AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 18, Number 17, 2002, pp. 1327–1332
© Mary Ann Liebert, Inc.

Sequence Note

Novel Evolutionary Analyses of Full-Length HIV Type 1 Subtype C Molecular Clones from Cape Town, South Africa

JAN ZUR MEGEDE,[1] SUSAN ENGELBRECHT,[2] TULIO DE OLIVEIRA,[3] SHARON CASSOL,[3] THOMAS J. SCRIBA,[2] ESTRELITA JANSE VAN RENSBURG,[2] and SUSAN W. BARNETT[1]

ABSTRACT

Understanding the origin, distribution, and evolving dominance of HIV-1 subtype C strains is an important component in the design and evaluation of a globally effective AIDS vaccine. To better understand subtype C viruses, we constructed complete molecular clones of primary, CCR-5-using isolates from South Africa and analyzed the molecular phylogenies of these clones using best fitting evolutionary substitution models. Analyses were performed on three full-length sequences, and on the individual genes. All clones were nonrecombinant, and although two of three had open reading frames and intact splice sites, they were not infectious. At the genomic level, the models demonstrated the increasing variability of subtype C in South Africa. At the subgenomic level, they revealed marked differences in the evolutionary patterns of individual genes, a finding that suggests that the genes are under different selective pressures and constraints. These data underscore the dynamic nature of the subtype C epidemic and emphasize the need for continuous monitoring of local strains.

The World Health Organization (WHO) estimated that in the year 2001 more than 40 million people worldwide were infected with human immunodeficiency virus 1 (HIV-1), the causative agent of acquired immunodeficiency syndrome (AIDS). The majority of these infected persons, an estimated 28.1 million, live in sub-Saharan Africa,[1] a region that has a high prevalence and incidence of HIV-1 subtype C viruses. In South Africa, HIV-1 prevalence rates are highest in KwaZulu-Natal (36.2%) and lowest in the Western Cape Province (8.7%).[2]

The rapid escalation of HIV-1 C infections in sub-Saharan Africa, and in several major regions of India and China,[3] makes the development of a subtype C vaccine an international public health priority. An important element of vaccine design is the construction and phylogenetic analysis of full-length sequences of frequently transmitted HIV-1 strains. To date, more than 60 subtype C full-length clones have been published with the majority of these clones being from Botswana.[4,5] Despite the dramatic impact of HIV-1 in South Africa, only five near full-length sequences have been published, four of them originating from Durban, KwaZulu-Natal.[6,7] In this study, we present the cloning and phylogenetic characterization of three full-length molecular clones generated from patients visiting the Infectious Diseases Clinic at Tygerberg Hospital, Cape Town in the Western Cape Province. We also describe the first application of best fitting models to the analysis of full-length sequences.

The primary isolates used in these studies have been well characterized with regard to growth kinetics and coreceptor usage.[8] Three isolates, TV001, TV002, and TV012, have been previously analyzed and classified as subtype C[9–11] based on sequence analysis of the *gag*, *env*, and accessory/regulatory genes. To generate full-length constructs, the peripheral blood mononuclear cell (PBMC) proviral DNA was extracted and

[1]Vaccines Research, Chiron Corporation, Emeryville, California 94608.
[2]Department of Medical Virology, University of Stellenbosch and Tygerberg Hospital, Tygerberg, South Africa.
[3]Africa Centre, University of Natal, Durban, South Africa.

