch lengths from individual sequences to the ancestral node on phylogenies reconstructed from distance data (optimized by maximum likelihood and the appropriate substitution model in PAUP* and extracted from the tree file with p4); and (iv) branch lengths from individual sequences to the ancestral node of the Bayesian phylogeny reconstructed with MrBayes.
env), as were estimates from plotting 500 bootstrap trees (1964 to 1974 for gag, 1954 to 1974 for env).

Individual divergent sequences had a large effect when each year contributed equally to the analysis (rather than each individual sequence contributing equally), and we therefore believe that ancestral dates estimated from regression analyses of mean values per sampling year are less accurate than those where all sequences contribute equally. Our ability to estimate the date of the most recent subtype C common ancestor did not rely on the presence of the small number of sequences dating from the early 1980s. However, the estimated date of the most recent common ancestor of the Karonga epidemic with env sequences was severely affected when some of the early sequences were removed (Fig. 1C). Dates estimated from

FIG. 1. Sensitivity plots for Karonga gag (A), entire gag (B), Karonga env (C), and entire env (D). Alignments that had all sequences dating from 1983 and 1984 removed and that had every possible combination of these sequences included were produced. For each alignment, a distance-based heuristic strategy was used (starting trees obtained by random stepwise addition of taxa, 10 replicates) and branch swapping as implemented in PAUP* (35). Branch lengths from individual sequences to the reconstructed ancestor were optimized with PAUP* (35), extracted from the tree file with p4 (8), and plotted against the sampling date.
the gag representative alignment were pushed slightly forward towards the present, while estimates with env sequences were pushed further back into the past. Only 59 sequences were used in each alignment, and the smaller numbers widened the confidence intervals and reduced the significance of each correlation. The large numbers of available sequences from the Karonga data set, which are spread over a number of sampling years, contributed hugely to our ability to estimate the year of the most recent subtype C common ancestor. However, the rate of dS substitutions over time appeared to be slower overall for the Karonga gag data set than for other subtype C sequences.

In general, estimates of the date of origin of the most recent subtype C common ancestor were more consistent between methods used for the gag than for the env alignments. The dates estimated with env ranged from 1945 to 1975, whereas the dates ranged from 1966 to 1974 for gag. Estimates derived from dS data (expected to appear clock-like) ranged from 1967 to 1969 for gag and 1970 to 1975 for env. The later dates inferred with env can be explained by the higher rate of substitution in env (tests for substitution saturation with DAMBE [40] indicated no substitution saturation within the env data). The shorter fragment size used for env (420 bp) compared to gag (620 bp) may also have contributed to this.

Attempts to estimate the date of origin of the most recent common ancestor of the different subtypes have been hampered by a lack of early sequences, lack of enough sequences with known sampling dates, and lack of sequences representing a range of sampling years. Population-based studies in Karonga District, Malawi, have provided a large number of subtype C sequences ranging in date between 1983 and 2000 which have largely overcome these problems for estimating the date of origin of this subtype. While full gene or genome sequences were not available, and while we accept that the large number of Karonga sequences may have biased the reconstruction of the subtype C ancestral sequence as well as the date of the most recent subtype C common ancestor, our data indicate that the most recent common ancestor of subtype C probably originated in the 1960s. The gag data (which we believe is the more reliable) suggest a date between 1966 and 1969 (95% confidence interval, 1959 to 1973). The Karonga epidemic is very diverse and may be a good model for the subtype C epidemic.

ACKNOWLEDGMENTS

Many thanks to Peter Foster for assistance in the use of p4 and to Marco Salemi for critical reading of the manuscript. Thanks also to the two independent reviewers for helpful comments and suggestions.

Until 1996, the Karonga Prevention Study was funded primarily by LEPRA (The British Leprosy Relief Association) and ILEP (The International Federation of Anti-Leprosy Organizations) with contributions from the WHO/UNDP/World Bank Special Programme for Research and Training in Tropical Diseases. Since 1996, the Wellcome Trust has been the principal funder. J.R.G. is funded by the United Kingdom Department of Health (Public Health Career Scientist award).

Permission for the study was received from the National Health Sciences Research Committee, Malawi, and the Ethics Committee of the London School of Hygiene & Tropical Medicine, United Kingdom.