JOURNAL OF VIROLOGY, Dec. 2002, p. 12890–12899 Vol. 76, No. 24
0022-538X/02/$04.00+0 DOI: 10.1128/JVI.76.24.12890–12899.2002

Early Evolution of the Human Immunodeficiency Virus Type 1 Subtype C Epidemic in Rural Malawi

Grace P. McCormack,[1]* Judith R. Glynn,[2] Amelia C. Crampin,[2,3] Felix Sibande,[3] Dominic Mulawa,[3] Lyn Bliss,[2] Philip Broadbent,[2] Katia Abarca,[4] Jorg M. Pönnighaus,[3] Paul E. M. Fine,[2] and Jonathan P. Clewley[1]*

Sexually Transmitted and Blood Borne Virus Laboratory, Central Public Health Laboratory, [1] *and Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine,*[2] *London, United Kingdom; Karonga Prevention Study, Chilumba, Malawi*[3]*; and Departamento de Pediatra, Pontificia Universidad Católica de Chile, Santiago, Chile*[4]

Received 20 June 2002/Accepted 12 September 2002

We have tracked the early years of the evolution of the human immunodeficiency virus type 1 (HIV-1) epidemic in a rural district of central east Africa from the first documented introductions of subtypes A, D, and C to the present predominance of subtype C. The earliest subtype C sequences ever reported are described. Blood samples were collected on filter papers from 1981 to 1984 and from 1987 to 1989 from more than 44,000 individuals living in two areas of Karonga District, Malawi. These samples included HIV-1-positive samples from 200 people. In 1982 to 1984, HIV-1 subtypes A, C, and D were all present, though in small numbers. By 1987 to 1989, 152 (90%) of a total of 168 sequences were subtype C and AC, AD, and DC recombinants had emerged. Four of the subtype C sequences from 1983 to 1984 were closely related and were found at the base of a large cluster of low diversity that by the late 1980s accounted for 40% of C sequences. The other two early C sequences fell into a separate and more diverse cluster. Three other clusters containing sequences from the late 1980s were identified. Each cluster contained at least one sample from a person who had recently arrived in the district. From 18 HIV-1-positive spouse pairs, 12 very closely related pairs of sequences were identified. We conclude that there were multiple introductions of HIV-1 with limited spread, followed by explosive growth of a subtype C cluster, probably arising from a single introduction in or before 1983.

Human immunodeficiency virus type 1 (HIV-1) subtype C is the most prevalent subtype of HIV-1 worldwide (29), yet little is known of its origins. In some countries the HIV-1 epidemic is almost entirely subtype C, which could be due to a "founder" effect and/or to preferential spread (1, 25). Few early HIV-1 samples from Africa have been sequenced, and few molecular-epidemiology studies of HIV-1 have sampled populations rather than groups of patients (who are often unrepresentative). Using population samples collected in two areas of a rural central east African community we have tracked the spread of HIV-1 from the first introductions of subtypes A, D, and C in the early 1980s to the state of subtype C predominance in the late 1980s.

Two detailed total population surveys were carried out in Karonga District, northern Malawi (Fig. 1), in 1981 to 1984 and in 1986 to 1989 to study the prevalence, incidence, and risk factors for leprosy and other mycobacterial infections (32, 33). Finger prick blood spots were collected on filter papers from all individuals of all ages in two areas of the district, one in the north and one in the south (9). They were dried and stored at −20°C. Subsequently, with permission from the Malawi Health Sciences Research Committee, all blood spots dating from 1984 and 1987 to 1989 and those collected from adults between 1981 and 1983 were tested for antibodies to HIV-1 (14). The first HIV-1-positive samples were identified from the 1982 specimens and are the earliest evidence of HIV-1 infection in this region of Africa. The HIV-1 prevalence in those aged 15 to 49 years was 0.1% in 1982 to 1984 and 2% in 1987 to 1989. Currently it is between 10 and 15% (14). In a pilot study we showed that stored filter paper blood spots are suitable for subtyping and sequencing and that the predominant subtype in the area in the1990s was C (J. R. Glynn, K. Abarca, J. P. Clewley, B. Ngwira, S. Malema, D. K. Warndorff, A. C. Crampin, and P. E. M. Fine, Abstr. XIth Int. Conf. AIDS STDs Africa, abstr. 13PT51-3, 1999). This is in line with findings from Lilongwe and Blantyre, the major cities of Malawi (28, 31). We here describe the molecular epidemiology of HIV-1 from two areas of Karonga District in the 1980s.

MATERIALS AND METHODS

Study population. Karonga, the northernmost lakeshore district of Malawi, is a predominately rural district, with the majority of residents being subsistence farmers. Two areas of Karonga District were included in this study, one in the north bordering on Tanzania and the other 70 km to the south, near to but not including the small village and port of Chilumba (Fig. 1). A total of 44,150 finger prick blood spot specimens were collected on filter papers in the 1980s from individuals in these two areas. Some individuals, including three who were HIV-1 positive, were included in both surveys. Eleven individuals were first identified as HIV positive in 1982 to 1984, and 189 were first identified as HIV positive in 1987 to 1989 (14).

* Corresponding author. Mailing address for Grace P. McCormack: Biology Department, National University of Ireland, Maynooth, County Kildare, Ireland. Phone: 353 17083855. Fax: 353 17083845. E-mail: grace.p.mccormack@may.ie. Mailing address for Jonathan P. Clewley: Sexually Transmitted and Blood Borne Virus Laboratory, Central Public Health Laboratory, 61 Colindale Ave., London NW9 5HT, United Kingdom. Phone: 44 20 8200 4400. Fax: 44 20 8200 1569. E-mail: jclewley@phls.org.uk.

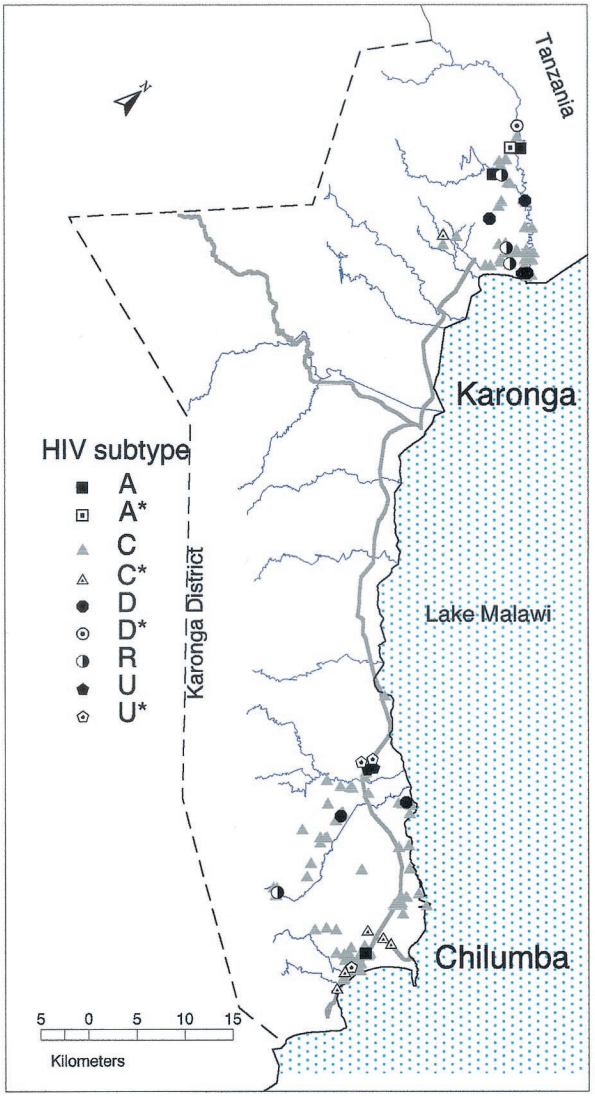


FIG. 1. Map of Karonga District, Malawi, showing the areas included in the study and the distribution of HIV-1 subtypes. *, subtype isolated in 1982 to 1984; the others were isolated in 1987 to 1989. Subtype U, unclassifiable sequences; subtype R, recombinants. Many data points are superimposed on each other because of their proximity, making it difficult to resolve them all.