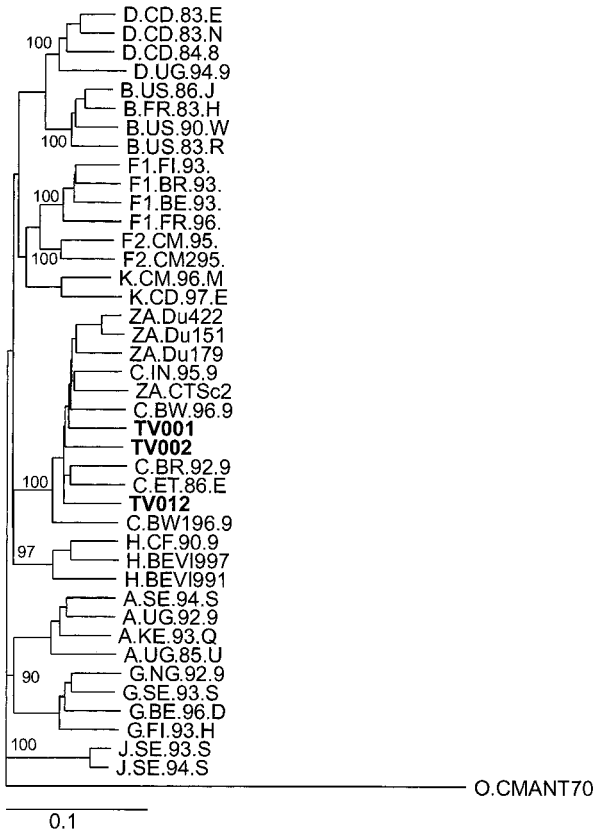

FIG. 1. Maximum likelihood phylogenetic tree analysis of full-length South African HIV-1 sequences and the Los Alamos database reference sequences for the different subtypes. The South African Tygerberg Virology (TV) sequences are indicated in bold lettering. An indication of the degree of sequence dissimilarity is shown on the horizontal axis and the subtypes are indicated on the vertical axis. The percentage of bootstrap trees out of 1000 replications supporting a particular phylogenetic group by more than 80% is placed alongside the node considered. The evolutionary model used was GTR+I+G. The log likelihood score for the phylogram was −83245.1250.

polymerase chain reaction (PCR) amplified using the Expand High Fidelity PCR System (Roche Molecular Biochemicals, Mannheim, Germany). The primers were designed to obtain complete full-length sequences. The 5- and 3-halves of the genome were amplified separately in two different reactions. Primers for the 4.4-kb 5-half (5-LTR to *pol*) were S1FsacTA 5-GTTTCTTGAGCTCTGGAAGGGTTAATTT ACTCCAAGAA-3 and S1RsalTA 5-GTTTCTTGTCGACTTGTCC ATGTATGGCTTCCCCT-3. Primers for the 5.4-kb 3-half (*pol* to 3-LTR) were S2FsalTA 5-GTTTCTTGTCGACTGTAGTCC AGGAATATGGCAATTAG-3 and S2_Full NotTA 5-GTTTCTTGCGGCCGCTGCTA GAGATTTTCCACACTACCA-3. The resultant PCR products were cloned into pCR-XL-TOPO (InVitrogen, Carlsbad, CA) using the TA-cloning system. Positive clones were sequenced on an ABI 310 Genetic Analyzer (Applied Biosystems).

The overlapping half-genomes were combined using an inserted *Sal*I site in a highly conserved region of *pol* for subsequent *in vitro* expression analysis. All three full-length clones, each derived from a different patient, were analyzed for p24 antigen and infectious particle production. Following transfection into human 293 (ATCC CRL-1573) or African green monkey COS-7 cells (ATCC CRL-1651), the transfected cells were cocultivated for 72 hr with phytohemagglutinin (PHA)-stimulated donor PBMCs. After 3 days, the culture supernatants of these cultures were removed and added to fresh PBMC cocultures. This process of removing the supernatant and adding it to fresh PBMCs was repeated once a week for 5 weeks. At various time points, the supernatants were analyzed for cell-free virus using the Coulter HIV-1 p24Gag Core Assay and for RT activity using a nonradioactive reverse transcriptase assay (Boehringer RT assay). During the first 2 weeks of culture, p24Gag was detected in the supernatant of all cultures. However, the expression of p24 antigen was transient, and supernatants collected after 2 weeks (up to 35 days) tested p24Gag negative. Reverse transcription activity was not detected in any of the culture supernatants at any time point. The transient expression of HIV-1 Gag suggested that 5 LTR and *gag* reading

