JOURNAL OF VIROLOGY, Feb. 2003, p. 2587–2599 Vol. 77, No. 4
0022-538X/03/$08.00+0 DOI: 10.1128/JVI.77.4.2587–2599.2003

Molecular Characteristics of Human Immunodeficiency Virus Type 1 Subtype C Viruses from KwaZulu-Natal, South Africa: Implications for Vaccine and Antiretroviral Control Strategies

M. Gordon,[1] T. De Oliveira,[1] K. Bishop,[1] H. M. Coovadia,[2] L. Madurai,[3] S. Engelbrecht,[4] E. Janse van Rensburg,[4] A. Mosam,[5] A. Smith,[6] and S. Cassol[1,7]*

HIV-1 Molecular Virology and Bioinformatics Laboratories, Africa Centre for Health and Population Studies and the Nelson R. Mandela School of Medicine,[1] Centre for HIV/AIDS Networking,[2] Department of Dermatology,[5] and Department of Virology,[6] University of Natal, and Medical Research Council,[3] Durban, and Department of Medical Virology, University of Stellenbosch and Tygerberg Hospital, Tygerberg,[4] South Africa, and Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom[7]

Received 29 July 2002/Accepted 7 November 2002

The KwaZulu-Natal region of South Africa is experiencing an explosive outbreak of human immunodeficiency virus type 1 (HIV-1) subtype C infections. Understanding the genetic diversity of C viruses and the biological consequences of this diversity is important for the design of effective control strategies. We analyzed the protease gene, the first 935 nucleotides of reverse transcriptase, and the C2V5 envelope region of a representative set of 72 treatment-naïve patients from KwaZulu-Natal and correlated the results with amino acid signature and resistance patterns. Phylogenetic analysis revealed multiple clusters or "lineages" of HIV-1 subtype C that segregated with other C viruses from southern Africa. The same pattern was observed for both black and Indian subgroups and for retrospective specimens collected prior to 1990, indicating that multiple sublineages of HIV-1 C have been present in KwaZulu-Natal since the early stages of the epidemic. With the exception of three nonnucleoside reverse transcriptase inhibitor mutations, no primary resistance mutations were identified. Numerous accessory polymorphisms were present in the protease, but none were located at drug-binding or active sites of the enzyme. One frequent polymorphism, I93L, was located near the protease/reverse transcriptase cleavage site. In the envelope, disruption of the glycosylation motif at the beginning of V3 was associated with the presence of an extra protein kinase C phosphorylation site at codon 11. Many polymorphisms were embedded within cytotoxic T lymphocyte or overlapping cytotoxic T-lymphocyte/T-helper epitopes, as defined for subtype B. This work forms a baseline for future studies aimed at understanding the impact of genetic diversity on vaccine efficacy and on natural susceptibility to antiretroviral drugs.

One of the most dramatic changes in the global AIDS pandemic has been the rapid emergence and devastating spread of human immunodeficiency virus type 1 (HIV-1) subtype C (8, 34, 55, 60, 69). As a result of this rapid escalation, HIV-1 C viruses now account for more than 56% of all global infections (13). First identified in retrospective specimens from Ethiopia and South Africa (25, 57, 77), subtype C began a devastating spread across southern Africa in the late 1980s (8). Major outbreaks have now occurred in every country of southern Africa, with some regions reporting adult prevalence rates as high as 40% (10, 55, 70).

Recent studies suggest that subtype C is spreading northward into the Congo, Tanzania, Burundi, and Kenya, where it is becoming increasingly predominant relative to other subtypes (24, 28, 54). C viruses also dominate the rapidly expanding epidemic in India (59) and are increasing in frequency in China (15, 54, 76) and Brazil (4, 64). C/D recombinants have been identified in several countries, including Tanzania, Kenya, and India (18, 33, 52), and C/B recombinants have been detected in China (73).

The reasons for the increase in HIV-1 C are not known but may be related to host, viral, or socioeconomic factors. At the viral level, it has been suggested that an extra NF-κB binding site in the long terminal repeat may enhance gene expression, altering the transmissibility and pathogenesis of C viruses (66). Others have suggested that C viruses may be more stable and that their protease genes may have increased catalytic activity relative to other subtypes (72). Additional features of subtype C include a five-amino-acid insertion in the transmembrane domain of Vpu (42), a prematurely truncated second exon of *rev* (15, 54, 78), and an increase in amino acid variation at protease cleavage sites (T. de Oliveira et al., submitted for publication).

Recent advances in sequencing and bioinformatics (9, 48, 49, 74) make it easier to analyze full-length HIV-1 sequences and correlate the genetic information with the immunological and biological properties of the virus. These advances, combined with the development of promising vaccine candidates and simplified, more affordable drug regimens, are paving the way for enhanced prevention and treatment efforts in southern Africa. As with HIV-1 B, it is expected that safe and efficacious treatment of C infections will not only reduce the morbidity

* Corresponding author. Mailing address: HIV-1 Molecular Virology and Bioinformatics Laboratories, Africa Centre for Health and Population Studies, Nelson R. Mandela School of Medicine, University of Natal, Congella 4013, Durban, South Africa. Phone: 27 31 260 4013. Fax: 27 31 260 4015. E-mail: sharon.cassol@mrc.ac.za.

and premature death associated with HIV-1 and AIDS (16, 22, 27, 46) but will also play a role in reducing transmission (23).

Since we are on the brink of implementing intervention strategies in a region of the world where subtype C infections predominate, it is urgent that we collect information that will help define the phylogenetic relationships, transmissibility, and drug responsiveness of C viruses. In this study, we analyzed the C2V5 and *pol* subgenomic regions of 72 contemporary viruses from KwaZulu-Natal and compared the results with those for 18 retrospective C isolates from South Africa.

MATERIALS AND METHODS

Specimen collection and processing. A total of 72 treatment-naive HIV-1-infected children ($n = 16$) and adults ($n = 56$) representing different ethnicities, genders, age groups, and stages of disease were selected for study. Samples were obtained in Durban and surrounding areas, including Ulundi and the Hlabisa region of northern Kwazulu-Natal and Tongaat and Phoenix in the coastal region north of Durban. Participants were recruited from among symptomatic and asymptomatic adult patients, tuberculosis patients, women and children attending district health clinics, and children being treated for pneumonia. After obtaining informed consent, blood samples were collected in EDTA anticoagulant tubes (most adult patients) or as dried blood spots (most pediatric patients). Plasma was isolated within 6 h of collection; dried blood spots were stored with desiccant at −20°C until analyzed.

Viral load and CD4$^+$ cell counts. RNA was extracted from plasma and dried blood spots with a guanidinium-silica method (Nuclisens isolation kit; Organon Teknika) and an automated extractor (Organon-Teknika). Virus levels were measured with the Nuclisens HIV-1 QT kit, an assay with a quantitative range of 40 to >500,000 copies of HIV-1 RNA/ml of plasma. When applied to 50 μl of dried blood, the lower limit of detection is 1,600 HIV-1 RNA copies/ml of blood. Specificity of the method has been previously assessed and shown to be greater than 98.9% (6). CD4$^+$ cell counts in venous blood were determined according to a standard FACSCount method.

Sequencing of the envelope C2V5 region. Sequencing of *env* was performed directly on a 621-bp PCR product generated from the C2V5 region (nucleotides 7026 to 7646, relative to HXB2) (31). RNA was extracted from plasma with the ViroSeq method (Applied Biosystems). Plasma RNA and Nuclisens-extracted dried blood spot RNA were reverse transcribed to cDNA with Superscript II and random hexamer primers (Invitrogen Corp., San Diego, Calif.). The RNA template and random primers (100 ng) were heated to 70°C for 10 min, chilled on ice, and reverse transcribed at room temperature in a 20-μl reaction volume containing 1× reaction buffer, 10 mM dithiothreitol, 0.5 mM each deoxynucleoside triphosphate, and 200 U of Superscript reverse transcriptase (Invitrogen) at 42°C for 50 min, followed by 15 min at 70°C.

The C2V5 *env* region was amplified from the cDNA with MK605 (5′-AATG TCAGCACAGTACAATGTACAC-3′; positions 6945 to 6969) and CD4R2 (5′-TATAATTCACTTGTCCAATTGTCC-3′; positions 7652 to 7675) as outer primers (5) and (M13F)-ES7 (5′-tgtaaaacgacggccagtCTGTTAAATGGCAGTC TAGC-3′; positions 7002 to 7021) and (M13R)-ES8 (5′-caggaaacagctatgaccCA CTTCTCCAATTGTCCCTCA-3′; positions 7648 to 7668) as inner primers. The first and second PCR steps were carried out in final volumes of 25 μl and 50 μl, respectively, containing 1× PCR buffer, 2.0 mM $MgCl_2$, 0.2 mM each deoxynucleoside triphosphate, 2.5 pmol of each primer, and 1.25 U of Amplitaq Gold. The PCR conditions were 95°C for 13 min, followed by six cycles at 95°C for 30 s, 65°C for 45 s, and 72°C for 60 s, with a decrease of 1°C per cycle. This was followed by 29 cycles at 95°C for 30 s, 60°C for 45 s, and 72°C for 60 s, with an increase of 5 s for each extension cycle, and a final extension of 72°C for 10 min. Amplified DNA was visually quantified by agarose gel electrophoresis, purified on a Microcon (Amicon) spin column, and sequenced on an automated 3100 genetic analyzer (Applied Biosystems Inc., Foster City, Calif.) with M13 sequencing primers and a Big-Dye terminator cycle sequencing kit.

Sequencing of reverse transcriptase and protease. Sequencing of *pol* (nucleotides 2253 to 3485, relative to HXB2) (31) was performed with the ViroSeq HIV-1 genotyping system (Applied Biosystems). Plasma and dried blood spot RNAs were reverse transcribed with Moloney murine leukemia virus reverse transcriptase. A 1.8-kb fragment containing the protease (amino acids 1 to 99) and reverse transcriptase (amino acids 1 to 312) regions was then amplified in a 40-cycle PCR with Amplitaq Gold DNA polymerase and AmpErase dUTP/uracil-*N*-glycosidase to minimize the risk of cross-contamination. PCR products were visually quantified by agarose gel electrophoresis. Following purification, the products were sequenced with six of the seven kit primers (primer D was not used) and Big-Dye terminator reagents and run on a 3100 genetic analyzer as described above. Sequences were assembled, translated, and analyzed for the presence of amino acid polymorphisms. A report was generated for each sequence, with mixtures of wild-type and mutant bases being classified as mutant.