DNA extraction, PCR, and sequencing. Genomic DNA was extracted from the dried blood spots by a modified Chelex resin method (6). Briefly, a 1-cm^2 portion of a dried blood spot was placed in 1.5 ml of Amplicor wash buffer (Roche Diagnostics, Lewes, United Kingdom) and incubated for 30 min with shaking at 1,000 rpm. The supernatant was removed, and the previous step was repeated until most of the hemoglobin was removed. Five hundred microliters of a 5% Chelex 100 resin solution was added and incubated at 95°C with shaking at 1,000 rpm for 1 h. Samples were spun at 13,000 rpm for 5 min in a Heraeus Biofuge (Fisher Scientific, Loughborough, United Kingdom), and the supernatant was transferred to a Centricon 100 column (Amicon) and concentrated to 40 to 50 μl. One-half of this volume was used in each nested PCR for *env* (C2-to-V3 region; 549 bp) and *gag* (p17 to p24 region; 750 bp) gene fragments (3). PCR products were gel purified with Ultrafree columns (Amicon) and sequenced directly with a Beckman CEQ2000 automatic capillary sequencer (Beckman, High Wycombe, United Kingdom). Sequences were managed with the Lasergene suite of programs (Dnastar Inc., Madison, Wis.).

Sequence alignment and phylogenetic analyses. Multiple alignments were assembled, aligned, and optimized with Bioedit (version 5; T. A. Hall, Department of Microbiology, North Carolina State University, [http://www.mbio.ncsu.edu/BioEdit/bioedit.html]). To assign a subtype, two alignments (*env* and *gag*) were assembled. Each alignment contained all of the Karonga sequences together with 10 reference sequences from the Los Alamos HIV database. All non-C, or putative non-C, sequences from Karonga were assembled in new alignments containing the entire reference alignment sequences from the HIV database and other selected sequences. All Karonga subtype C sequences were assembled in an alignment with African and Indian subtype C sequences, with two subtype A sequences as outgroups.

Phylogenetic reconstructions were then carried out with PAUP*, version 4.0 1. (D. L. Swofford, Sinauer Associates, Sunderland, Mass.). Due to the number of taxa involved, phylogenetic relationships were inferred by using distance matrix analysis. However, for each data set evolutionary models were evaluated with Modeltest (34), which employs likelihood methods to select an optimal model and estimate model parameters. The optimal model of evolution for the alignment of the *gag* subtype C sequences was the Kimura three-parameter model (19) with unequal base frequencies (A, 0.42; C, 0.21; G, 0.21; T, 0.16). The proportion of invariable sites was estimated at 0.53, rate variation among sites was found to follow a gamma distribution with the shape parameter estimated at 0.93, and the rate matrix values were estimated as follows: A-C and G-T, 1.00; A-G and C-T, 3.13; A-T and C-G, 0.64. The chosen model of evolution for the alignment of *env* subtype C sequences was the general time-reversible model (38) with unequal base frequencies (A, 0.41; C, 0.19; G, 0.19; T, 0.22). The proportion of invariable sites was estimated at 0.20, rate variation among sites was found to follow a gamma distribution with the shape parameter estimated at 0.62, and the rate matrix values were estimated as follows: A-C, 1.31; A-G, 4.51; A-T, 0.45; C-G, 0.65; C-T, 3.55; G-T, 1.00.

Tree searches were carried out by using a heuristic search strategy, with both neighbor joining and random stepwise addition of taxa with 10 replicates and branch swapping. Bootstrapping was employed to examine tree robustness, with 1,000 replicates performed on starting trees obtained by neighbor joining. Extensive phylogenetic reconstructions were made with various subsets of sequences by using agreement between the *env* and *gag* trees as a guide to true relationships among sequences. At each step alternative phylogenies were examined by using SplitsTree (D. H. Huson, Faculty of Computer Science, University of Tübingen, [http://www-ab.informatik.uni-tuebingen.de/software/splits/welcome_en.html]), and the reliability of these reconstructions was evaluated by bootstrapping. Diversity of clusters was measured by using mean pair-wise genetic distances (excluding sequences from individuals known to be epidemiologically linked). Pair-wise genetic distances between epidemiologically linked individuals and between the early sequences were also examined.

Epidemiological analyses. Clusters within subtype C were defined from the molecular data independently of the epidemiological information. The distributions of subtypes, clusters, and genetic distances were then compared with demographic characteristics including data on spouses, areas of residence, and travel history.

Nucleotide sequence accession numbers. Sequences have been submitted to the GenBank database under accession no. AY146085 to AY146391.

RESULTS

Subtyping. Extraction and amplification of DNA from all 203 HIV-1-positive dried blood spots from 200 people were attempted. *gag* PCR products were obtained from 170 samples (85% recovery), and *env* PCR products were obtained from 151 samples (74% recovery). Amplicons that showed evidence of deletions and those that were very weakly positive were not sequenced. In total, 159 *gag* products and 141 *env* products from 179 of the 203 samples, including all 11 HIV-positive samples from the early 1980s and repeat samples from the late 1980s from three of these individuals, were sequenced. Sequences from 157 people (158 samples) were subtype C. The other samples were subtypes A and D and recombinants of A, C, and D. Six sequences were of uncertain affinity and will be described elsewhere. The numbers of sequences produced for each subtype and recombinant and unclassifiable strains for each gene region are presented in Table 1.

In 1982 to 1984 subtypes A, D, and C were all present. The

TABLE 1. Numbers of sequences produced during this study for each HIV-1 subtype and recombinant and unclassifiable strains for each time period

Sample period	Region(s) sequenced	No. of sequences that were:					
		Subtype A	Subtype D	Subtype C	Unclassifiable	Recombinant	Total
1981–1984	*gag* only	0	0	1	1	NA[a]	2
	env only	0	0	2	0	NA	2
	gag and *env*	1	1	3	2	0	7
	Total	1	1	6	3	0	11
1986–1989	*gag* only	1	2	32	1	NA	36
	env only	0	1	17		NA	18
	gag and *env*	2	3	103	2	4	114
	Total	3	6	152	3	4	168

[a] NA, not applicable.

