Updated Proposal of Reference Sequences of HIV-1 Genetic Subtypes

Thomas Leitner,[1] **Bette Korber,**[1] **David Robertson**[2] **, Feng Gao,**[3] **Beatrice Hahn**[3]

[1] *Theoretical Biology and Biophysics, Group T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545;*
[2] *Laboratory of Structural & Genetic Information, CNRS-EP 91, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France;*
[3] *Department of Medicine and Microbiology, University of Alabama at Birmingham, 701 S. 19th Street, LHRB 613, Birmingham, AL 35294.*

The known universe of HIV-1 viruses is divided into two **groups**, a major "M" group and an outlier "O" group, (reviewed in [9]). HIV-1 **subtypes** are clusters of sequences within the M group that are defined by phylogenetic analysis [6, 12]. In the *Human Retroviruses and AIDS 1996* compendium, a reference set of HIV-1 sequences representing the different subtypes was proposed. Since that issue of the compendium, many more full length sequences from the various HIV-1 M group subtypes have been produced. Partly as a consequence of the availability of these new sequences, and partly due to further analysis of pre-existing data, new information has been accumulating concerning subtyping and HIV-1 alignments and hybrid genomes [3, 6, 8, 14]. In response we are updating the proposed list of reference sequences that appeared in Table 1 page III-30 of *Human Retroviruses and AIDS 1996*. We anticipate updating this table and the accompanying alignments on a regular basis in the future as a service to the scientific community, particularly since new submissions are likely to include a considerable number of complex mosaic genomes.

Table 1 lists reference sequences for 9 group M subtypes in major coding regions, i.e., *gag*, *pol*, *env* and *nef*. The criteria for inclusion of a reference sequence have changed since 1996, with a shift in emphasis to using full length HIV genomic sequences to serve as representatives for the subtypes, when such sequences are available. These are supplemented with sequences spanning intact coding regions, or else the longest gene fragments currently available. Table 1A summarizes a basic list of reference sequences for international variation and subtyping efforts; Table 1B provides additional information, including GenBank accession numbers, citations describing the respective sequence, sampling year, and the country where the virus was collected. Protein and nucleotide sequence alignments are available from the HIV database for each of the four coding regions on the web site at http://hiv-web.lanl.gov. These alignments also contain the HIV-1 O group sequences ANT70 and MVP5180 and chimpanzee viral sequences CPZ-GAB and CPZ-ANT. Full length nucleotide genome alignments are also available in Part I of this volume. The alignments were generated using first a hidden Markov model [4] at the nucleotide level, and then manually corrected at the amino acid level to keep open reading frames in frame, and finally back-translated into nucleotides. We include here maximum likelihood trees [5, 15] showing the phylogenetic relationships of the representative sequences selected for the *gag*, *pol*, *env* and *nef*. Trees for V3 and p17 sequences are also included, because: 1) these genomic regions are commonly sequenced; 2) sequences are available for all of the selected reference strains, including the shorter fragments; and 3) the organization of the V3 region tree is somewhat altered relative to the full length *env* tree.

Some of the HIV-1 subtypes are more clearly defined than others. Each subtype A, B, C, D, F, and H, contains at least one full length, apparently non-recombinant genome available as a reference sequence, as well as multiple additional full length *env* and *gag* sequences. Non-recombinant means there are no identified conflicting subtype associations in different regions of the sequence. Our current database and methods, however, have limitations, and additional sequences and analyses in the future may change our current understanding of subtype relationships. Short fragments of genes are unreliable for subtyping a sequence. For example, while the full-length representative subtype H sequence is essentially subtype H throughout the genome (see trees), phylogenetic analyses of V3 sequences occasionally produce trees that show a close association of subtype H and A sequences. Since this

discordance is not always observed but depends on the particular alignments and programs used, it is not sufficient evidence for recombination. Nevertheless, the fact that it can occur, underscores the possibility that ambiguities can arise with subtyping efforts, particularly those focused on short regions of the genome.