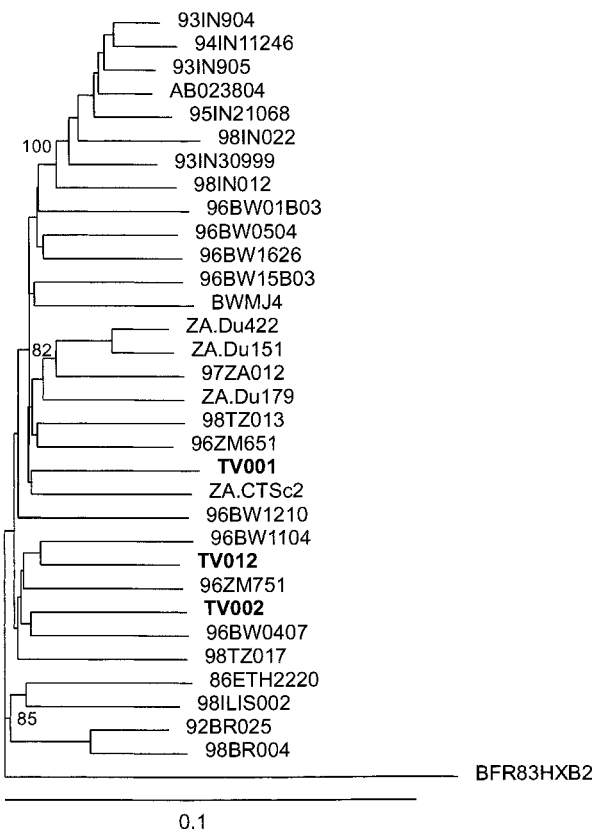

FIG. 2. Maximum likelihood phylogenetic tree analysis of full-length HIV-1 subtype C sequences. The South African Tygerberg Virology (TV) sequences are indicated in bold lettering. An indication of the degree of sequence dissimilarity is shown on the horizontal axis and the subtypes are indicated on the vertical axis. The percentage of bootstrap trees out of 1000 replications supporting a particular phylogenetic group by more than 80% is placed alongside the node considered. The evolutionary model used was GTR+I+G. The log likelihood score for the phylogram was −65376.1367.

Table 1. Evolutionary Model of Substitution Selected for Each Genome Region for Subtype C Sequence Analysis

	gag	*pol*	*vif*	*vpr*	*tat*	*rev*	*vpu*	*env*	*nef*
Model[a]	REV + I + G	TrN + I + G	K81uf + I + G	HKY + I + G	REV + I + G	REV + G	TrN + G	REV + I + G	HKY + I + G
−In L	9133.60	2373.74	3194.85	1856.04	1970.96	2182.11	2019.12	22120.50	4498.09
Pinv	0.3489	0.4872	0.4096	0.3580	0.4341	0	0	0.2661	0.2715
Gamma shape	0.7334	0.7978	0.8595	0.8665	0.8731	0.3731	0.4222	0.4837	0.6889
Ti/Tv				2.7680					1.6505
p value	<0.000001	0.005228	0.000308	0.000644	0.004612	<0.000001	0.273682	<0.000001	0.001843

[a]REV + I + G, general reversible (REV) + gamma distribution; TrN + I + G, Tamura Nei + Pinv + gamma distribution; K81uf + I + G, Kimura, 1981; HKY + I + G, Hasegawa, Kiskino, and Yano, 1985; TVM + G, maximum likelihood; TrN + G, Tamura Nei + gamma distribution.

frame are functional, but that the replication and/or infection process was interrupted. The reading frames of clones TV001 and TV002 were intact as indicated by full-length sequencing. TV012 had a premature stop in ENV that could explain the noninfectivity. Splice site analysis showed no abnormalities when compared to known infectious subtype B and C molecular clones (data not shown). Analysis of protease cleavage sites revealed a higher site-specific variability when compared to other M-group HIV-1 strains (DeOliveira et al., unpublished).

The assembled full-length sequences exhibited slight variations in size: 9781 bp for TV001 clone 8/5_5, 9752 bp for TV002 clone 12/5_1, and 9691 bp for TV012 clone 2. To exclude the possibility of inter- and intrasubtype recombinant forms, recombination analyses were performed using RIP (Recombination Identification Program),[12] and a method based on bootscanning called SIMPLOT.[13] Data analyses using these methods suggested that all three molecular clones were nonrecombinant.