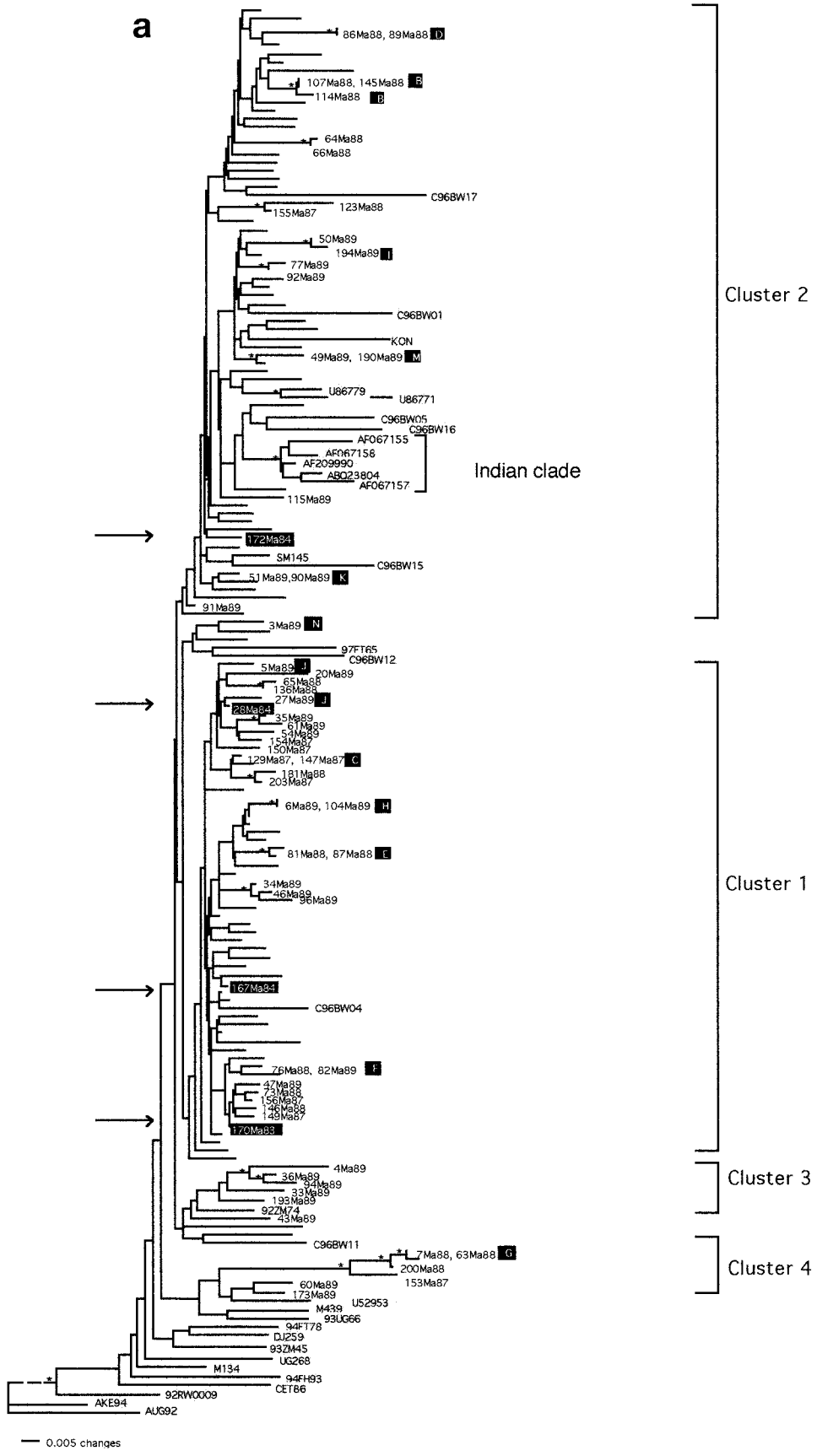
first four positive specimens, from 1982, were one A and one D from the northern area and two from the south that were unclassifiable. Subtype C was first identified in 1983 to 1984 in one individual in the north and five in the south. There was one further unclassifiable infection in the south. Subtype C therefore accounted for 6 of 11 (55%) of the early samples. The individuals with subtypes A and D were both young men who had come into the district in 1982 from Tanzania. Of the six subtype C-infected individuals only one was a long-term resident. The others had arrived within the previous 2 years: two had been born in Zambia and had been living in Blantyre, two had moved from elsewhere in Malawi, and one came from Zambia.

By 1987 to 1989 subtype C accounted for 90% of the cases of HIV infection (152 of 168; $P = 0.004$ compared with the earlier period). There were also three subtype A (including the individual identified in 1982), six subtype D, and three unclassifiable sequences (including one of those from the first survey and her husband). In addition, there were four individuals with *gag-env* recombinants: two AC, one AD, and one DC. All but one of the individuals with the non-C subtypes had direct links with Tanzania or Zambia or lived in the north of the district, near the Tanzanian border. The individuals with unclassifiable sequences were found only in the south.

Phylogenetic reconstruction of the Karonga subtype C sequence data set. The phylogenetic trees produced from analysis of *env* and *gag* sequence alignments are presented in Fig. 2a and b. All subtype C sequences were monophyletic with 100 bootstrap proportion (BP) on both trees. The mean genetic distances between Karonga subtype C sequences were 5.22% (range, 0.16 to 15.11%; standard deviation [SD], 2.17%) for the *gag* gene region and 9.75% (range, 0.5 to 25.2%, SD, 3.22%) for the *env* gene region. The structures of the *gag* and *env* gene trees were very similar, with most of the Karonga sequences holding similar positions on both. A number of clusters could be identified on both trees, although these clusters were not supported by bootstrapping. The majority of Karonga sequences fell into a main clade that was characterized by short internal branch lengths and long terminal branches. This main clade included sequences from the Los Alamos database derived from Malawi and other African countries such as Botswana, Zambia, and South Africa, as well as from India. Outside the main clade was a small outlying cluster of Karonga sequences (cluster 4; see below) in addition to a number of sequences from Uganda, Rwanda, and Djibouti, the 1986 Ethiopian reference sequence, and also a Brazilian reference sequence. Within the large main clade all of the Indian sequences formed a subcluster on each tree, which was supported by 92 BP on the *gag* tree.

Analysis of sequences dating 1982 to 1984. Six HIV-1-positive individuals with subtype C virus were identified in the 1982-to-1984 survey; one sample (sequence 170) was collected in 1983, and the others (sequences 28, 167, 168, 169, and 172) were collected in 1984. These represent the earliest subtype C sequences known to date. From these early subtype C specimens four *gag* sequences were retrieved (sequences 170, 28, 167, and 172). When these four sequences were analyzed alone, in an alignment of 642 bp there was only one informative site. At this site (base position 107) sequences 170 and 167 had an adenine base while 28 and 172 had a thymine base.

FIG. 2. Phylogenetic (neighbor-joining) tree for subtype C sequences from Karonga District, Malawi, based on *gag* (a) and *env* (b) gene sequences. Sequences produced during this study are identified by a number (sample number) followed by Ma (Malawi) and another number (year). Dark grey shading, sequences derived from samples collected in the first survey (1982 to 1984); arrows, early subtype C sequences; white letters in black boxes, spouse pairs (e.g., sequences 7 and 63 in cluster 1 belong to spouse pair G). Sequences MaK30, MaK07, MaF33, MaF37, MaF53, MaF48, MaF65, MaF06, and MaF14 (open boxes) were derived from samples collected in the 1990s from individuals living in the Karonga District. Sequences from the Los Alamos database (light grey shading) are as follows: Malawi, MW965, MW954, MW960, and U07237 (11) and SH750 (4); Senegal, SE365 (21, 22); Somalia, SM145 (21, 22); Zambia, L22954 and SM145 (21, 22), U86770 and U86778 (39), and 93ZM45 (7); Ethiopia, U15061 and CET86 (40) and 94ET93, 94ET78, and 97ET65 (7); South Africa, GOM and KON (4); Brazil, U52953 (12); India, AF067155, AF067158, and AF067157 (20), ABO23804 (24), AF209990 (S. Gupta, K. Arora, A. Gupta, and V. K. Chaudhary; direct on-line submission to GenBank), HIV1D760 and HIV1D747 (8), and HIV100710 (15); Botswana, C96BW11, C96BW02, C96BW16, C96BW17, C96BW04, C96BW15, C96BW01, and C96BW12 (27); Kenya, MM1324 and MM9846 (26); Djibouti, DJ259 (21, 22); Burundi, BU9107, BU9105, and BU9103 (30); Rwanda, 92RW009 (11) and M134 and M439 (18); Uganda, UG268 (21, 22) and 93UG66 (7). AUG92 (13) and AKE94 (35) are subtype A reference sequences used as outgroup sequences; the true lengths of their branches are not indicated on the trees. *, branches supported by bootstrapping (i.e., >70 BP). For clarity, taxa not mentioned in the text or not supported by bootstrapping are not labeled.