Genetic subtyping and phylogenetic analysis. To rule out contamination between samples, each new sequence was compared to other sequences amplified at the same time, as well as to other sequences previously amplified in our laboratory and published sequences in the Los Alamos BLAST search database (2). The sequences were aligned with CLUSTAL W (67) and manually edited with the codon alignment of the Genetic Data Environment (GDE 2.2) program (63). New sequences were then compared to subtype reference strains in the Los Alamos subtype database (http://hiv-web.lanl.gov/content/hiv-db/SUBTYPE_REF/align.html). Following degapping with the degapped option in PAUP*, phylogenetic trees were generated on a Linux computer with the F84 model of substitution and the neighbor-joining method (version 4.0b2a) of PAUP* (65). Trees were rooted with a homologous region of HIV-1 group O (O-CM_MP5180).

To examine intrasubtype relationships, each KwaZulu-Natal sequence was analyzed against a subset of published C sequences from Zimbabawe, South Africa, Brazil, Tanzania, Zambia, Ethiopia, Israel, and eastern India. Appropriate evolutionary models were selected with the Akaike identification system (1), implemented in MODELTEST 3.0 (48). With this method, a pairwise distance matrix was calculated and used to construct neighbor-joining maximum likelihood trees. Parameters of the reverse transcriptase/protease model, TVM + I + G, were: $f_A = 0.3986$, $f_C = 0.1653$, $f_G = 0.2033$, and $f_T = 0.2328$; R matrix values, $R_{A\to C} = 2.7534$, $R_{A\to G} = 10.1383$, $R_{A\to T} = 0.9138$, $R_{C\to G} = 1.3684$, $R_{C\to T} = 13.5383$, and $R_{G\to T} = 1.0000$; proportion of invariable sites = 0.4263; and heterogeneous variable site distribution (gamma) with alpha shape = 0.8233. Parameters of the *env* model, GTR + I + G, were: $f_A = 0.3801$, $f_C = 0.1838$, $f_G = 0.2890$, $f_T = 0.1472$; R matrix values, $R_{A\to C} = 3.3002$, $R_{A\to G} = 8.3576$, $R_{A\to T} = 3.7717$, $R_{C\to G} = 1.9646$, $R_{C\to T} = 23.3707$, $R_{G\to T} = 1.0000$; proportion of invariable sites = 0.1534; and heterogeneous variable site distribution (γ) with alpha shape (α) = 0.7332. Trees were viewed with Treetool and Treeview.

Genetic diversity and intersubtype recombination analysis. Mean genetic distances were measured with the Kimura-2 parameter model implemented in MEGA (35). To investigate whether the sequences were recombinant forms of subtype C, recombination analyses were performed with the recombination identification program (62), Bootscanning (56), recombination detection program (53), and Simplot (39), a method that uses a sliding-window approach to calculate bootstrap plots for constructing neighbor-joining trees with the DNADIST, NEIGHBOR, or CONSENSE programs of the PHYLIP package (14).

Nucleotide and amino acid sequence analysis. Nucleic acid sequences were also analyzed with SNAP (http://hiv-web.lanl.gov) (32) and Codeml, a program from the PAML software package (51). Various software programs were then used to calculate the ratio of synonymous to nonsynonymous amino acid substitutions as a measure of natural selection pressure at the protein level. Programs included SNAP and MEGA (35), which calculate a synonymous-to-nonsynonymous (d_s/d_n) substitution ratio, and Codeml, which calculates a w (d_n/d_s) value. High rates of synonymous mutation are indicative of conservation and a strict requirement for biological function, while high rates of nonsynonymous substitution are indicative of adaptive change, presumably in response to host selection pressure.

To identify amino acid patterns that are characteristic of KwaZulu-Natal viruses, nucleotide sequences were translated and aligned and the consensus was analyzed by viral epidemiology signature pattern analysis (32). Consensus sequences were screened for the presence of biologically important sites with Prosite, a database of protein families and domains.

Identification of resistance mutations and correlation with phenotype. The Stanford HIV-SEQ and β-test programs were used to identify and assess the impact of resistance-associated mutations and polymorphisms on phenotypic resistance. Each reverse transcriptase and protease sequence was compared to that of a subtype B reference strain, HXB2, in the Stanford HIV reverse transcriptase and protease sequence database (http://hivdb.Stanford.Edu/hiv/). Mutations associated with reduced sensitivity to antiretroviral drugs were assigned a drug penalty score based on genotypic-phenotypic correlative data.

Nucleotide sequence accession numbers. GenBank accession numbers for sequences obtained in this study including information on the year of specimen collection and risk category are provided in Table 1.

TABLE 1. GenBank accession numbers and year of sampling[a]

pol sequence name	GenBank accession no.	Yr of sampling	Transmission[b]	*env* sequence name[c]	GenBank accession no.
ZA004p01	AY136957	2001	P	NA	
ZA005p01	AY136958	2001	A	ZA005e01	AY137011
ZA006p01	AY136959	2001	P	NA	
ZA007p01	AY136960	2001	A	ZA007e01	AY137012
ZA008p01	AY136961	2001	A	NA	
ZA009p01	AY136962	2001	A	ZA009e01	AY137013
ZA010p01	AY136963	2001	A	ZA010e01	AY137014
ZA011p01	AY136964	2001	A	ZA011e01	AY137015
ZA012p01	AY136965	2001	A	NA	
ZA013p01	AY136966	2001	A	ZA013e01	AY137016
ZA014p01	AY136967	2001	A	ZA014e01	AY137017
ZA015p01	AY136968	2001	A	NA	
ZA016p01	AY136969	2001	A	ZA016e01	AY137018
ZA017p01	AY136970	2001	A	ZA017e01	AY137019
ZA018p01	AY136971	2001	A	ZA018e01	AY137020
ZA019p01	AY136972	2001	A	ZA019e01	AY137021
ZA020p01	AY136973	2001	A	ZA020e01	AY137022
ZA021p01	AY136974	2001	A	ZA021e01	AY137023
ZA022p01	AY136975	2001	A	ZA022e01	AY137024
ZA025p01	AY136978	2001	A	NA	
ZA024p01	AY136977	2001	A	ZA024e01	AY1370026
ZA023p01	AY136976	2001	A	ZA023e01	AY137025
ZA026p01	AY136979	2001	P	NA	
ZA027p01	AY136980	2001	P	ZA027e01	AY137027
ZA028p01	AY136981	2001	A	ZA028e01	AY137028
ZA029p01	AY136982	2001	A	NA	
ZA030p01	AY136983	2001	A	ZA030e01	AY137029
ZA031p01	AY136984	2001	A	ZA031e01	AY137030
ZA032p01	AY136985	2001	A	ZA032e01	AY137031
ZA033p01	AY136986	2001	P	ZA033e01	AY137032
ZA034p01	AY136987	2001	P	ZA034e01	AY137033
ZA035p01	AY136988	2001	P	ZA035e01	AY196496
ZA036p01	AY136989	2001	P	ZA036e01	AY137034
ZA037p01	AY136990	2001	P	ZA037e01	AY137035
ZA038p01	AY136991	2001	P	ZA038e01	AY137036
ZA039p01	AY136992	2001	P	ZA039e01	AY137037
ZA040p01	AY136993	2001	P	ZA040e01	AY137038
ZA041p01	AY136994	2001	P	NA	
ZA042p01	AY136995	2001	P	ZA042e01	AY137039
ZA043p01	AY136996	2001	P	ZA043e01	AY137040
ZA044p01	AY136997	2001	P	ZA044e01	AY137041
ZA045p01	AY136998	2001	A	ZA045e01	AY137042
ZA046p01	AY136999	2001	A	NA	
ZA047p01	AY196498	2001	A	ZA047e01	AY137043
ZA048p01	AY196499	2001	A	ZA048e01	AY137044
ZA049p01	AY196500	2001	A	ZA049e01	AY137045
ZA050p01	AY196501	2001	A	ZA050e01	AY137046
ZA051p01	AY196502	2001	A	ZA051e01	AY137047
ZA052p01	AY196503	2001	A	ZA052e01	AY137048
ZA053p01	AY196504	2001	A	ZA053e01	AY137049
ZA054p01	AY196505	2001	A	ZA054e01	AY137050
ZA055p01	AY196506	2001	A	ZA055e01	AY137051
ZA057p01	AY196507	2001	A	ZA057e01	AY137053
ZA058p01	AY196508	2001	A	ZA058e01	AY137054
ZA059p01	AY196509	2001	A	ZA059e01	AY137055
ZA060p01	AY196510	2001	A	ZA060e01	AY137056
ZA061p01	AY196511	2001	A	ZA061e01	AY137057
ZA062p01	AY196512	2001	A	ZA062e01	AY137058
ZA063p01	AY137008	2001	A	ZA063e01	AY137059
ZA064p01	AY137006	2001	A	ZA064e01	AY137060
ZA065p01	AY137007	2001	A	ZA065e01	AY137061
ZA066p01	AY137004	2001	A	ZA066e01	AY137062
ZA068p01	AY196513	2001	A	ZA068e01	AY137064
ZA069p01	AY196514	2001	A	ZA069e01	AY137065
ZA071p02	AY137000	2001	A	ZA071e01	AY137067
ZA073p01	AY196515	2001	A	ZA073e01	AY137070
ZA074p01	AY196516	2001	A	ZA074e01	AY137071
Za075p01	AY196517	2001	A	ZA075e01	AY137069
ZA077p02	AY137001	2001	A	ZA077e01	AY137073
ZA078p02	AY137003	2001	A	ZA078GRe02	AY137072
ZA079p02	AY137002	2002	A	ZA079GRe02	AY196497
ZA080p01	AY137005	2001	A	NA	

[a] *pol* and *env* sequences from the same virus and individual are shown on the same line. All samples were collected in Durban, South Africa, and surrounding regions.
[b] A, adult (heterosexual) transmission; P, pediatric (perinatal) transmission.
[c] NA, not available.