Multiple alignments were performed with CLUSTAL X[14] and the final alignment was manually adjusted. Phylogenetic analysis was performed with the three full-length sequences and the Los Alamos subtype reference sequences. The first tree was plotted with sequences representing group M (Fig. 1), and the second one with sequences representing subtype C (Fig. 2). An appropriate evolutionary model for these sequences was selected using the Akaike information criterion[15] as implemented in Modeltest 3.0.[16] For phylogenetic analysis, the best fitting model was GTR+I+G, a six base reversible substitution model that takes into account the base frequency, the proportion of invariable sites, and the gamma rate of heterogeneity among different genomes with an estimated alpha shape. Parameters of the best fitting model were as follows: equilibrium nucleotide frequencies of $f_A = 0.3792$, $f_C = 0.1806$, $f_G = 0.2257$, $f_T = 0.2145$; proportion of invariable sites (Pinv) = 0.2400; variable sites (G) gamma distribution shape parameter = 0.7170; and R matrix values, $R_{A\text{-}C} = 1.8120$, $R_{A\text{-}G} = 4.5559$, $R_{A\text{-}T} = 0.8489$, $R_{C\text{-}G} = 0.9473$, $R_{C\text{-}T} = 5.7537$, $R_{G\text{-}T} = 1$. A pairwise distance matrix was calculated based on this model and used in the construction of a neighbor-joining tree in version 4.0b2a of PAUP.[17] The log likelihood (−ln L) for this tree was −83245.1250. The topology of this tree was very well supported with high bootstrap values (Fig. 1).

To further examine relationships within the trees, our new South African sequences were compared to a subset of subtype C reference sequences from the Los Alamos database[18] using maximum likelihood analysis. Again, an appropriate evolutionary model (GTR+I+G) for these 31 sequences was selected using the Akaike information criterion. Parameters of this model were as follows: $f_A = 0.3670$, $f_C = 0.1788$, $f_G = 0.2366$, $f_T = 0.2176$; _ = 0.5996; plnv = 0.3356; G = 0.5996; and the substitution model rate matrix of $R_{A\text{-}C} = 1.9411$, $R_{A\text{-}G} = 4.9012$, $R_{A\text{-}T} = 0.9445$, $R_{C\text{-}G} = 0.556$, $R_{C\text{-}T} = 6.0959$, $R_{G\text{-}T} = 1$. The log likelihood of this tree was −65376.1367 (Fig. 2). As expected, the percentage of invariant sites was higher for subtype C viruses, and the alpha shape of the gamma rate of heterogeneity was smaller. An increase in invariant sites is normal and agrees with analyses performed on other similar databases. The small gamma rate of heterogeneity suggests that subtype C viruses contain a small number of sites that evolve quickly, and a large number of sites that evolve slowly.

Figure 2 is a representative tree showing the different subclustering patterns of subtype C viruses. The Indian subcluster was very well supported with a maximal bootstrap value of 100%. Although only two sequences from Brazil were included in the analyses, these sequences also clustered together. Sequences from Israel and Ethiopia formed a third subcluster, a finding that may be due to the phylogenetic noise produced by the large number of African and Indian sequences, rather than a reflection of the relatedness of the viruses. Further analyses are required to differentiate between these two possibilities. In contrast to the Indian and Brazilian sequences, no distinct geographic subclusters were visualized among any of the sequences from southern or eastern Africa, including those from South Africa.

These new sequences from South Africa were interdispersed among other subtype C clusters from Africa, supporting the concept that these infections represent a more longstanding epidemic with multiple introductions from different geographic areas.[4,19,20] The distance matrix revealed an average diversity of 10% between different TV (Tygerberg Virology) isolates. The full-length intrasubtype nucleotide diversity between all described South African isolates ranged from 4% (Du151/Du422) to 13% (97ZA012/CTSc2).