a
86Ma88, 89Ma88 D
107Ma88, 145Ma88 B
114Ma88 B
64Ma88
66Ma88
C96BW17
155Ma87
123Ma88
50Ma89
194Ma89 J
77Ma89
92Ma89
C96BW01
KON
49Ma89, 190Ma89 M
U86779
U86771
C96BW05
C96BW16
AF067155
AF067158
AF209990
ABO23804
AF067157
Indian clade
115Ma89
172Ma84
SM145
C96BW15
5Ma89,90Ma89 K
91Ma89
3Ma89 N
97FT65
C96BW12
5Ma89 J
20Ma89
65Ma88
136Ma88
27Ma89 J
28Ma84
35Ma89
61Ma89
54Ma89
154Ma87
150Ma87
129Ma87, 147Ma87 C
181Ma88
203Ma87
6Ma89, 104Ma89 H
81Ma88, 87Ma88 C
34Ma89
46Ma89
96Ma89
167Ma84
C96BW04
76Ma88, 82Ma89 F
47Ma89
73Ma88
156Ma87
146Ma88
149Ma87
170Ma88
4Ma89
36Ma89
94Ma89
33Ma89
193Ma89
92ZM74
43Ma89
C96BW11
7Ma88, 63Ma88 G
200Ma88
153Ma87
60Ma89
173Ma89
U52953
M439
93UG66
94FT78
DJ259
93ZM45
UG268
M134
94FH93
CET86
92RW0009
AKE94
AUG92
Cluster 2
Cluster 1
Cluster 3
Cluster 4
0.005 changes

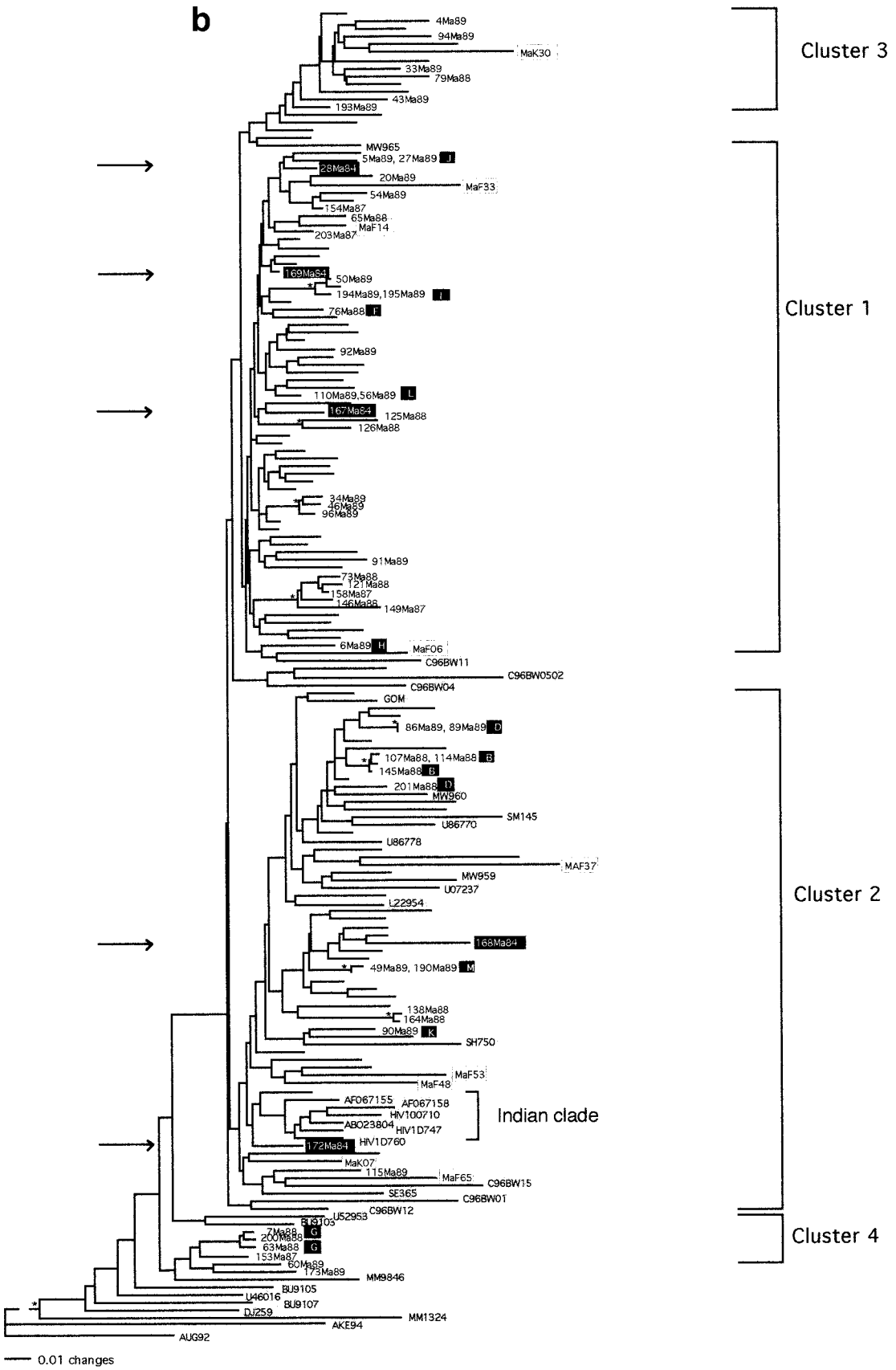


FIG. 2—*Continued.*

TABLE 2. Genetic distances within and between early subtype C sequences from the Karonga 1982-to-1984 survey

Cluster	Genetic distance (%) (range [mean] or mean) for:	
	gag	*env*
1	0.94–1.26 (1.15)	2.69–4.47 (3.45)
2		8.3
1 versus 2	3.38	5.0–11.29 (8.10)

However, sequence 172 was far more divergent than the other three, lacking the codon for an amino acid (positions 238 to 240) and having 17 unique base pair changes compared to 3 or 4 unique changes for the remaining three sequences. The genetic distance between sequences 28, 167, and 170 ranges from 0.9% to 1.2%, while the distances from each of these to sequence 172 is approximately 3.4% (Tables 2 and 3). Five early subtype C sequences were available for analysis of the *env* gene (sequences 28, 167, 168, 169, and 172). Of 13 informative sites in a 420-bp alignment, eight joined sequences 172 and 168 to the exclusion of the other three, and these two sequences also had more unique base changes than the other three, resulting in greater terminal branch lengths. The genetic distances between sequences 28, 167, and 169 ranged from 2.7 to 4.4%, while the distance between 172 and 168 was 8.3%. The distances between the two groups of sequences ranged from 5 to 11% (Table 2).

TABLE 3. Pairwise comparisons of sequences from epidemiologically linked individuals from the Karonga 1986-to-1989 survey, with regard also to the clusters in which they fall

Sequences compared (spouse pair)	Cluster[a] or subtype	Mean genetic distance (%) for:	
		gag	*env*
Spouses			
84 and 85 (a)	Unclassifiable	1.82	7.44
107 and 114 (b)	2	0	0.5
107 and 145 (B)	2	0.48	0.5
129 and 147 (c)	2	0.8	
89 and 201 (d)	2		5
89 and 86 (d)	2	0	0
87 and 81 (e)	1	0.64	
76 and 82 (f)	1	1.64	
7 and 63 (g)	4	0.32	1
6 and 104 (h)	1	0	
194 and 195 (i)	1		1
5 and 27 (j)	1	2.29	5
51 and 90 (k)	2	2	
110 and 56 (l)	1		3.8
190 and 49 (m)	2	1.65	0.5
3 and 98 (n)	1/2		
53 and 182 (p)	2	2.36	
130 and 144 (q)	D	0.63	
Siblings			
91 and 92	3	1.2	7.43
91 and 115	3/4	3.2	11.26
92 and 115	3/4	3.9	10.81
Mother and daughter			
154 and 155	3/4	5.4	4

[a] Two cluster numbers separated by a shill indicate that one sequence is in one cluster, and one is in another.