TABLE 2. Characteristics of and laboratory results for children and adults in the study

Variable	All patients (n = 72)	Adults (n = 56)	Children (n = 16)
Mean age ± SD		38.4 yr	15.72 mo
Sex, no. of subjects (%)			
Male		19/56 (34)	NA[a]
Female		37/56 (66)	NA
Ethnicity, no. of subjects/total (%)			
Black	46/72 (64)	30/56 (54)	16/16 (100)
White	2/72 (3)	2/56 (4)	0/16 (0)
Colored	2/72 (3)	2/56 (4)	0/16 (0)
Indian	22/72 (31)	22/56 (39)	0/16 (0)
CD4 cell count, no. of subjects/total (%)			
≤200 cells/mm^3		12/40 (30)	NA
201 to 500 cells/mm^3		17/40 (43)	NA
≥ 501 cells/mm^3		11/40 (28)	NA
Avg CD4 cell count, cells/mm^3		366	NA
Plasma HIV RNA, no. of subjects/total (%)			
≤400 copies/ml	0/50 (0)	0/36 (0)	0/14 (0)
401 to ≤10^4 copies/ml	7/50 (14)	7/36 (19)	0/14 (0)
>10^4 to ≤10^5 copies/ml	16/50 (32)	14/36 (39)	2/14 (14)
>10^5 copies/ml	27/50 (54)	15/36 (42)	12/14 (86)
Avg plasma HIV RNA (copies/ml)	248,260	113,020	383,500
HIV-1 subtype, no. of subjects/total (%)			
Subtype C	60/61 (98)	48/49 (98)	12/12 (100)
D/C recombinant	1/61 (2)	1/49 (2)	0/12 (0)

[a] NA, not available.

RESULTS

Study population. Demographic and laboratory results for the 72 KwaZulu-Natal patients are summarized in Table 2. Six specimens came from three sets of epidemiologically linked sex partners. After recording the genotype, the male partner of each couple was excluded from further analysis. Many adult patients had HIV-1-related symptoms; 10 had tuberculosis. Two children were asymptomatic; the remaining 14 children had a variety of symptoms, ranging from pneumonia to weight loss, hepatomegaly, splenomegaly, and diarrhea. All of the children were black. To investigate genetic change over time, 18 retrospective samples (8 from KwaZulu-Natal and 10 from Cape Town) were sequenced and included in the analysis.

Genetic divergence, subtyping, and phylogenetic tree analysis. As shown in Table 3, the average intersequence divergence among KwaZulu-Natal sequences was significantly higher than among subtype C sequences from Brazil and India, but comparable to values observed for Botswana and other countries in southern Africa. There was no measurable difference in diversity between Indian and black or between adult and pediatric subgroups. Eleven *env* samples carried insertions and deletions and could not be sequenced directly from the PCR product. Overall, KwaZulu-Natal *env* sequences differed from the reference sequences of subtypes A, B, and D by 30.4%, 29.3%, and 32.2%, respectively. Pol sequences differed from subtype A, B, and D reference strains by 11.6%, 11.08%, and 11.01%, respectively.

Maximum-likelihood and neighbor-joining distance methods were used to determine subtype. As expected, 60 of 61 (98.4%) matched *env-pol* sequence pairs and all of the retrospective sequences grouped as subtype C. These phylogenetic relationships were supported by bootstrap values of >95%. One sample, ZA021p01, had different *env* and *pol* subtypes, suggesting recombination between these two regions. Further analysis by recombination identification program, recombination detection program, Bootscanning, and Simplot confirmed that ZA021p01 was an intersubtype recombinant that typed as subtype C in *env* and as a C/D recombinant in *pol*, with the

TABLE 3. DNA distances between subtype C sequences from different population groups

Country and ethnic group (no. of viruses)	Mean % distance (SE)			
	pol	*pol* 1st and 2nd positions	*env*	*env* 1st and 2nd positions
South Africa (73)	4.93 (0.27)	2.39 (0.24)	19.18 (1.0)	19.3 (1.1)
Indian (from KZN) (19)	5.07 (0.34)		19.99 (1.1)	
Black (from KZN) (44)	4.9 (0.3)		19.38 (1.1)	
Botswana (51)	5.92 (0.30)	2.86 (0.24)	19.25 (0.99)	18.7 (1.0)
India (9)	3.44 (0.29)	2.22 (2.28)	11.78 (0.84)	11.7 (0.9)
Tanzania (4)	4.86 (0.47)	2.08 (0.43)	17.81 (1.5)	17.8 (1.6)
Zambia (2)	5.16 (0.6)	2.23 (0.5)	20.3 (2.0)	20.9 (2.4)
Brazil (2)	2.65 (0.45)	1.85 (0.47)	12.65 (1.4)	12.1 (1.6)

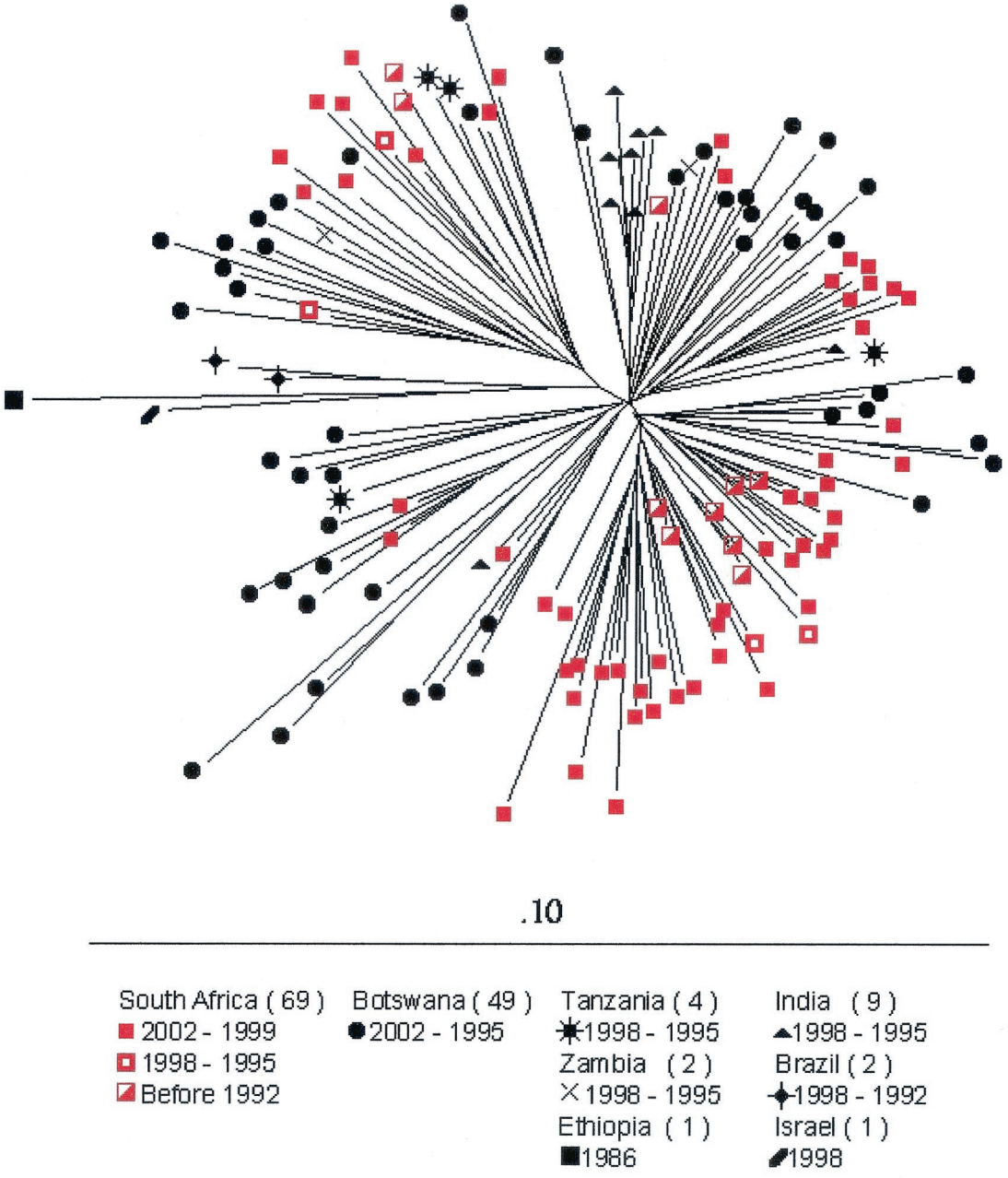


FIG. 1. Representative *pol* tree showing the relationships between retrospective and contemporary sequences from South Africa, Botswana, and other countries affected by the subtype C epidemic. The sequences are coded by the country of origin and year of isolation. The following sequences were included in the analysis: 49 previously described isolates from Botswana (accession numbers AF110960, AF110963, AF110967, AF110970, AF110972, AF110973, AF110978, and AF443074 to AF443115), 9 sequences from India (accession numbers AF286232, AF286223, AF286231, AB023804, AF067159, AF067155, AF067154, AF067157, and AF067158), 4 sequences from Tanzania (accession numbers AF286234, AF286235, AF361874, and AF361875), 2 sequences from Zambia (AF286224 and AF286225), 2 sequences from Brazil (U52853 and AF2862228), 1 sequence from Ethiopia (U46016), 1 sequence from Israel (AF286233), and 69 sequences from South Africa, including 5 previously described sequences (AF286227, AY043173, AY043174, AY043175, and AY043176), 3 sequences from another study (71), and 61 sequences newly generated from this study (14 retrospective and 47 contemporary strains).

breakpoint for recombination occurring approximately two-thirds of the way along the reverse transcriptase gene.