All full length representatives of subtypes E and G that have been sequenced to date represent mosaic genomes, with parts of the viral genome clustering with the A subtype in phylogenetic analysis, and parts of the genome forming the two clearly distinguishable clades designated either E or G. The E subtype is clearly "E" in envelope, appears to be an A/E mosaic in the regulatory regions, and is essentially A-associated in *gag* and *pol* [3, 8]. All of the longer subtype G sequences available to date have stretches of subtype A-associated sequence interspersed. In contrast to the A/E mosaics, however, these A/G mosaic sequences have many different patterns of A-like sequence, suggesting that they resulted from independent recombination events [6, 13, 14]. Subtype I sequences [10] have also been reanalyzed [7, 16] and there is now evidence from analysis of a complete genome that they are multiply mosaic and composed of 3 (or of possibly 4) subtypes [7]. Further analysis is necessary to determine the potential they have to represent a distinct subtype. Only fragments of subtype J *env* and *gag* genes are available at this time [11], so further work is needed in this case as well to fully characterize these strains. Nevertheless, these sequences are included as reference sequences to assist in identification of possible J subtype related strains in the future. In addition to the two subtype J sequences listed in the table, three other sequences (GM4, GM5, GM7) have been found to cluster close to this subtype over the *env* V3 region [1]. However, the sequence of GM4 is suggested to be a G/?/C recombinant [2], where the question mark represents a $\sim$600 bp section that includes the J-like V3 region.

REFERENCES

[1] Blouin, J. C., E. A. Guzman, and B. T. Foley. 1996. Global variation in the HIV-1 V3 region, p. III-77–III-201. In G. Myers, B. Korber, B. Foley, K.-T. Jeang, J. W. Mellors, and S. Wain-Hobson (ed.), *Human Retroviruses and AIDS: a compilation and analysis of nucleic and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM.

[2] Bobkov, A., R. Cheingsong-Popov, M. Salminen, F. McCutchan, J. Louwagie, K. Ariyoshi, H. Whittle, and J. Weber. 1996. Complex mosaic structure of the partial envelope sequence from a Gambian HIV type 1 isolate. *AIDS Res. Hum. Retrovirus*. **12**:169–171.

[3] Carr, J. K., M. O. Salminen, C. Koch, D. Gotte, A. W. Artenstein, P. A. Hegerich, D. St. Louis, D. S. Burke, and F. E. McCutchan. 1996. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J. Virol*. **70**:5935–5943.

[4] Eddy, S. 1995. HMMER Hidden Markov Models of Protein and DNA Sequence, 1.8 ed. Washington University School of Medicine, St. Louis, MO.

[5] Felsenstein, J. 1993. PHYLIP: Phylogeny Inference Package, 3.52c ed. University of Washington, Seattle, WA.

[6] 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. H. Hahn. submitted 1997. Non-recombinant reference clones and sequences for human immunodeficiency virus type 1 subtypes A, C, D, F, and H.

[7] Gao, F., D. L. Robertson, and B. H. Hahn. 1997, unpublished data.

[8] 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.

[9] Korber, B., I. Loussert-Ajakai, J. Blouin, and S. Saragosti. 1997. A comparison HIV-1 group M and group O functional and immunogenic domains in the *gag* p24 protein and the C2V3 region of the envelope protein, p. III-41–III-56. In G. Myers, B. Korber, B. Foley, K.-T. Jeang, J. W. Mellors, and S. Wain-Hobson (ed.), *Human Retroviruses and AIDS 1996: a compilation and analysis of nucleic acid and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM.

[10] Kostrikis, L. G., E. Bagdades, Y. Cao, L. Zhang, D. Dimitriou, and D. D. Ho. 1995. Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. *J. Virol*. **69**:6122–6130.

[11] Leitner, T., A. Alaeus, S. Marquina, E. Lilja, K. Lidman, and J. Albert. 1995. Yet another subtype of HIV type 1? *AIDS Res. Hum. Retrovirus*. **11**:995-997

[12] Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, B. E. Sanders, 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.

[13] McCutchan, F. 1997, personal communication.

[14] McCutchan, F. E., M. O. Salminen, J. K. Carr, and D. S. Burke. 1996. HIV- 1 genetic diversity. *AIDS*. 10 (suppl 3):S13-S20.