Four of the early sequences (sequences 28, 167, 169, and 170) were part of cluster 1 (Fig. 2) and were found in both the north and the south of the district. Three of these early cluster 1 sequences came from individuals who had come from other regions of Malawi, while sequence 170 came from a long-term resident of Karonga. Two early sequences were part of cluster 2 (Fig. 2). Sequence 168 came from a young woman who was born in Zambia but had been living in Blantyre; the other (172) came from a young man from Zambia. Both were living in the south of the district.

Analysis of entire sequence data set, 1982 to 1989. The structures of the two gene trees were similar, and four clusters containing many of the same sequences could be identified on each. However, when all sequences were included, there was no support for any of the clusters and most internal branches collapsed under bootstrapping.

Cluster 1 accounted for 40% of the HIV-1 strains identified in the 1987-to-1989 survey. Most sequences formed complex sets of relationships with other sequences, which resulted in low bootstrap support for relationships within the cluster. The grouping of sequences 34, 46, and 96 was highly supported on both trees (Fig. 2). The only possible epidemiological link between these is that samples 34 and 46 came from the same village. The early *gag* sequence 28 was found to be at the center of a network containing sequences 5, 20, 27, 28, 35, 54, 61, 65, 136, 150, and 154, and sequence 170 was found basal to a group containing sequences 47, 146, 149, 156, and 73. However sequences 28 and 170 were also very closely related to each other, and networks linked all of above groups, which resulted in low or no bootstrap support. With the *env* gene a group containing sequences 73, 146, 149, 158, and 121 was highly supported. The *gag* gene for sample 158 was not amplified, and sample 121 is a recombinant (being subtype A in *gag*). The only non-Malawi sequence in cluster 1 on the *gag* tree was C96BW04 (Fig. 2a). This sequence was more divergent on the *env* tree, grouping outside and basal to cluster 1 with sequences 89 and C96BW0502 (Fig. 2b). On the *env* tree (Fig. 2b) the sequence C96BW11 was the only non-Malawi sequence in cluster 1, and on the *gag* tree (Fig. 2a) this sequence was placed with sequences from cluster 3. The mean genetic distance between epidemiologically unlinked sequences in cluster 1 was 3.13% (range, 0.31 to 9.22%; SD, 1.33%) for *gag* and 6.54% (range, 1.58 to 14.8%; SD, 2%) for *env*.

Smaller groups of sequences were also apparent in cluster 2, but the sequences were more diverse. The mean genetic distance between epidemiologically unlinked sequences in this cluster was 4.33% (range, 0.15 to 8.76%; SD, 1.08%) for *gag* and 9.23% (range, 0.48 to 20.57%; SD, 2%) for *env*. The majority of relationships within this cluster were not at all resolved, and bootstrap support was only present, for the most part, for the spouse pairs. Three pairs of sequences were highly supported on the *gag* tree, and one was highly supported on the *env* tree; all had unknown epidemiological links. All of the sequences branched off separately from any network that we drew and did not cluster according to village, region, or date. This cluster contained most of the Los Alamos database sequences included in the study; these were not included in the calculations of genetic distances.

Clusters 3 and 4 were first identified only in the second survey (1987 to 1989). Cluster 3 (containing sequences 4, 33,

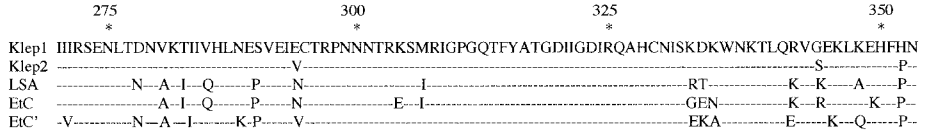


FIG. 3. Amino acid sequence comparison between subtype C consensus sequences of the Karonga 1982-to-1984 survey (Klep1), the Karonga 1987-to-1989 survey (Klep2), the Los Alamos database subtype C (LSA) sequences, and the Ethiopian clusters (EtC and EtC′). The latter three were taken from the study of Abebe et al. (2).

43, 94, and 193 on both trees) was found only in the south and included sequences from two immigrants to the district and from two who had been living outside the district in the early 1980s. Sequences 33 and 43 came from a male immigrant and female resident from the same village. However the relationships within the cluster in general were not highly supported and did not reflect the location of the individuals. In cluster 4 individuals with sequences 7 and 63 were spouses and sequence 200 came from a female living in the same village in the north. This small group had high bootstrap support. Sequence 153, which came from an individual from Tanzania who lived in the north, was also associated with this cluster on both *gag* (99 BP) and *env* trees. Two other Karonga sequences, 60 and 173, were always in a similar, but unsupported, position in this cluster. These samples came from two women in the southern part of Karonga District, one of whom had come from Tanzania in 1986.

Eighteen sequences were found to form sets of relationships on the *gag* tree very different from those on the *env* tree. When these 18 sequences were excluded from the *gag* alignment, cluster 1 had 64 BP and cluster 3 had 63 BP, although relationships within each cluster remained largely undetermined. Cluster 4 sequences were quite divergent from the rest of the Karonga sequences, and, when they were removed, support for cluster 3 increased to 74 BP, while support for clusters 1 and 3 as a monophyletic group increased to 80 BP. There was no support at any stage for cluster 2 as a monophyletic group.

Spouses and households. Sequence data for 18 spouse pairs (including, as four pairs, two men who each had two wives, pairs b and d) were available. One pair was subtype D (pair q) and another was unclassifiable (pair a). The rest were subtype C, and for all but one of these 16 pairs the members were in the same cluster (Fig. 2). For the one pair with differing clusters, only *env* (sequence 98) was available for one member and only *gag* (sequence 3) was available for the other, so no direct comparison could be made. Where direct comparison was possible, the mean distance in *gag* was 1.0% (SD, 0.92%) and the mean distance in *env* was 1.9% (SD, 2.1%). This compares with a mean distance of 5.3% (SD, 1.6%) in *gag* and 9.2% (SD, 2.8%) in *env* for unrelated pairs. As seen in Table 3 the genetic distances ranged from 0 to 2.36% for *gag* sequences and from 0 to 7.44% for *env* sequences. For 12 subtype C spouse pairs the genetic distances between their sequences were very small, for *gag* ranging from 0 to 2% and for *env* ranging from 0 to 1%. In spouse group d, sequences from the man and his first wife (sequences 89 and 86) were identical in both *env* and *gag*, suggesting very recent transmission. However, the husband's sequence differed by 5% in *env* from that of his second wife (sequence 201). The pair-wise distances between sequences from another two spouse pairs (pairs j and p) were over twice the mean spouse pair distance. While spouse pair j grouped together on the *env* tree, neither this pair nor spouse pair p (53 and 182) was grouped on the *gag* tree. While sequences within each of these three spouse pairs did not group directly with each other, they were present in the same small group of sequences within a cluster.

Three other households contained more than one individual with HIV-1 sequence data available. Two adult sisters (sequences 91 and 92) had similar sequences (1.2% difference in *gag*, 7.43% difference in *env*), but the sequences did not group together on either tree, while the sequence from their brother (115) was unrelated at 3 to 4% difference in *gag* and 11% difference in *env* (Table 3). The sequence differences between a mother and her adult daughter (sequences 154 and 155) were 5.4 (in *gag*) and 4% (in *env*), and these did not group together on either tree. A father and his 8-year-old son had different subtypes. The individuals with sequence data available came from 42 different villages (defined by allegiance to a village headman). In some cases sequences from epidemiologically unlinked individuals were closely related to spouse pairs. Sequence 200 is associated with spouse pair g and comes from a female living in the same village, and sequence 50 is associated with pair l and comes from a male from the same village (Fig. 2b). There was no correlation between clusters and village. All but one of the villages containing more than two HIV-1-positive individuals included multiple clusters. The one exception contained four subtype C, cluster 2, sequences, which were at least 3% apart in the *gag* gene.