To investigate within-subtype clustering, trees were constructed with published C sequences from eight different countries (Fig. 1 and 2). Full-length reference sequences were selected because these strains contained both the *env* and *pol* genes. Unlike sequences from India, where seven out of nine (77.7%) samples grouped as a single monophyletic group, KwaZulu-Natal sequences were widely dispersed across multiple clusters, or sublineages. The topology of samples within

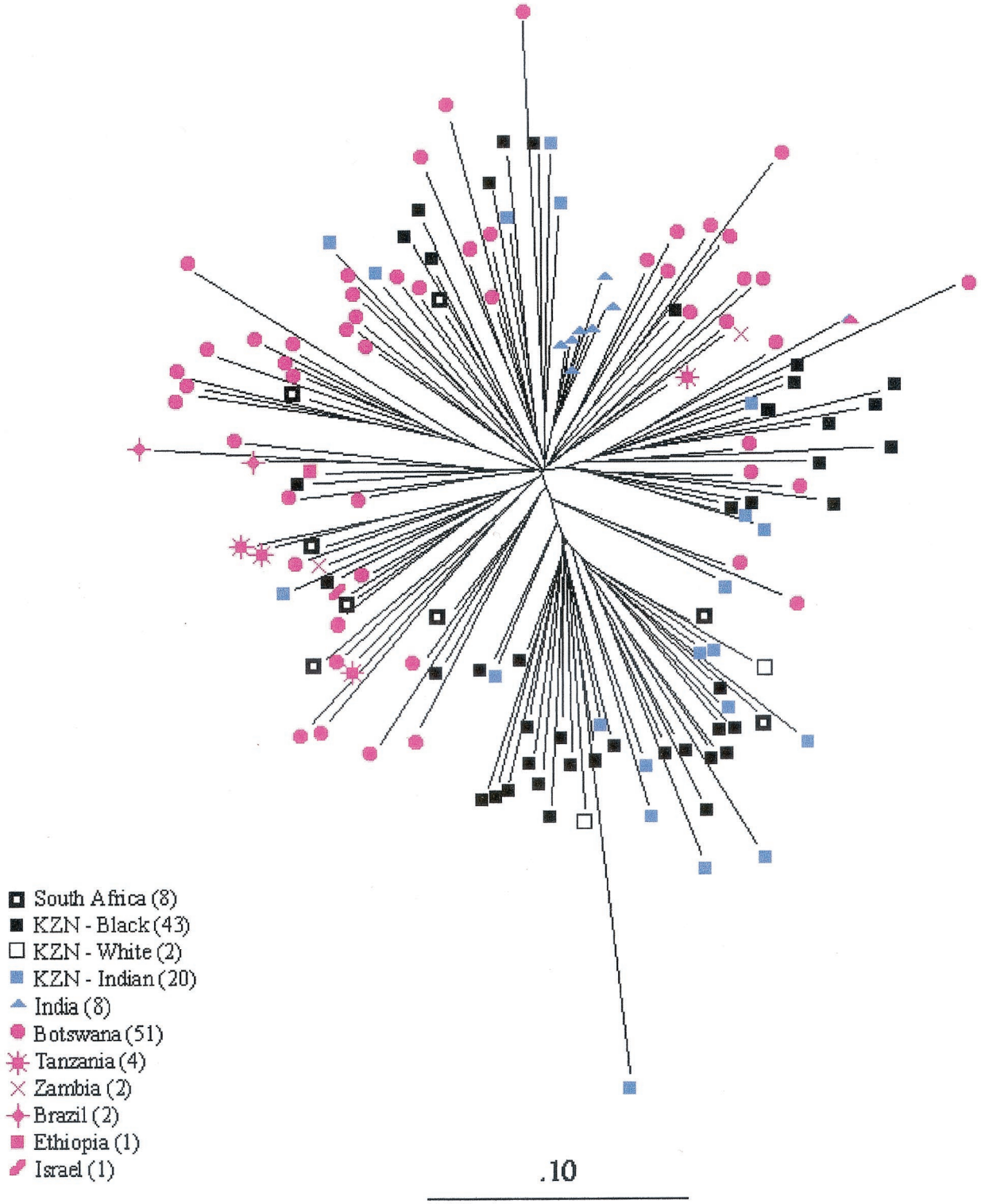


FIG. 2. Phylogenetic relationship of C2V5 envelope sequences from KwaZulu-Natal, Botswana, Zambia, and Tanzania. Non-KwaZulu-Natal strains are the same as those described in Fig. 1.

these maximum-likelihood and neighbor-joining trees was similar for both *env* and *pol* and for retrospective specimens collected prior to 1992.

As shown in Fig. 1, many of the retrospective sequences were localized internally, closer to the root of the sublineage. For most trees, the bootstrap support for the delineated sublineages was higher than 70%. Overall, 58 of the 69 (84.0%) KwaZulu-Natal sequences grouped within sublineages. The

number of samples within each sublineage ranged from 3 to 14 for *env* and from 4 to 12 for *pol*. One of the largest sublineages consisted of only KwaZulu-Natal sequences. The remaining sublineages contained sequences from other countries in southern Africa (primarily Botswana, but also Tanzania and Zambia), in addition to those from KwaZulu-Natal. One KwaZulu-Natal sample clustered with sequences from Brazil, Ethiopia, Zambia, and Israel. Sequences from different patient groups were distributed across the phylogenetic tree and showed no obvious evidence of geographic or subgroup clustering based on ethnicity, age, or sex (Fig. 2).

KwaZulu-Natal and subtype-specific signature motifs. The KwaZulu-Natal protease consensus sequence was identical to the consensus sequence of subtype C at 100% of 99 amino acids, but differed from the consensus of subtypes A, B, and D at seven, eight, and six positions, respectively (Fig. 3). Compared to the B consensus, amino acid substitutions were identified at 32 different positions. The mean number of substitutions was nine, with 65 (94.2%) isolates having eight or more substitutions relative to subtype B.

The reverse transcriptase consensus differed from the subtype C consensus at only one position, codon V60I. Forty-three (62.3%) of the KwaZulu-Natal sequences had an isoleucine at reverse transcriptase codon 60 rather than the valine residue that is typical of B and C subtypes. This polymorphism was not present in the consensus sequence of retrospective samples. Comparison of the KwaZulu-Natal consensus to that of subtype B revealed 19 different amino acid substitutions. Thirty-six (52.2%) had 21 to 25 substitutions and another 33 (47.8%) had 14 to 20 substitutions relative to HIV-1 B. The most frequent substitutions are shown in Table 4.

The characteristic GPGQ motif at the tip of the V3 loop was conserved in 98.4% of KwaZulu-Natal samples (Fig. 4). The RIGPGQTFYATG dodecapeptide (amino acids 13 to 24 of V3), previously identified in 69.1% of V3 sequences from Calcutta, India (C_{IN}), was detected in only 28.6% of KwaZulu-Natal specimens. One of the most variable amino acids was the C-terminal glycine (G). Overall, 34.9% of KwaZulu-Natal sequences had an asparagine substitution at this position; 19% had a deletion mutation. Most (87.7%) of the substitutions and deletions were present within the black subgroup. The deletion mutation, present in 24.2% of black and 5.3% of Indian sequences, caused a decrease in length of the V3 loop from 35 to 34 amino acids. With the exception of a single D25K mutation, no basic amino acid substitutions were detected at V3 loop positions 11, 24, or 25.

Amino acid substitutions associated with drug resistance. The amino acid sequence of each KwaZulu-Natal sequence was compared to sequences in the Stanford University HIV reverse transcriptase and protease sequence database in order to identify polymorphisms and mutations previously associated with drug resistance in HIV-1 B infections. No primary resistance mutations to protease inhibitors were detected in any of the KwaZulu-Natal samples. However, a substantial number of accessory (secondary) mutations were found at the following positions, in order of decreasing frequency: I93L (97.1%), M36I (85.5%), M63P/S/I/V/H (37.7%), K20R (13.0%), V77I (7.2%), and L10I (1.4%). Similarly, no primary or accessory mutations to resistance against nucleoside reverse transcriptase inhibitors were identified. However, three patients were found to harbor resistance mutations to nonnucleoside reverse transcriptase inhibitors: patient ZA024p01 had a K103N mutation, and her male partner, ZA023p01, carried a G190A mutation in addition to K103N. A third patient, ZA010p01, had a single A98G mutation. Table 5 summarizes the frequency and pattern of these mutations.

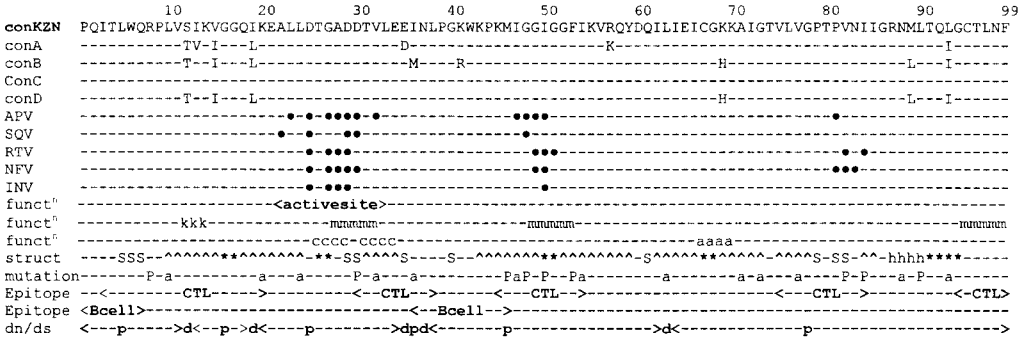
Amino acid substitution and selection pressure. KwaZulu-Natal sequences were then compared internally to assess the mutational behavior of reverse transcriptase and protease in the absence of drug therapy. Analysis by the likelihood ratio method of Yang (74) indicated that both genes were under strong purifying (negative) selection pressure (d_n/d_s or $w < 1$), with >95% of sites having $w1 = 0.019$ and $w2 = 0.395$. In contrast, only 5 (5.1%) amino acids in protease (codons 12, 19, 35, 37, and 63) and 15 (4.8%) amino acids in reverse transcriptase (codons 36, 39, 123, 135, 162, 166, 174, 196, 207, 211, 214, 245, 272, 277, and 286) were found to be under strong positive Darwinian (d) selection ($w = 2.055$). As shown in Fig. 3, these amino acids were not randomly distributed but were located at discrete loci along the reverse transcriptase and protease genes. Seven (35%) amino acids that were under positive (diversifying) selection pressure (protease positions 12S and 19I; reverse transcriptase positions 36A, 39E, 123G, 211K, and 245Q) were present in both the KwaZulu-Natal and subtype C consensus sequence but not in the consensus sequences of subtypes A, B, and D, suggesting that these signature residues may offer a subtype-specific fitness advantage to C viruses.