[15] Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci*. **10**:41–48.

[16] Salminen, M. 1996, personal communication.

Table 1A Updated Proposal of Reference Sequences of HIV-1 Genetic Subtypes

Subtype	*gag*	*pol*	*env*	*nef*
A	U455 92UG037.1 K89 VI32	U455 92UG037.1	U455 92UG037.1 K89(KENYA) SF170	U455 92UG037.1
B	HXB2 JRFL OYI RF	HXB2 JRFL OYI RF	HXB2 JRFL OYI RF	HXB2 JRFL OYI RF
C	ETH2220 92BR025.8 UG268 ZAM18	ETH2220 92BR025.8	ETH2220 92BR025.8 UG268 ZAM18	ETH2220 92BR025.8
D	NDK Z2Z6 ELI 94UG114.1[1]	NDK Z2Z6 ELI 94UG114.1[1]	NDK Z2Z6 ELI 94UG114.1[1]	NDK Z2Z6 ELI 94UG114.1[1]
E[2]			CM240 TN235 90CR402.1 93TH253.3	
F	93BR020.1 BZ162 VI69 VI174	93BR020.1	93BR020.1 BZ163 BZ126 RJI03	93BR020.1
G[3]	92NG003.1[4] 92NG083.2 SE6165[5]	92NG083.2 SE6165[5]	92NG003.1[4] 92NG083.2 92UG975.10 92RU131.9	92NG003.1[4] 92NG083.2
H	90CF056.1 VI557[5]	90CF056.1	90CF056.1 VI557[5] CA13[5]	90CF056.1
J	SE7022[5] SE7887[5]		SE7022[5] SE7887[5]	

[1] The sequence 94UG114.1 is the most distant complete genome D subtype sequence (see trees), tending to branch off closest to the B/D root in most analyses. In some subgenomic regions, it may even move outside the B/D cluster.

[2] E in most of *env*, A in *gag* and *pol* , mixture of A & E in regulatory genes [3, 8, 14].

[3] The G reference sequences may show resemblance to subtype A in regions of *pol* and *vif*, see also footnote 4.