The lack of a discrete South African subcluster contrasts with a recent report by Novitsky *et al.*[21] These investigators analyzed 51 near-full-length sequences from Botswana and five sequences from South Africa, four from KwaZulu-Natal, and one from the Western Cape. In this study, the Botswana sequences formed multiple distinct clusters, or lineages, while the South African sequences segregated as a separate, distinct cluster. This apparent subclustering of South African strains was attributed

Table 2. HIV-1 Subtype C Genome Regions That Support the Complete Genome Tree Topology (Indicated with X)

	gag	*pol*	*vif*	*vpr*	*tat*	*rev*	*vpu*	*env*	*nef*	*Complete genome*
Indian subcluster	X	X	X		X			X		X
Ethiopia/ Brazilian subcluster	X	X	X	X	X		X	X	X	X
African subcluster	X	X	X		X		X	X	X	X
South Africa subcluster						X				

to a phylogenetic founder effect rather than to a bias introduced by including a disproportionate number of sequences from Botswana. Larger sample sizes are required to determine whether full-length sequences from South Africa form country-specific lineages or cluster under the Botswana lineages.

Based on the initial data reported in this study, we suggest that continuous surveillance of local South African strains will reveal multiple cocirculating lineages rather than a single founder effect as has been recently observed in India[21,22] and Brazil.

To determine whether the same subclustering patterns were conserved along the entire genome, best fitting evolutionary methods were applied to each individual gene and the results compared to those obtained for the full-length sequences. Table 1 summarizes the substitution models selected for likelihood analysis of the different genomic regions. Table 2 shows the level of support for each HIV-1 subtype C subcluster analysis across different genomic regions. As shown, different genes showed different patterns of evolution.

The *gag*, *pol*, *vif*, *tat*, and *env* genes fell into three discrete subclusters (Brazil, India, and Africa), similar to those observed for the full-length genomes. None of these subgenomic regions supported a monophyletic South African lineage. *Nef* and *vpr* analysis did not support the Indian subcluster, and *rev* did not support any of the three clustering patterns. Possible reasons for these different subgenomic clustering patterns include low variability in the regions, the existence of gene-specific selective pressures and constraints, and/or intraclade recombination events.

In summary, accumulated evidence suggests that HIV-1 subtype C is rapidly becoming a dominant strain in the global AIDS epidemic. In this study, we have described the first applications of evolutionary models to the analysis of full-length subtype C genomes, as well as the individual genes. We detected marked differences in the evolution of individual genes and suggest that these gene-specific differences may be the result of different selection pressures. The finding that *gag*, *pol*, *vif*, *tat*, and *env* support the same topology as the complete genome may be indicative of a founder effect. At the genomic level, the models provide evidence that African subtype C epidemic is more heterogeneous and of older origin than the Indian and Brazilian epidemics. The lack of a discrete South African subcluster suggests that there are multiple lineages and that the local epidemic is being continuously imported from other African regions.

Given the rapidly evolving nature of the HIV-1 epidemic in South Africa, it will become increasingly important to monitor local strains on an ongoing basis. Knowledge of the changing phylogeny is needed to design vaccines that are directed against epidemiologically important contemporary strains of the virus. Finally, to be successful, any new candidate vaccine or control strategy will need to take into account the changing genetics and epidemic behavior of HIV-1 C viruses. The use of new evolutionary models will provide a valuable tool for tracking these changes and assessing their impact on the behavior of the epidemic.

SEQUENCE DATA

Nucleotide sequences were submitted to GenBank under accession numbers AY162223–2225.

ACKNOWLEDGMENTS

This work was supported by grants from the Wellcome Trust (UK) Grant 061238/2/00/2 (S.C.), the NIH HIV Vaccine Design and Development Team Contract No. NO1-AI-05396 (S.W.B.), the Poliomyelitis Research Foundation (PRF), and the Harry Crossley Foundation (S.E.).