Impact of substitution on functional motifs. Naturally occurring polymorphisms also resulted in significant variation in the number and type of phosphorylation sites. Overall, 17

TABLE 4. Frequency of the most common amino acid substitutions in the *pol* gene

Protein	Amino acids	% of strains
Protease	H69K/Q/Y	100
	I93L	97.1
	L19I/V/E/T/A	97.1
	I15V	92.8
	M36I/L/T	91.3
	R41K	89.8
	L89M	89.8
	T12S/A/P	81.2
Reverse transcriptase	V35T/I/K/M/Q	100
	Q207E/D/G/N/S/R/K	100
	V245Q/K/L/H	98.6
	T39E/D/A/K/N	98.5
	I293V	97.1
	V292I	97.1
	T200A/I/E	95.7
	E291D	95.7
	K173A/T/V/G/I	94.2
	S48T/E	92.7
	K122E/Q	93.2
	D177E/G/N	89.9
	A272P/Q/S/R	89.9
	T286A/V	81.1
	E36A/T/V	78.3
	D123G/N/S	73.9
	K277R/S	66.7
	V60I	62.3
	R211K	56.5

Protease



Reverse Transcriptase

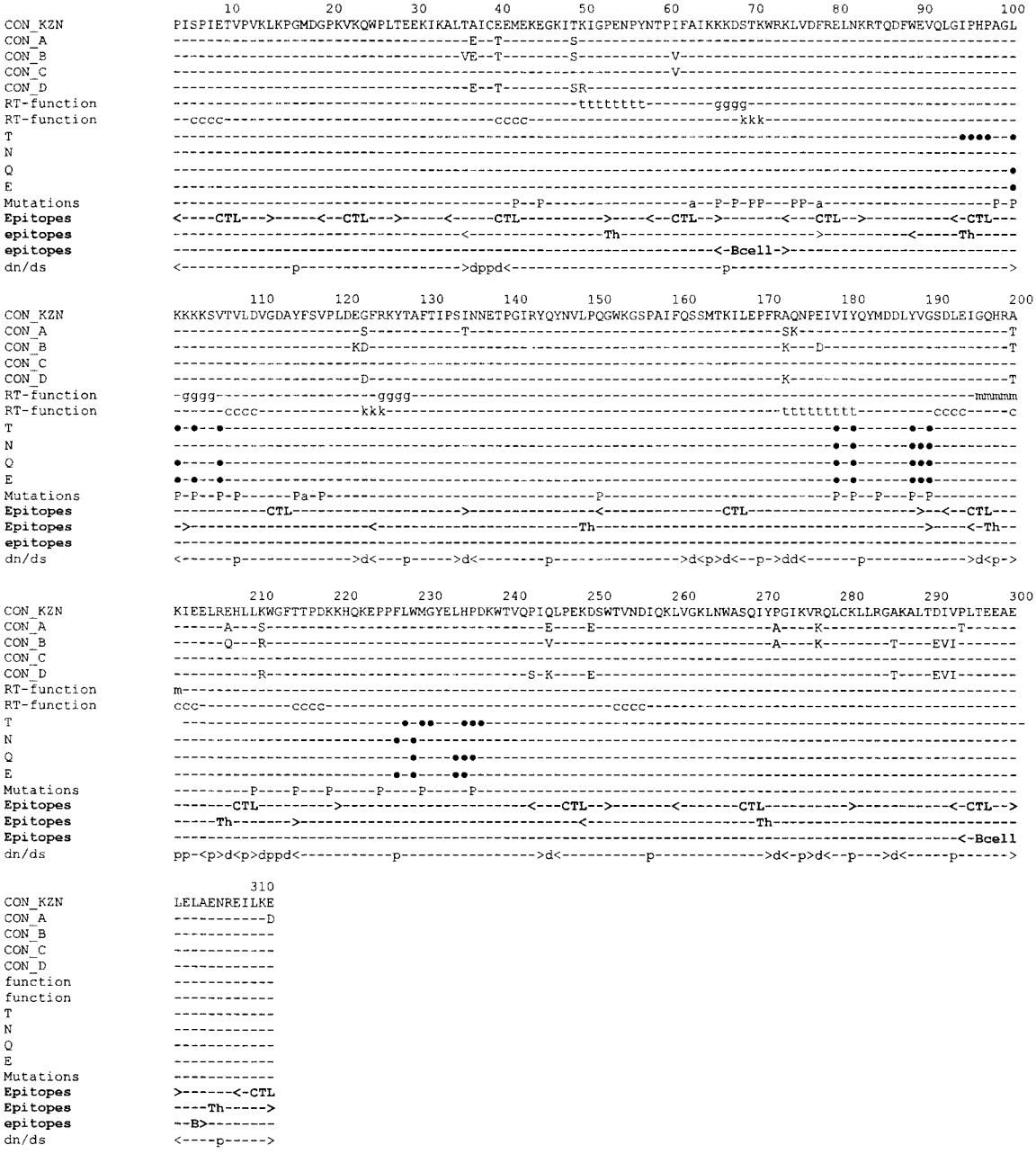


TABLE 5. Amino acid substitutions at codons associated with drug resistance[a]

Patient no.	Protease substitution(s)	Reverse transcriptase substitution(s)
21	M36I	
36, 29, 62	I93L	
7, 8, 11, 13, 14, 15, 16, 19, 27, 30, 31, 32, 33, 34, 37, 39, 40, 41, 42, 47, 51, 53, 54, 55, 57, 59, 61, 73, 78, 79, 80	M36I, I93L	
20	V771, I93L	
10	M36I, I93L	A98G
35, 66, 74	L63P, I93L	
49	K20R, V77I, I93L	
5, 9, 22, 38, 58, 68, 69	K20R, M36I, I93L	
4, 12, 17, 18, 25, 26, 28, 45, 46, 48, 50, 52, 63, 64, 65, 71, 75, 77	M36I, L63P/T/S/H, I93L	
6	M36I, V77I, I93L	
23	L63T, V771, I93L	K103N*, G190A
24	L63T, V771, I93L	K103N*
44	L10I, M36I, L63P, I93L	
60	M36I, L63P, V77I, I93L	
43	K20R, M36I, L63V, I93L	

[a] Amino acid substitutions relative to the North American/European HIV-1 B subtype. *, primary mutation associated with resistance to nevirapine, delavirdine, and efavirenz.

potential phosphorylation sites were identified in the *pol* gene, 3 in the protease and 14 in the reverse transcriptase. Twelve of the *pol* sites were conserved among KwaZulu-Natal patients and in the consensus sequences for subtypes A, B, C, and D (Fig. 3). These included the predicted protein kinase C site at codons 12 to 14 near the N terminus of the protease and the two casein kinase II phosphorylation motifs at the active site. Most KwaZulu-Natal sequences had an S-X-K rather than a T-X-K motif at protease codons 12 to 14. Conserved phosphorylation sites in reverse transcriptase included protein kinase C codons 68 to 70; tyrosine kinase codon 49 to 56; cyclic AMP/cyclic GMP-dependent codons 65 to 67, 102 to 105, and 125 to 128; and CKII codons 3 to 6, 107 to 110, 191 to 194, 215 to 218, and 253 to 256.

Two KwaZulu-Natal patients lacked a cyclic AMP phosphorylation site at reverse transcriptase codons 102 to 105 due to the presence of a K103N mutation. Some phosphorylation sites, such as the CKII sites at reverse transcriptase positions 39 to 41 and 200 to 203, were present in subtypes A, B, and D but absent from most of the KwaZulu-Natal and subtype C sequences. Other differences included the absence of an internal myristoylation site (41) at reverse transcriptase codons 196 to 201 in nine patients and the presence of an amidation site at protease codons 67 to 70 in subtype A, subtype C, and all but two of the KwaZulu-Natal sequences. With a single exception, all of the natural reverse transcriptase mutations were embedded within cytotoxic T-lymphocyte, T-helper, or overlapping cytotoxic T-lymphocyte/T-helper epitopes, as defined for B viruses.

Of particular interest, with respect to the *env* gene, was a cluster of substitutions located at or in close proximity to the bottom of the V3 loop, a region known to play a major role in viral tropism and coreceptor usage. This cluster included amino acid −1, immediately upstream from the cysteine residue at the beginning of V3, and amino acid positions 11 and 13 within the V3 loop itself. In common with other C viruses, strains from 89.0% of KwaZulu-Natal patients had amino acid substitutions that resulted in elimination of the N-linked glycosylation site at position −1 (amino acid 301 according to the numbering of Korber at al. [31]). In 91.0% of patients, loss of glycosylation was associated with a serine (S) substitution at position 11 and the presence of a positively charged arginine (R) residue at V3 position 13. The resultant S-X-R motif gave rise to a second, alternative protein kinase C site immediately adjacent to the phosphorylation site at amino acids 8 to 10. These findings suggest a potential linkage between deglycosylation and phosphorylation in the V3 loop of C viruses.

Most A variants also carried the extra protein kinase C site at position 11 to 13 but lacked the N-linked glycan at position −1. Instead, a more distal N-X-S glycosylation site (positions −7 to −5) was frequently absent in A viruses. Another protein kinase C site, located downstream from the C terminus of V3 at positions 45 to 47 (relative to V3), was missing in most KwaZulu-Natal viruses. This site is highly conserved among subtype B viruses. In common with subtype B, KwaZulu-Natal and other C viruses contained a highly conserved CKII site at amino acids 68 to 71.