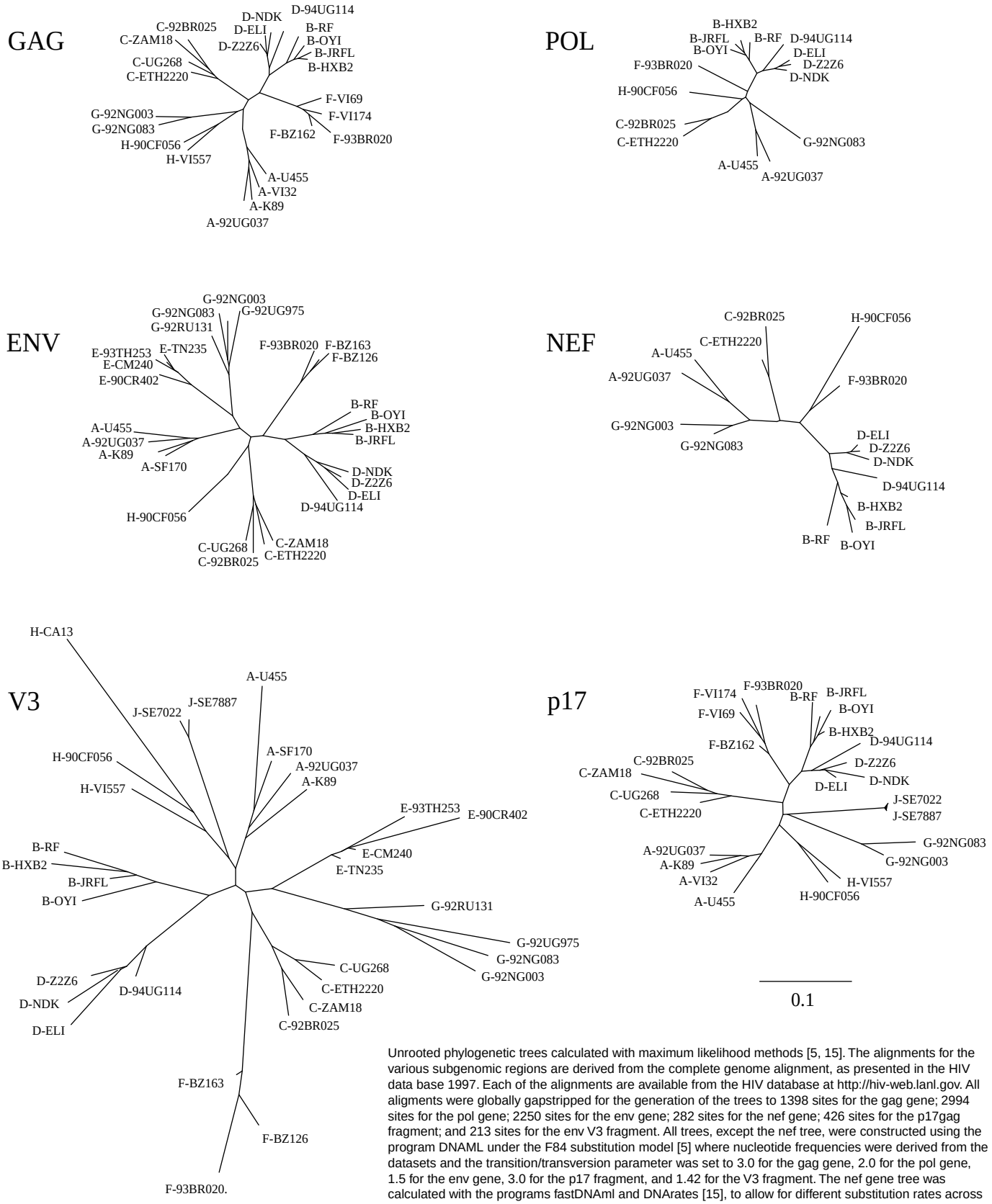
[4] 92NG003.1 is a full length sequence, but is not included in the *pol* alignment because of A-like regions in this gene [6].

[5] Full length gene sequences of *gag*, *pol*, or *env* are not yet available, see Table 1B.