V3 loop. The predicted consensus amino acid sequences of the envelope gp120 V3 region for 1984 and the period 1987 to 1989 are shown in Fig. 3 aligned with those of the Los Alamos subtype C, Ethiopian subtype C, and Ethiopian subcluster C′ consensus sequences (2). For the Karonga sequences, the most frequent amino acid was included in the consensus for each sequence position, even if it was not present in the majority of sequences. The consensus sequences were found to vary at 19 positions (Fig. 3). At each of these positions, each displayed amino acid was present at that position in at least one of the Karonga sequences at each time period and in most cases was also common. For example, at position 278 in Fig. 3, both D and N were present among the sequences collected in the 1982-to-1984 and 1986-to-1989 time periods. The 1984 consensus sequences differed from the 1987-to-1989 sequences at three positions (294, 344, and 352). At positions 294 and 352 sequences 168 and 172 differed from the other three early sequences and had the same amino acids as those present in the 1987-to-1989 consensus sequence. For position 344 sequences 28 and 167 had the same amino acid as those in the later survey. The GPGQ motif, typical of subtypes A and C, at the crown of the V3 loop (positions 309 to 312; Fig. 3) was

conserved in the Karonga data set, being present in all of the sequences. While the majority of sequences had the alanine-to-threonine substitution at the carboxy-terminal part of the GPGQ sequence, many sequences retained the alanine residue. An isoleucine (I) substitution was present in one sequence, and a serine (S) substitution was present in two sequences. Again, as found in most subtype C sequences, none of the Karonga sequences showed an excess of basic substitutions in the V3 region, suggesting that they all have a non-syncytium-inducing phenotype. Between positions 332 and 333 of the alignment shown in Fig. 3 an extra amino acid was present in two sequences. The husband (sequence 86) and wife (sequence 89) pair had an insert of AGT (serine) at this position, which was not present in the sequence of the husband's second wife (201) or any other subtype C sequence examined.

DISCUSSION

We have studied the early evolution of the HIV-1 epidemic in the population of two areas in rural Malawi, characterized the earliest HIV-1-positive samples in the region, and found the earliest subtype C samples so far recorded. Other studies have described the distribution and characteristics of subtypes and strains, and some have compared them in different groups in the population (36, 41, 42), but this is one of the first studies to link sequence analysis with detailed population-based epidemiological data (43).

We have shown that subtypes A and D were present in northern Malawi as early as 1982, whereas the earliest subtype C sequence was from a sample collected in 1983. Five more subtype C samples were collected in 1984. Among the first previously recognized subtype C sequences were those found in neighboring Zambia in the early 1990s (23). It is reasonable to assume that subtype C spread from central Africa to both eastern and southern Africa (and subsequently to Asia) (8, 37), and it has been estimated that subtype C was introduced into Ethiopia in 1983 (1). Subtypes A, C, and D predominate in southern Tanzania (16), and A and D predominate in most of east Africa (17).

The 1987-to-1989 survey found only eight more individuals who had subtype A or subtype D virus, and by that time there were also four recombinant viruses of subtypes A, D, and C present in the region. On the other hand, 90% of sequences were subtype C, and non-C subtypes were more common in the north, bordering on Tanzania. During the 1980s there were many construction activities in the district involving temporary immigrants, including construction of a main road, completed in 1988, which runs through Karonga district to the Tanzanian border. Traffic between Karonga and Tanzania is regular. The greater success of subtype C in establishing itself in the district could be due to a greater initial number of subtype C viruses in the susceptible population (a stronger founder effect) (10) or could involve other factors such as higher transmissibility (5). Although this study does not resolve these alternatives, the early presence of a mixed population of subtypes and the later predominance of subtype C support the view that subtype C is more successful than other subtypes.

The mean genetic distance between Karonga *gag* sequences (5.22%) was lower than that between the sequences from Botswana (9.4%) but higher than that between Indian sequences (3.7%). Similarly, for the *env* C2-V3 region, the Karonga data set had a mean distance of 9.75% while the Botswana and the Indian sequences had mean distances of 15.7 and 6.2%, respectively. The Karonga sequences were collected in the 1980s, while the other subtype C sequences were collected in the 1990s. This may be reflected in the lower genetic-distance estimates and may further suggest that in the 1980s the epidemic in Malawi progressed a little further than the epidemic in India in the 1990s. However, the range of distances in the Karonga data set was relatively large: from <1 to >15% for *gag* and >25% for *env*. The maximum genetic distance among Botswana sequences, in comparison, was 11% for *gag* and 21% for *env*, suggesting that some of the Karonga sequences were very diverse (e.g., sequences in cluster 4) but that there are also many closely related sequences present, giving an overall lower mean value.

Similar groups of sequences were evident on both *env* and *gag* gene trees and were labeled clusters 1 to 4 even though their existence did not have strong bootstrap support when all Karonga and database sequences were included in the analysis. However, when database sequences, which largely dated from the 1990s, were removed together with the more divergent Karonga sequences (e.g., cluster 4) there was some high support for clusters 1 and 3 but none for cluster 2.

The six early subtype C sequences in Karonga fell into two groups, which corresponded to clusters 1 and 2 and which were separated with 99 BP when analyzed on their own. The genetic distances between four of the early sequences (sequences 28, 167, 169, and 170), all positioned in cluster 1, were less than the distances between some spouse pairs in the study (Table 2; approximately 1% for *gag* and 2.7 to 4.5% for *env*), which may, therefore, be consistent with a local transmission event. If this is true, then 40% of HIV-1 infections present in the survey areas in the late 1980s may have arisen from a single subtype C introduction in 1983 or earlier. Their positions on the trees and their behavior during split-decomposition analyses suggest that sequence 170 is basal to one sequence group and that sequence 28 is central to another. Whether this illustrates some earlier spread of the epidemic among the infected individuals is arguable. Of potential interest is the fact that all of these early samples originated from individuals who had come from Malawi and that sequences in cluster 1 were confined, for the most part, to those from Malawi. The only exception on the *gag* tree was a divergent clone from Botswana, CW98B04, and the only exception on the *env* tree was divergent sequence CW96B11. The positions of these divergent sequences with the two gene regions were not congruent, however, and their true status remains undetermined.

The other two early sequences, 172 and 168, were diverse (differing by 8% in the *env* gene), and both came from new arrivals to the district. It is likely, therefore, that these sequences, both dating from 1984, represent two separate introductions of HIV-1 into the district. These sequences fell into cluster 2 and did not appear basal or central to any sequence groups present in the cluster. This cluster also contained many of the database sequences (e.g., from Zambia, Botswana, Somalia, and India) included in the study, and, in general, the range and mean of genetic distances among sequences in cluster 2 were greater than those in cluster 1.

Clusters 3 and 4 were identified for the first time from

sequences obtained in the late 1980s. Each contained a small number of sequences and was restricted to one area of Karonga district. Cluster 3, related to clusters 1 and 2, fell within the main subtype C clade, but cluster 4 occupied a position outside this main group. This cluster contained two sequences from individuals who had traveled from Tanzania. This might explain their position away from the main clade. The finding that these viruses had not spread further into the population may be due to their more recent introduction or to the presence of more-limited sexual networks.