DISCUSSION

Despite the dramatic impact of HIV-1 and AIDS on the KwaZulu-Natal region of South Africa, few studies have examined the genetic diversity and molecular phylogeny of KwaZulu-Natal viruses. To date, only eight full-length South African sequences have been published (71, 78). The primary goals of this study were to identify regions of high variability, characterize amino acids that are unique to local strains, and identify sites that are highly conserved and thus likely to be impor-

FIG. 3. Correlation of signature patterns with structure and function for protease and reverse transcriptase. conKZN, KwaZulu-Natal consensus; conA, conB, conC, and conD, consensus sequences for subtypes A, B, C, and D, respectively; APV, SQV, RTV, NFV, INV, drug binding sites for amprenavir, saquinavir, ritonavir, nelfinavir, and indinavir, respectively; funct^n, RT, reverse transcriptase; CTL, cytotoxic T-lymphocyte epitope; ●, drug-binding site; k, protein kinase C phosphorylation site; c, casein kinase phosphorylation site; m, myristoylation site; aaaa, amidation site; t, tyrosine kinase phosphorylation site; g, cyclic AMP- and cyclic GMP-dependent protein kinase site; T, thiocarboxanilide UC-781; N, nevirapine; Q, quinoxaline HBY 097; E, efivirenz; a, accessory mutation; P, primary mutation; caret, extended β-strand; S, bend; star, hydrogen-bonded turn; h, helix; p, purifying selection pressure; d, Darwinian (positive) selection pressure.

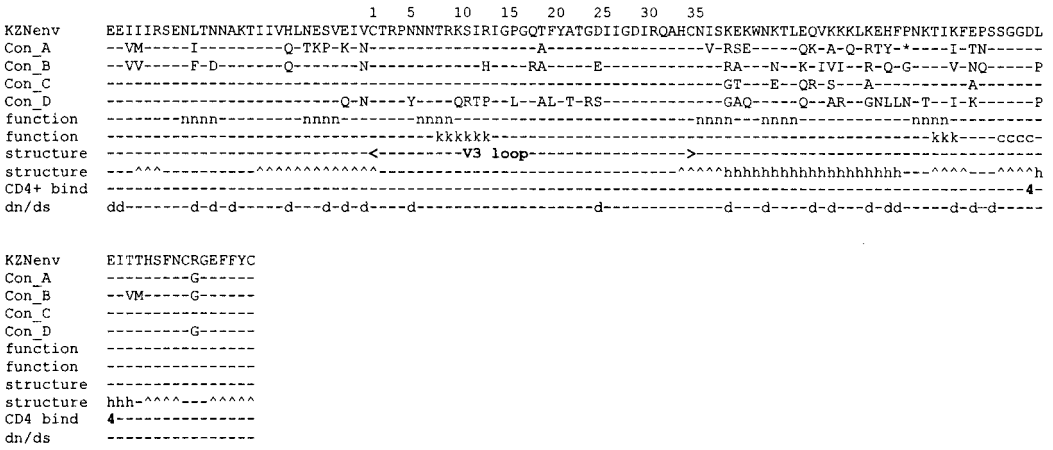


FIG. 4. Correlation of signature patterns with structure and function of V3 loop. KNZenv, KwaZulu-Natal consensus; Con_A, Con_B, Con_C, and Con_D, consensus sequences for subtypes A, B, C, and D, respectively; k, protein kinase C phosphorylation site; c, casein kinase phosphorylation site; n, N-linked glycosylation site; caret, extended β-strand; h, helix; 4, $CD4^+$ binding site; d, Darwinian (positive) selection pressure.

tant for vaccine development and the assessment of antiretroviral therapy.

Our results indicate that C viruses in KwaZulu-Natal have a higher level of nucleotide diversity than previously reported (70, 71) and that the epidemic, in its explosive phase, is characterized by multiple circulating sublineages in both the Indian and black communities. The restricted distribution of subtype C viruses from India compared to the multilineage pattern of Indian viruses from Africa indicates that the two Indian epidemics have different origins and different evolutionary histories. The presence of retrospective samples (collected prior to 1990) at internal (basal) branches in three of the sublineages suggests that each lineage is derived from a different founder variant and that these variants have been cocirculating in South Africa for at least 10 years. Of significant note was the cosegregation and close relatedness of sequences from KwaZulu-Natal black and Indian inhabitants, not only to each other, but also to published sequences from Botswana. This close relationship with sequences from Botswana was not observed in a previous study (45), presumably because of the small number of samples included from South Africa ($n = 5$). Taken together, our findings confirm the existence of multiple HIV-1 C sublineages in southern Africa and demonstrate that the spread of these different lineages has been substantial.

The finding that C viruses from KwaZulu-Natal are substantially more diverse than those in India and Brazil is consistent with other studies and has been attributed to the longer duration of the AIDS epidemic in Africa (4, 59). The overall evolutionary rate of *pol* and *env* sequences, as measured by a dated-tip likelihood method (51), was 35% and 68% higher than that of subtype B. Despite the high level of diversity, KwaZulu-Natal viruses were remarkably well conserved at the amino acid level, both within subtype C and among different individuals. This is due to the fact that a large number of the nucleotide substitutions are silent (synonymous) mutations that cause no change in the amino acid sequence. As a result, the consensus sequence for the KwaZulu-Natal protease was identical to the consensus sequence for subtype C, while the reverse transcriptase consensus sequences differed from the C consensus at a single amino acid, codon 60.

High rates of synonymous-to-nonsynonymous nucleotide change have also been observed among subtype C isolates from Zimbabwe (58) and Ethiopia (38). This inherent property of African subtype C viruses is a reflection of the differential pressure exerted on the three positions of the amino acid code. For the KwaZulu-Natal reverse transcriptase gene, the mutation rate for the third position of the codon was four times higher than that observed for the second position and 30 times higher than for the first codon position (data not shown).

The conservation of subtype C at the amino acid level offers considerable promise for the development of a consensus- or ancestor-based "supervaccine" (17, 45). Recent primate studies suggest that it may be possible to overcome diversity and achieve cross-protection against different HIV-1 variants (12, 61). However, it should be stressed that the long-term impact of silent mutations on vaccine efficacy is not known.

In the context of antiretroviral therapy, one recent study found that, despite numerous naturally occurring mutations in reverse transcriptase, C viruses from Zimbabwe were as susceptible as subtype B viruses to commonly used nucleoside and nonnucleoside reverse transcriptase inhibitors (58). However, another recent study found that, although C viruses in Ethiopia were susceptible to reverse transcriptase inhibitors, the presence of silent mutations led to a more rapid emergence of resistance (38). These data emphasize the need for carefully designed prospective trials to determine whether existing polymorphisms influence the development of resistance in C-infected patients.

With the exception of two primary resistance mutations, K103N and G190A, which occurred in a single husband-wife pair, none of the reverse transcriptase or protease polymorphisms occurred at drug-binding sites or at active sites of the enzymes. Both mutations are known to cause high-level resistance to nevirapine in persons infected with subtype B (50). Although believed to be naturally occurring, the possibility that these mutations represent treatment-induced changes cannot be excluded. As many as 15% of patients in the private sector in South Africa have received or are currently receiving some form of antiretroviral therapy. Many protocols include nevirapine because of its low cost and long half-life. Nevirapine is

also being increasingly used for the prevention of mother-to-child HIV-1 transmission in KwaZulu-Natal and other regions of Africa (23).

All of the remaining *pol* polymorphisms occurred in regions involved in the three-dimensional configuration of reverse transcriptase and protease. One such polymorphism, which occurred in a single patient, was A98G in the reverse transcriptase. This mutation was also detected in a treatment-naive patient from Ethiopia (38). In persons infected with subtype B, A98G has been associated with low-level resistance to nonnucleoside reverse transcriptase inhibitors. Other polymorphisms were localized within the hinge region of protease, a region that induces conformational changes during drug binding. A subset of these mutations, M36I/R41K/H69/L89 M, has been linked to increased catalytic activity in subtypes A and C (72). Another series of polymorphisms, at codons 12, 15, 19, and 93, occurred in >80% of KwaZulu-Natal viruses and formed a KwaZulu-Natal/subtype C signature motif. The first three amino acids of this motif are located near the N terminus of protease, in an extended β-strand; the fourth, I93L, is located in a hydrogen-bonded turn, immediately upstream of the protease/reverse transcriptase cleavage site. The marked dominance of I93L among C viruses, its close proximity to the protease/reverse transcriptase cleavage site, and its linkage to the T12S/T15V/L19I signature warrant further investigation. Studies of HIV-1 B have reported that mutations in the protease and Gag-Pol cleavage sites contribute to drug resistance, are specifically selected during therapy, and can lead to improved enzyme kinetics (7, 11).

The observed natural polymorphisms did not occur at random but were clustered in specific functional domains of the reverse transcriptase, protease, and *env* genes. Overall, >95% of KwaZulu-Natal *pol* codons were found to be under strong purifying (negative) selection pressure ($d_n/d_s < 1.0$) and thus were unlikely to undergo nonsynonymous substitution. These conserved codons were concentrated within active sites and at drug-binding sites in reverse transcriptase and protease and at nucleoside triphosphate binding sites in reverse transcriptase. The remaining 5% of amino acids were under strong positive selection pressure and were concentrated in regions associated with maintaining the tertiary structure and facilitating conformational changes. Some positively selected codons, such as protease 63 and reverse transcriptase 123 and 174, showed extensive interpatient and intersubtype variation. Other codons (such as protease 12S and reverse transcriptase 39E, 245Q, 272P, and 277R) were highly conserved among KwaZulu-Natal and subtype C sequences and formed part of an HIV-1 C signature sequence. The conservation of codons in the face of strong diversifying pressure suggests that they may play an important role in the evolutionary, structural, and phenotypic properties of C viruses. A few positively selected codons were conserved across several subtypes, suggesting that they may contribute to the evolutionary history of group M viruses.