REFERENCES

1. UNAIDS/WHO Report on the global AIDS epidemic—update December 2001. UNAIDS, Geneva, 2001. (http://www.unaids.org/epidemic_update/report_dec01/index.html).
2. Department of Health/Directorate Health Systems Research: Seventh national HIV survey of women attending antenatal clinics of the public health service in South Africa. October/November 2000. Directorate Health Systems Research. Department of Health, Pretoria, South Africa, 2001.
3. Esparza J and Bhamarapravati N: Accelerating the development and future availability of HIV-1 vaccines: Why, when, where, and how? Lancet 2000;355:2061–2066.
4. Novitsky VA, Montano MA, McLane MF, Renjifo B, Vannberg F, Foley BT, Ndung'u TP, Rahman M, Makhema MJ, Marlink R, and Essex M: Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: A set of 23 full-length clones from Botswana. J Virol 1999;73:4427–4432.
5. Ndung'u T, Renjifo B, Novitsky VA, McLane MF, Gaolekwe S, and Essex M: Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana. Virology 2000; 278:390–399.
6. Rodenburg CM, Li Y, Trask SA, Chen Y, Decker J, Robertson DL, Kalish ML, Shaw GM, Allen S, Hahn BH, and Gao F: Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res Hum Retroviruses 2001;17:161–168.
7. Van Harmelen J, Williamson C, Kim B, Morris L, Carr J, Abdool Karim SS, and McCutchan F: Characterization of full-length HIV type 1 subtype C sequences from South Africa. AIDS Res Hum Retroviruses 2001;17:1527–1531.
8. Treurnicht F, Smith T, Engelbrecht S, Claassen M, Robson B, Zeier M, and van Rensburg E: Genotypic and phenotypic analysis of the env gene from South African HIV-1 subtype B and C isolates. J Med Virol 2001;68:141–146.
9. Engelbrecht S, de Villiers T, Sampson CC, zur Megede J, Barnett SW, and van Rensburg EJ: Genetic analysis of the complete gag and env genes of HIV type 1 subtype C primary isolates from South Africa. AIDS Res Hum Retroviruses 2001;17:1533–1547.
10. Scriba TJ, Treurnicht FK, Zeier M, Engelbrecht S, and van Rensburg EJ: Characterization and phylogenetic analysis of South African HIV-1 subtype C accessory genes. AIDS Res Hum Retroviruses 2001;17:775–781.
11. Scriba TJ, de Villiers T, Treurnicht FK, zur Megede J, Barnett SW, Engelbrecht S, van Rensburg EJ: Characterization of South African HIV type 1 subtype C complete 5 long terminal repeat, nef, and regulatory genes. AIDS Res Hum Retroviruses 2002;18:149–159.
12. Siepel AC, Halpern AL, Macken C, and Korber BT: A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res Hum Retroviruses 1995;11: 1413–1416.
13. Salminen MO, Carr JK, Burke DS, and McCutchan FE: Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses 1995;11: 1423–1425.
14. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin, and Higgins

DG: The ClustalX-Windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
15. Akaike H: A new look at statistical model identification. IEEE Trans Automatic Control 1974;19:716–723.
16. Posada D and Crandall KA: MODELTEST: Testing the model of DNA substitution. Bioinformatics 1998;14:817–818.
17. Swofford DL: PAUP* Phylogenetic analysis using Parsimony (*and other methods), Version 4. Sinauer Associates, Sunderland, MA, 2000.
18. Human Retroviruses and AIDS 2000: *A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences* (Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, and Wolinsky S, eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 2000.
19. Van Harmelen JH, Van der Ryst E, Loubser AS, York D, Madurai S, Lyons S, Wood R, and Williamson C: A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. AIDS Res Hum Retroviruses 1999;15:395–398.
20. Abebe A, Lukashov VV, Pollakis G, Kliphuis A, Fontanet AL, Goudsmit J, and de Wit TF: Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. AIDS 2001;15(12):1555–1561.
21. Novitsky V, Smith UR, Gilbert P, McLane MF, Chigwedere P, Williamson C, Ndung u T, Klein I, Chang SY, Peter T, Thior I, Foley BT, Gaolekwe S, Rybak N, Gaseitsiwe S, Vannberg F, Marlink R, Lee TH, and Essex M: Human immunodeficiency virus type 1 subtype C molecular phylogeny: Consensus sequence for an AIDS vaccine design? J Virol 2002;76:5435–5451.
22. Shankarappa R, Chatterjee R, Learn GH, Neogi D, Ding M, Roy P, Ghosh A, Kingsley L, Harrison L, Mullins JI, and Gupta P: 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 2001;75:10479–10487.

Address reprint requests to:
Jan zur Megede
Vaccines Research
Chiron Corporation/Mailstop 4.3
4560 Horton Street
Emeryville, CA 94608

E-mail: Jan_zur_Megede@Chiron.com