Of the spouse pairs sequenced, the majority had closely related sequences consistent with transmission between spouses. Five spouse groups (a, j, n, p, and the second wife in spouse group d) had less closely related sequences. They may have been infected from different sources, as has been found previously (43), but this would be surprising at a time when HIV prevalence was low. Genetic distance between spouse sequences is also influenced by the interval of time since transmission occurred, which is largely unknown. However, spouse pair j had been seen 5 years previously, in 1984. At this time they did not live together but had had a child together. She was HIV positive at this time (sequence 28), and he was negative. It is likely that the large genetic distance between sequences 5 and 27 is due to evolution proceeding for some time within each individual between 1984 and 1989. Since spouse pair a married in 1982 and the wife was then positive (sampled during the 1981-to-1984 survey; sequences were in the unclassifiable group), it is also possible that her husband may have become positive quite some time before they were both seen again in 1988. Transmission of minor viral variants could also result in a different viral spectrum in the recipient, and this may account for the large genetic distances seen in pairs p and d (89 and 201) although we cannot rule out the possibility that transmission from different sources occurred. The epidemiological history of the father-and-son pair who, by phylogenetic analysis, had different HIV-1 subtypes of DNA suggested that the son contracted HIV-1 from his mother and that his father contracted it from a different source. Unfortunately a sample could not be collected from the mother as she never left Zambia and is reported to have died.

We would expect that the genetic distances measured from the *env* gene would be higher than those from the *gag* gene because of the higher rate of substitution occurring in this gene. However for spouse pair m the opposite is evident, as the genetic distance is 1.65% for *gag* but only 0.5% for *env*. Spouse group b also shows some inconsistencies. The genetic distance between sequences from the two wives (sequences 114 and 145) calculated from the *env* gene is less than the distance between the sequence from each wife and that from the husband, sequence 107 (0.2 versus 0.5%; Table 1), but the genetic distance calculated from the *gag* gene is larger (1% compared with 0 and 0.48%). These inconsistencies may be due simply to sampling effects or may point to methodological problems that should be taken into account when using genetic distances (especially from a single gene region) to investigate viral transmission events.

In summary, we have identified an emergent subtype C cluster (cluster 1), which may have had a single introduction into Karonga district in the early 1980s and then diversified to account for 40% of HIV-1 infections in the area by the late 1980s. Although subtypes A and D were present early in the epidemic, they did not spread in a comparable fashion. We have also identified a more diverse subtype C cluster, probably of Tanzanian origin (cluster 4), which may have been introduced twice. Cluster 3 may have arisen from within cluster 1, while cluster 2 probably arose through multiple introductions of subtype C from neighboring regions. Many local epidemics are dominated by one or a few subtypes of HIV-1 (e.g., subtypes E and B in Thailand and subtype B in the United States), a restriction which is attributable to a combination of chance, the initial or repeated introduction of virus strains of limited diversity, or a slight selective advantage of one or another strain. The general predominance of subtype C in Africa may suggest a selective advantage, and in this study we have traced its rise and identified a particularly successful cluster.

ACKNOWLEDGMENTS

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 and funded this study. J.R.G. is supported in part by the British Department for International Development.

We thank the Government of the Republic of Malawi for their interest in and support of the Project and the Malawi Health Sciences Research Committee for permission to publish the paper. We thank Keith Branson for the map.

REFERENCES

1. **Abebe, A., V. V. Lukashov, G. Pollakis, A. Kliphuis, A. L. Fontanet, J. Goudsmit, and T. F. R. de Wit.** 2001. Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. AIDS **15:**1555–1561.
2. **Abebe, A., G. Pollakis, A. L. Fontanet, B. Fisseha, B. Tegbaru, A. Kliphuis, G. Tesfaye, H. Negassa, M. Cornelissen, J. Goudsmit, and T. F. R. de Wit.** 2000. Identification of a genetic subcluster of HIV type 1 subtype C (C′) widespread in Ethiopia. AIDS Res. Hum. Retrovir. **16:**1909–1914.
3. **Barlow, K. L., I. D. Tatt, P. A. Cane, D. Pillay, and J. P. Clewley.** 2001. Recombinant strains of HIV-1 in the UK. AIDS Res. Hum. Retrovir. **17:** 467–474.
4. **Becker, M. L. B., G. De Jager, and W. B. Becker.** 1995. Analysis of partial *gag* and *env* gene sequences of HIV type 1 strains from southern Africa. AIDS Res. Hum. Retrovir. **11:**1265–1267.
5. **Bjorndal, A., A. Sonnerborg, C. Tscherning, J. Albert, and E. M. Fenyo.** 1999. Phenotypic characteristics of human immunodeficiency virus type 1 subtype C isolates of Ethiopian AIDS patients. AIDS Res. Hum. Retrovir. **15:**647–653.
6. **Cassol, S., B. G. Weniger, G. Babu, M. O. Salminen, X. Zheng, M. T. Htoon, A. Delaney, M. O'Shaughnessy, and C.-Y. Ou.** 1996. Detection of HIV type 1 *env* subtypes A, B, C, and E in Asia using dried blood spots: a new surveillance tool for molecular epidemiology. AIDS Res. Hum. Retrovir. **12:**1435–1441.
7. **De Baar, M. P., A. De Ronde, B. Berkhout, M. Cornelissen, K. H. van der Horn, A. M. van der Schoot, F. De Wolf, V. V. Lukashov, and J. Goudsmit.** 2000. Subtype-specific sequence variation of the HIV type 1 long terminal repeat and primer-binding site. AIDS Res. Hum. Retrovir. **16:**499–504.
8. **Dietrich, U., M. Grez, H. von Briesen, B. Panhans, M. Geißendorfer, H. Kuhnel, J. Maniar, G. Mahambre, W. B. Becker, M. L. B. Becker, and H. Rübsamen-Waigmann.** 1993. HIV-1 strains from India are highly divergent from prototypic African and US/European strains, but are linked to a South African isolate. AIDS **7:**23–27.
9. **Fine, P. E. M., J. M. Pönnighaus, P. Burgess, J. A. Clarkson, and C. C. Draper.** 1988. Seroepidemiological studies of leprosy in northern Malawi based on an enzyme-linked immunosorbent assay using synthetic glycoconjugate antigen. Int. J. Lepr. **56:**243–254.
10. **Foley, B., H. Pan, S. Buchbinder, and E. L. Delwart.** 2000. Apparent founder effect during the early years of the San Francisco HIV type 1 epidemic (1978–1979). AIDS Res. Hum. Retrovir. **16:**1463–1469.
11. **Gao, F., S. G. Morrison, D. L. Robertson, C. L. Thornton, S. Craig, G. Karlsson, J. Sodroski, M. Morgado, B. Galvao-Castro, H. von Briesen, S. Beddows, J. Weber, P. M. Sharp, G. M. Shaw, B. H. Hahn, and the WHO and**