Although many factors contribute to the generation of new variants, one of the most important is related to cytotoxic T lymphocytes and the role they play in recognizing epitopes presented by major histocompatibility complex class I molecules. With a single exception, all of the naturally occurring reverse transcriptase mutations were embedded within cytotoxic T-lymphocyte, T-helper, or overlapping cytotoxic T-lymphocyte/T-helper epitopes as previously defined for B viruses (30). Several signature sequences in *env* also mapped to known subtype B cytotoxic T-lymphocyte epitopes, including the heavily glycosylated regions at the bottom of V3 and the associated protein kinase C phosphorylation site at V3 position 11. Information on subtype C epitopes is just beginning to emerge and, when combined with novel methods of analysis, may lead to new insights into the immune selection pressures occurring during seroconversion and in response to therapy. By examining sites under positive selection pressure, we may be able to identify targets of the host immune system and select appropriate epitopes for inclusion in a subtype C vaccine.

Although it is well known that most C viruses lack a V3 glycosylation site and a basic amino acid residue at position 11, the biological significance of these findings remains unclear. Disruption of V3 glycosylation has also been reported to occur in 52%, 34%, and 20% of subtype G, A, and D viruses, respectively. Studies of subtype B have suggested that this N-linked glycan may play a role in the interaction of gp120 with its coreceptors (37) and in perinatal transmission. Nakayama et al. (43) found that absence of this V3 glycan caused a marked reduction in CXCR4-dependent but not CCR5-dependent viral entry. Others have suggested that the V3 glycan is not necessary for CXCR4 usage (40) and that its absence leads to enhanced infectivity of CXCR4-expressing cells (47).

Li et al. (37) found that multiple factors contribute to coreceptor usage and that the effects exerted by the V3 glycan are both isolated and context dependent. Similarly, the absence of a basic amino acid at position 11 of V3 and at positions 24 and 25 has been associated with a non-syncytium-inducing phenotype and CCR5 coreceptor-using properties, while the presence of basic charge has been correlated with CXCR4 and syncytium-inducing phenotypes (19, 20, 21, 26, 43). As with deglycosylation, these correlations have been imprecise.

Our findings, showing a potential linkage between V3 deglycosylation and the presence of a serine phosphorylation site at position 11, suggest that factors other than glycosylation and charge may have to be taken into account when assessing the function of V3. Based on the knowledge that C viruses are almost exclusively non-syncytium inducing and CCR5 using, it is tempting to speculate that deglycosylation may allow better access to the CCR5 coreceptor, while phosphorylation may alter the conformation of gp120, exposing retroviral sites that are needed for efficient CCR5-mediated viral entry. Although highly speculative, this possibility warrants further study given the critical importance of V3 for host cell recognition and viral entry.

Differences were also observed in the number and position of phosphorylation sites in reverse transcriptase and protease. Phosphorylation is known to modulate the activity of many proteins that interact with nucleic acids, including DNA and RNA polymerase. It is also known that, in addition to reverse transcriptase and protease, several protein kinases are incorporated into mature HIV-1 virions (68), where they are available not only to regulate the activity of reverse transcriptase and protease, but also to participate in interactions with the host cell. Phosphorylation of threonine residue at reverse transcriptase codon 215 has been shown to increase discrimination against azidothymidine, leading to drug resistance (36), and

phosphorylation of protease substrates can lead to impaired proteolytic cleavage (68). Our data indicate that several phosphorylation sites in the *pol* gene of KwaZulu-Natal and subtype C viruses are highly conserved and positively selected. It will be important to determine whether these sites play a significant role in the replicative capacity and proteolytic processing of C viruses.

ACKNOWLEDGMENT

This work was supported by Programme Grant 061238 from the Wellcome Trust, United Kingdom.

REFERENCES

1. **Akaike, H.** 1997. A new look at statistical model identification. IEEE Trans. Automatic Control **19:**716–723.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
3. **Biggar, R. J., Janes, M., Pilon, R., Roy, R., Broadhead, R., Kumwenda, N., Taha, T. E., and S. Cassol.** 2002. Human immunodeficiency virus type 1 infection in twin pairs infected at birth. J. Infect. Dis. **186:**281–285.
4. **Brindeiro, R., B. Vanderborght, E. Caride, L. Correa, R. M. Oravec, O. Berro, L. Stuyver, and A. Tanuri,** 1999. Sequence diversity of the reverse transcriptase of human immunodeficiency virus type 1 from Brazilian untreated individuals. Antimicrob. Agents Chemother. **43:**1674–1680.
5. **Cassol, S., B. G. Weniger, G. Babu, et al.** 1996. Detection of HIV-1 env subtypes A, B, C, and E in Asia with dried blood spots: a new surveillance tool for molecular epidemiology. AIDS Res. Hum. Retrovir. **12:**1435–1441.
6. **Cassol, S., M. J. Gill, R. Pilon, M. Cormier, R. Voight, B. Willoughby, and J. Forbes.** 1997. Quantification of human immunodeficiency virus type 1 RNA from dried plasma spots collected on filter paper. J. Clin. Microbiol. **35:**2795–2801.
7. **Cote, H. C. F., Z. L. Brumme, and R. Harrigan.** 2001. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by Indinavir, Ritonavir, and/or Saquinavir. J. Virol. **75:**589–594.
8. **De Oliveira, T., K. Bishop, S. Danaviah, and S. Cassol.** 2001. Changing dynamics of HIV-1 subtype diversification in Africa. 8th HIV Dynamics and Evolution Meeting, Paris, 27–29 April 2001.
9. **De Oliveira, T., R. Miller, M. Tarin, and S. Cassol.** An integrated genetic data environment (GDE)-based LINUX interface for the analysis of HIV-1 and other microbial sequences. Bioinformatics, in press.
10. **Department of Health/Directorate Health Systems Research.** 2001. Seventh national HIV survey of women attending antenatal clinics of the public health services in South Africa, October/November 2000. Directorate Health Systems Research, Department of Health, Pretoria, South Africa.
11. **Doyon, L. G., D Thibeault, F. Poulin, et al.** 1996. Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors. J. Virol. **70:**3763–3769.
12. **Dunn, D. S., B. Hurtel, C. Beyer, et al.** 1997. Protection of SIVmac-infecTed macaque monkeys against superinfection by a simian immunodeficiency virus expressing envelope glycoproteins of HIV type 1. AIDS Res. Hum. Retrovir. **13:**913–922.
13. **Esparza, J., and N. Bhamarapravati.** 2000. Accelerating development and future availability of HIV-1 vaccines. Why, when, where and how? Lancet **355:**2061–2066.
14. **Felsenstein, J.** 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics **5:**164–166.
15. **Gao, F., D. L. Robertson, C. D. Carruthers, S. G. Morrison, B. Jian, Y. Chen, F. Barre-Sinoussi, M. Girard, A. Srinivasan, A. G. Abimiku, G. M. Shaw, P. M. Sharp, and B. 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.
16. **Garcia, F., M. Plana, G. M. Ortiz, et al.** 2002. Virological and immunological consequences of structured interruptions in chronic HIV-1 infection. AIDS **15:**F29–40.
17. **Gaschen, B., J. Taylor, K. Yusim, et al.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296:**2354–2360.
18. **Hoelscher, M., B. Kim, L. Maboko, F. Mhalu, R. von Sonnenburg, D. L. Birx, and F. E. McCutchan.** 2001. High proportion of unrelated HIV-1 intersubtype recombinants in the Mgeya region of southwest Tanzania. AIDS **15:** 1461–1470.
19. **Hoffman, T. L., and R. W. Doms.** 1999. HIV-1 envelope determinants for cell tropism and chemokine receptor use. Mol. Membr. Biol. **16:**57–65.
20. **Hoffman, N. G., F. Seillier-Moiseiwitsch, J. Ahn, J. M. Walker, and R. Swanstrom.** 2002. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. J. Virol. **76:**3852–3864.
21. **Hwang, S. S., T. J. Boyle, H. K. Lyerly, and B. R. Cullen.** 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science **253:**71–74.
22. **Inter-Agency Task Team on Mother-to-Child Transmission of HIV.** 2001. New data on the prevention of mother-to-child transmission of HIV and their policy implications: conclusions and recommendations. WHO. Technical Consultation on Behalf of UNFPAUNICEF/WHO./UNAIDS. World Health Organization, Geneva, Switzerland.
23. **Jackson, J., B. G. Becker-Pergola, L. A. Guay, et al.** 2001. Identification of the K103N mutation in Ugandan women receiving nevirapine to prevent HIV-1 vertical transmission. AIDS **14:**F111–F115.
24. **Janssens, W., A. Buve, and J. N. Nkengasong.** 1997. The puzzle of HIV-1 subtype in Africa. AIDS **11:**795–812.
25. **Johansson, B., K. Sherefa, and A. Sonnerborg.** 1995. Multiple enhancer motifs in HIV type 1 strains from Ethiopia. AIDS Res. Hum. Retrovir. **11:**761–764.
26. **Kato, K., H. Sato, and Y. Takebe.** 1999. Role of naturally occurring basic amino acid substitution in the human immunodeficiency virus type 1 subtype E envelope V3 loop on viral coreceptor usage and cell tropism. J. Virol. **73:**5520–5526.
27. **Kazatchkine, M. D., P. N. Van, D. Costagliola, et al.** 2000. Didanosine dosed once daily is equivalent to twice daily dosing of patients on double or triple combination antiretroviral therapy. A1454–147 Team. J. Acquir. Immune Defic. Syndr. **15:**418–424.
28. **Koch, N., J. B. Ndihokubwayo, N. Yahi, C. Tourres, J. Fantini, and C. Tamalet.** 2001. Genetic analysis of HIV type 1 strains in Bujumbura (Burundi): predominance of subtype C variant. AIDS Res. Hum. Retrovir. **17:**269–273.
29. **Korber, B.** 2001. HIV signature and sequence variation analysis, p 55–72. *In* A. G. Rodrigo and G. H. Learn (ed.), Computational analysis of HIV molecular sequences. Kluwer Academic Publishers, Dordrecht, The Netherlands.
30. **Korber, B., C. Brander, B. Haynes, et al.** 2000. HIV-1 molecular immunology. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.Mex.
31. **Korber, B., B. Foley, B. Kuiken, S. K. Pillai, and J. Sodroski.** 1998. Numbering positions in HIV relative to HXB2CG. *In* B. Korber et al. (ed.), Human retroviruses and AIDS. Los Alamos National Laboratory, Los Alamos, N.Mex.
32. **Korber, B., and G. Myers.** 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retrovir. **8:**1549–1560.
33. **Koulinska, I. N., T. Ndung'u, D. Mkakagile, et al.** 2001. A new human immunodeficiency virus type 1 circulating recombinant from Tanzania. AIDS Res. Hum. Retroviruses **17:**423–431.
34. **Kuiken, C. L., B. Foley, B. H. Hahn, et al.** 1999. Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.Mex.
35. **Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei.** 2001. MEGA2: molecular evolutionary genetics analysis software. Arizona State University, Tempe, Ariz.
36. **Lazaro, J. B., J. Boretto, B. Selmi, J. P. Capony, and B. Canard.** 2000. Phosphorylation of AZT-resistant human immunodeficiency virus type 1 reverse transcriptase by casein kinase II in vitro: effects on inhibitor sensitivity. Biochem. Biophys. Res. Commun. **275:**26–32.
37. **Li, Y., M.-A. Rey-Cuille, and S.-L. Hu.** 2001. N-linked glycosylation in the V3 region of HIV type 1 surface antigen modulates coreceptor usage in viral infection. AIDS Res. Hum. Retrovir. **17:**1473–1479.
38. **Loemba, H., B. Brenner, M. A. Parniak, et al.** 2002. Genetic divergence of human immunodeficiency virus type 1 Ethiopian clade C reverse transcriptase (reverse transcriptase) and rapid development of resistance against nonnucleoside inhibitors of reverse transcriptase. Antimicrob. Agents Chemother. **46:**2087–2094.
39. **Lole, K. S., R. C. Bollinger, R. S. Paranjape, et al.** 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.
40. **Losman, B., M. Biller, S. Olofsson, et al.** 1991. The N-linked glycan of the V3 region of HIV-1 gp120 and CXCR4-dependent multiplication of a human immunodeficiency virus type 1 lymphocyte-tropic variant. FEBS Lett. **454:** 47–52.
41. **Maurer-Stroh, S., B. Eisenhaber, and F. Eisenhaber.** 2002. N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. J. Mol. Biol. **317:**523–540.
42. **McCormick-Davis, C., S. B. Dalton, D. K. Singh, and E. B. Stephens.** 2000. Comparison of Vpu sequences from diverse geographical isolates of HIV type 1 identifies the presence of highly variable domains, additional invariant amino acids and a signature sequence motif common to subtype C isolates. AIDS Res. Hum. Retrovir. **16:**1089–1095.
43. **Nakayama, E. E., T. Shioda, M. Tatsumi, et al.** 1998. Importance of the

N-glycan in the V3 loop of HIV-1 envelope protein for CXCR-4 but not CCR-5-dependent fusion. FEBS Lett. **426:**367–372.
44. **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.
45. **Novitsky, V., U. R. Smith, P. Gilbert, et al.** 2002. Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? J. Virol. **76:**5435–5451.
46. **Palella, F. J., K. M. Delaney, A. C. Moorman, et al.** 1998. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. N. Engl. J. Med. **338:**853–860.
47. **Polzer, S., M. T. Dittmar, H. Schmitz, et al.** 2001. Loss of N-linked glycans in the V3-loop region of gp120 is correlated to an enhanced infectivity of HIV-1. Glycobiology **11:**11–19.
48. **Posada, D. K., and A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
49. **Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Homes, and P. H. Harvey.** 2001. The epidemic behaviour of the hepatitis C virus. Science **292:**2323–2325.
50. **Raffi, F., V. Reliquet, V. Ferre, et al.** 2000. The VIRGO study: nevirapine, didanosine and stavudine combination therapy in antiretroviral-naïve HIV-1-infected adults. Antivir. Ther. **5:**267–272.
51. **Rambaut, A.** 2000. Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16:**395–399.
52. **Renjifo, B., B. Chaplin, D. Mwakagile, P. Shah, F. Vannberg, G. Msamanga, D. Hunter, W. Fawzi, and M. Essex.** 1998. Epidemic expansion of HIV type 1 subtype C and recombinant genotypes in Tanzania. AIDS Res. Hum. Retrovir. **14:**635–638.
53. **Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn.** 1995. Recombination in HIV-1. Nature **374:**124–126.
54. **Rodenburg, C. M., Y. Li, S. A. Trask, et al.** 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.
55. **Rollins, N. C., M. Dedicoat, S. Danaviah, et al.** 2002. A new approach to monitoring trends in HIV-1 prevalence, incidence and mother-to-child transmission rates in rural Africa. Lancet **360:**389–390.
56. **Salminen, M., J. Carr, D. Burke, and F. McCutchan.** 1995. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. AIDS Res. Hum. Retrovir. **11:**1423–1425.
57. **Salminen, M. O., B. Johansson, A. Sonnerborg, S. Ayehunie, D. Gotte, P. Leinikki, D. Burke, and F. E. McCuthan.** 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.
58. **Shafer, R. W., J. A. Eisen, T. C.Merigen, and D. A. Katzstein.** 1997. Sequence and drug susceptibility of subtype C reverse transcriptase from human immunodeficiency virus type 1 seroconverters in Zimbabwe. J. Virol. **71:**5441–5448.
59. **Shankarappa, R., R. Chatterjee, and C. Learn.** 2001. Human immunodeficiency virus type 1 *env* sequences from Calcutta in Eastern India: identification of features that distinguish subtype C sequences in India from other subtype C sequences. J. Virol. **75:**10479–10487.
60. **Sharp, P. M., D. L. Robertson, F. Gao, and B. H. Hahn.** 1994. Origins and diversity of human immunodeficiency viruses. AIDS **8:**S27–42.
61. **Shibata, R., C. Siemon, S. C. Czajak, et al.** 1997. Live, attenuated simian immunodeficiency virus vaccines elicit potent resistance against a challenge with a human immunodeficiency virus type 1 chimeric virus. J. Virol. **71:**8141–8148.
62. **Siepel, A. C., A. L. Halpern, C. Macken, and B. Korber.** 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res. Hum. Retrovir. **11:**1413–1416.
63. **Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet.** 1994. The Genetic Data Environment: an expandable GUI for multiple sequence analysis. Comput. Appl. Sci. **10:**671–675.
64. **Soares, M. A., T. De Oliveira, R. M. Brindeiro, R. S. Diaz, E. C. Sabino, L. Brigido, I. L. Pires, M. G. Morgado, M. C. Dantas, D. Barreira, P. R. Teixeira, S. Cassol, A. Tanuri, and the Brazilian Network for Drug Resistance Surveillance.** 2003. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. AIDS **17:**1–11.
65. **Swafford, K. L.** 1999. PAUP 4.0: phylogenetic analysis with parsimony (and other methods), version 4.0b2a. Sinauer Associates Inc., Sunderland, Mass.
66. **Tatt, I. D., K. L. Barlow, A. Nicool, and J. P. Clewley.** 2001. The public health significance of HIV-1 subtypes. AIDS **15:**S59–S71.
67. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
68. **Tozser, J., P. Bagossi, P. Boross, et al.** 1999. Effect of serine and tyrosine phosphorylation on retroviral proteinase substrates. Eur. J. Biochem. **265:**423–429.
69. **UNAIDS/W.H.O. Working Group on Global HIV/AIDS and STD Surveillance.** 2000. Report on the global HIV/AIDS epidemic, December 2000. UN AIDS, Geneva, Switzerland.
70. **Van Harmelen, J. H., E. Van der Ryst, S. A. Loubser, et al.** 1999. A predominantly HIV type 1 subtype C-restricted epidemic in south African urban populations. AIDS Res. Hum. Retrovir. **15:**395–398.
71. **Van Harmelen, J., C. Williamson, B. Kim, L. Morris, J. Carr, S. S. Karim, and F. McCutchan.** 2001. Characterization of full-length HIV-1 type 1 subtype C sequences from South Africa. AIDS Res. Hum. Retrovir. **17:**1527–1531.
72. **Velazquez-Campoy, A., M. J. Todd, S. Vega, and E. Freire.** 2001. Catalytic efficiency and vitality of HIV-1 proteases from African viral subtypes. Proc. Natl. Acad. Sci. USA **98:**6062–6067.
73. **Yang, R., X. Xia, S. Kusagawa, C. Zhang, K. Ben, and Y. Takebe.** 2002. On-going generation of multiple forms of HIV-1 intersubtype recombinants in the Yunnan province of China. AIDS **16:**1401–1407.
74. **Yang, Z.** 2000. Phylogenetic analysis by maximum likelihood (PAML), version 3.0. University College London, London, England.
75. **Yang, Z., R. Nielsen, N. Goldman, and A.-M. Krabbe Pederson.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.
76. **Yu, X. F., J. Chen, Y. Shao, C. Beyrer, and S. Lai.** 1998. Two subtypes of HIV-1 among injection-drug users in southern China. Lancet **351:**1250.
77. **Zacharova, V., M. L. Becker, V. Zachar, P. Ebbesen, and A. S. Goustin.** 1997. DNA sequence analysis of the long terminal repeat of the C subtype of human immunodeficiency virus type 1 from Southern Africa reveals a dichotomy between B subtype and African subtypes on the basis of upstream NF-IL6 motif. AIDS Res. Hum. Retrovir. **13:**719–724.
78. **zur Megede, J., S. Engelbrecht, T. De Oliveira, S. Cassol, T. J. Scriba, E. Janse van Rensburg, and S. W. Barnett.** 2002. Novel evolutionary analyses of full-length HIV-1 subtype C molecular clones from Cape Town, South Africa. AIDS Res. Hum. Retrovir. **18:**1327–1332.