NIAID Networks for HIV Isolation and Characterization. 1996. Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. J. Virol. **70:**1651–1667.
12. **Gao, F., D. L. Robertson, C. D. Carruthers, S. G. Morrison, B. Jian, Y. Chen, F. Barré-Sinoussi, M. Girard, A. Srinivasan, A. G. Abimiku, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1998. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J. Virol. **72:**5680–5698.
13. **Gao, F., D. L. Robertson, S. G. Morrison, H. Hui, S. Craig, J. Decker, P. N. Fultz, M. Girard, G. M. Shaw, B. H. Hahn, and P. M. Sharp.** 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J. Virol. **70:**7013–7029.
14. **Glynn, J. R., J. M. Pönnighaus, A. C. Crampin, F. Sibande, L. Sichali, P. Nkhosa, P. Broadbent, and P. E. M. Fine.** 2001. The development of the HIV epidemic in Karonga District, Malawi. AIDS **15:**2025–2029.
15. **Grez, M., U. Dietrich, P. Balfe, H. von Briesen, J. K. Maniar, G. Mahambre, E. L. Delwart, J. I. Mullins, and H. Rübsamen-Waigmann.** 1994. Genetic analysis of human immunodeficiency virus type 1 and 2 (HIV-1 and HIV-2) mixed infections in India reveals a recent spread of HIV-1 and HIV-2 from a single ancestor for each of these viruses. J. Virol. **68:**2161–2168.
16. **Hoelscher, M., B. Kim, L. Maboko, F. Mhalu, F. von Sonnenburg, D. L. Birx, and F. E. McCutchan.** 2001. High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. AIDS **15:** 1461–1470.
17. **Janssens, W., A. Buvé, and J. N. Nkengasong.** 1997. The puzzle of HIV-1 subtypes in Africa. AIDS **11:**705–712.
18. **Kampinga, G. A., A. Simonon, P. van de Perre, E. Karita, P. Msellati, and J. Goudsmit.** 1997. Primary infections with HIV-1 women of and their offspring in Rwanda: findings of heterogeneity at seroconversion, coinfection and recombination of HIV-1 subtypes A and C. Virology **227:**63–76.
19. **Kimura, M.** 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78:**454–458.
20. **Lole, K. S., R. C. Bollinger, and R. S. Paranjape.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73:** 152–160.
21. **Louwagie, J., W. Janssens, J. Mascola, L. Heyndrickx, P. Hegerich, G. van der Groen, F. E. McCutchan, and D. S. Burke.** 1995. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. J. Virol. **69:**263–271.
22. **Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, E. Sanders-Buell, G. A. Eddy, G. van der Groen, K. Fransen, G.-M. Gershy-Damet, R. Deleys, and D. S. Burke.** 1993. Phylogenetic analysis of *gag* genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS **7:**769–780.
23. **McCutchan, F. E., B. L. P. Ungar, P. Hegerich, C. R. Roberts, A. K. Fowler, S. K. Hira, P. L. Perine, and D. S. Burke.** 1992. Genetic analysis of HIV-1 isolates from Zambia and an extended phylogenetic tree for HIV-1. J. Acquir. Immune Defic. Syndr. **5:**441–449.
24. **Mochizuki, N., N. Otsuka, K. Matsuo, T. Shiino, A. Kojima, T. Kurata, K. Sakai, N. Yamamoto, S. Isomura, T. N. Dhole, Y. Takebe, M. Matsuda, and M. Tatsumi.** 1999. An infectious DNA clone of HIV type 1 subtype C. AIDS Res. Hum. Retrovir. **15:**1321–1324.
25. **Morison, L., A. Buvé, L. Zekeng, L. Heyndrickx, S. Anagonou, R. Musonda, M. Kahindo, H. A. Weiss, R. Hayes, M. Laga, W. Janssens, and G. van der Groen.** 2001. HIV-1 subtypes and the HIV epidemics in four cities in sub-Saharan Africa. AIDS **15**(Suppl. 4)**:**S109–S116.
26. **Neilson, J. R., G. C. John, J. K. Carr, P. Lewis, J. K. Kreiss, S. Jackson, R. W. Nduati, D. Mbori-Ngacha, D. D. Panteleeff, S. Bodrug, C. Giachetti, M. A. Bott, B. A. Richardson, J. Bwayo, J. Ndinya-Achola, and J. Overbaugh.** 1999. Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. J. Virol. **73:**4393–4403.
27. **Novitsky, V. A., M. A. Montano, M. F. McLane, B. Renjifo, F. Vannberg, B. T. Foley, T. P. Ndung'u, M. Rahman, M. J. Makhema, R. Marlink, and M. Essex.** 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J. Virol. **73:**4427–4432.
28. **Orloff, G. M., M. L. Kalish, J. Chiphangwi, K. E. Potts, C. Y. Ou, G. Schochetman, G. Dallabetta, A. I. Saah, and P. G. Miotti.** 1993. V3 loops of HIV-1 specimens from pregnant women in Malawi uniformly lack a potential N-linked glycosylation site. AIDS Res. Hum. Retrovir. **9:**705–706.
29. **Osmanov, S., C. Pattou, N. Walker, B. Schwardlander, and J. Esparza.** 2002. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J. Acquir. Immune Defic. Syndr. **29:**184–190.
30. **Penny, M. A., S. J. Thomas, N. W. Douglas, S. Ranjbar, H. Holmes, and R. S. Daniels.** 1996. *env* gene sequences of primary HIV type 1 isolates of subtypes B, C, D, E, and F obtained from the World Health Organization Network for HIV Isolation and Characterization. AIDS Res. Hum. Retrovir. **12:**741–747.
31. **Ping, L. H., J. A. Nelson, I. F. Hoffman, J. Schock, S. L. Lamers, M. Goodman, P. Vernazza, P. Kazembe, M. Maida, D. Zimba, M. M. Goodenow, J. J. J. Eron, S. A. Fiscus, M. S. Cohen, and R. Swanstrom.** 1999. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. J. Virol. **73:**6271–6281.
32. **Pönnighaus, J. M., P. E. M. Fine, and L. Bliss.** 1993. The Karonga Prevention Trial: a leprosy and tuberculosis vaccine trial in northern Malawi. I. Methods of the vaccination phase. Lepr. Rev. **64:**338–356.
33. **Pönnighaus, J. M., P. E. M. Fine, L. Bliss, I. J. Sliney, D. J. Bradley, and R. J. W. Rees.** 1987. The Lepra Evaluation Project (LEP), an epidemiological study of leprosy in northern Malawi. I. Methods. Lepr. Rev. **58:**359–375.
34. **Posada, D., and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
35. **Poss, M., A. G. Rodrigo, J. J. Gosink, G. H. Learn, D. D. V. Panteleeff, H. L. Martin, Jr., J. Bwayo, J. K. Kreiss, and J. Overbaugh.** 1998. Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. J. Virol. **72:** 8240–8251.
36. **Rayfield, M. A., R. G. Downing, J. Baggs, D. J. Hu, D. Pieniazek, C.-C. Luo, B. Biryahwaho, R. A. Otten, S. D. K. Sempala, T. J. Dondero, and the HIV Variant Working Group.** 1998. A molecular epidemiologic survey of HIV in Uganda. AIDS **12:**521–527.
37. **Rodenburg, C. M., Y. Li, S. A. Trask, Y. Chen, J. Decker, D. L. Robertson, M. L. Kalish, G. M. Shaw, S. Allen, B. H. Hahn, F. Gao, and the UNAIDS and NIAID Networks for HIV Isolation and Characterization.** 2001. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res. Hum. Retrovir. **17:**161–168.
38. **Rodriguez, F., J. F. Oliver, A. Marin, and J. R. Medina.** 1990. The general stochastic model of nucleotide substitutions. J. Theor. Biol. **142:**485–501.
39. **Salminen, M. O., J. K. Carr, D. L. Robertson, P. Hegerich, D. Gotte, C. Koch, E. Sanders-Buell, F. Gao, P. M. Sharp, B. H. Hahn, D. S. Burke, and F. E. McCutchan.** 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. J. Virol. **71:**2647–2655.
40. **Salminen, M. O., B. Johansson, A. Sönnerberg, S. Ayehunie, D. Gotte, P. Leinikki, D. S. Burke, and F. E. McCutchan.** 1996. Full-length sequence of an Ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. AIDS Res. Hum. Retrovir. **12:**1329–1339.
41. **van Harmelen, J. H., E. van der Ryst, A. S. Loubser, D. York, S. Madurai, S. Lyons, R. Wood, and C. Williamson.** 1999. A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. AIDS Res. Hum. Retrovir. **15:**395–398.
42. **Weniger, B. G., Y. Takebe, C.-Y. Ou, and S. Yamazaki.** 1994. The molecular epidemiology of HIV in Asia. AIDS **8**(Suppl. 2)**:**S13–S28.
43. **Yirrell, D. L., H. Pickering, G. Palmarini, L. Hamilton, A. Rutemberwa, B. Biryahwaho, J. Whitworth, and A. J. Leigh Brown.** 1998. Molecular epidemiological analysis of HIV in sexual networks in Uganda. AIDS **12:**285–290.