Table 1B Sequence descriptions

Subtype	Sequence	Acc. No.	Source	Region	Sampling year	Sampling country (origin)
HIV-1 M Group sequences in alignments						
A	U455	M62320	Oram, J.D. et al, *ARHR* **6**:1073–1078 (1990)	complete	NA	Uganda
A	92UG037.1	U51190	Gao, F. et al, *J.Virol.* **70**:7013–7029 (1996)	complete	1992	Uganda
A	K89	L11774	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	NA	Kenya
A	VI32	L11788	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	NA	Burundi
A	SF170	M66533	Evans, L. et al, *PNAS* **85**:2815 (1988)	*env*	NA	Rwanda
B	HXB2	K03455, M38432	Wong-Staal, F. et al, *Nature* **313**:277–284 (1985)	complete	NA	France
B	JRFL	U63632	O'Brien, W.A. et al, *Nature* **348**:69 (1990)	complete	NA	US
B	OYI	M26727	Wain-Hobson, S. et al, *AIDS* **3**:707 (1989)	complete	NA	Gabon
B	RF	M17451, M12508	Starcich, B.R. et al, *Cell* **45**:637–648 (1986)	complete	1983	US (Haiti)
C	ETH2220	U46016	Salminen, M.O. et al, *ARHR* **12**:1329–1339 (1996)	complete	1986	Ethopia
C	92BR025.8	U52953	Gao, F. et al, in preparation (1997)	complete	1992	Brazil
C	UG268	L11799	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	1993	Uganda
C	UG268	L22948	Louwagie, J. et al, *J.Virol.* **69**:263–271 (1995)	*env-nef*	1993	Uganda
C	ZAM18	L03705	McCutchan, F. et al,*JAIDS* **5**:441–449 (1992)	*gag*	1989	Zambia
C	ZAM18	L22954	Louwagie, J. et al, *J.Virol.* **69**:263–271 (1995)	*env*	1989	Zambia
D	NDK	M27323	Spire, B. et al, *Gene* **81**:275–284 (1989)	complete	NA	Zaire
D	Z2Z6	M22639	Theodore, T. et al, unpublished (1988)	complete	NA	Zaire
D	ELI	K03454, X04414	Alizon, M. et al, *Cell* **46**:63–74 (1986)	complete	NA	Zaire
D	94UG114.1	U88824	Gao, F et al, in prepapration (1997)	complete	1994	Uganda
E	CM240	U54771	Carr, J.K. et al, *J.Virol.* **70**:5935–5943 (1996)	complete	1990	Thailand
E	TN235	L03698	McCutchan, F.E. et al, *ARHR* **8**:1887–1895 (1992)	*env*	NA	Thailand
E	90CR402.1	U51188	Gao, F. et al, *J.Virol.* **70**: 7013–7029 (1996)	complete	1990	Central African Republic
E	93TH253.3	U51189	Gao, F. et al, *J.Virol.* **70**: 7013–7029 (1996)	complete	1993	Thailand
F	BZ162	L11751	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	NA	Brazil
F	VI69	L11796	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	NA	Belgium (Rwanda)
F	VI174	L11782	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	NA	Zaire
F	BZ163	L22085	Louwagie, J. et al, *ARHR* **10**:561–567 (1994)	*env-nef*	NA	Brazil
F	BZ126	L22082	Louwagie, J. et al, *ARHR* **10**:561–567 (1994)	*env-nef*	NA	Brazil
F	93BR020.1	AF005494	Gao, F. et al, in preparation (1997)	complete	1993	Brazil
F	RJI03	U08974	Sabino, E.C., et al., *J.Virol.* **68**:6340–6346 (1994)	partial *env*	NA	Brazil
G	92NG003.1	U88825	Gao, F. et al, in preparation (1997)	complete	1992	Nigeria
G	SE6165	L40752, L40761	Leitner, T. et al, *Virology* **209**:136–146 (1995)	p17, RT	1993	Sweden (Central Africa)
G	92NG083.2	U88826	Gao, F. et al, in preparation (1997)	complete	1992	Nigeria
G	92UG975.10	U27426	Gao, F. et al, *J.Virol.* **70**:1651–1657 (1996)	*env*	1992	Uganda
G	92RU131.9	U30312	Gao, F. et al, *J.Virol.* **70**:1651–1657 (1996)	*env-nef*	1992	Russia
H	90CF056.1	AF005496	Gao, F. et al, in preparation (1997)	complete	1990	Central African Republic
H	VI557	U09666	Janssens, W. et al, *ARHR* **10**:877–879 (1994)	V3-V5	NA	Zaire
H	VI557	L11793	Louwagie, J. et al, *AIDS* **7**:769–780 (1993)	*gag*	NA	Zaire
H	CA13	U09667	Janssens, W. et al, *ARHR* **10**:877–879 (1994)	V3-V5	NA	Cameroon
J	SE7022	L41177, L41179	Leitner, T. et al, *ARHR* **11**:995–997 (1995)	V3, p17	1993	Sweden (Zaire)
J	SE7887	L41176, L41178	Leitner, T. et al, *ARHR* **11**:995–997 (1995)	V3, p17	1994	Sweden
Additional sequences available in alignments						
O Group	ANT70	L20587	Vanden Haesevelde, M. et al, *J.Virol.* **68**:1586–1596 (1994)	complete	NA	Cameroon
O Group	MVP5180	L20571	Gurtler, L. et al, *J.Virol.* **68**:1581–1585 (1994)	complete	1991	Cameroon
Chimpanzee	SIV-CPZANT	U42720	Vanden Haesevelde, M. et al,*Virology* **221**:346–350 (1996)	complete	1986	Zaire
Chimpanzee	SIV-CPZGAB	X52154	Huet, T. et al, *Nature* **345**:356–359 (1990)	complete	NA	Gabon

HIV-1 Genetic Subtypes



Unrooted phylogenetic trees calculated with maximum likelihood methods [5, 15]. The alignments for the various subgenomic regions are derived from the complete genome alignment, as presented in the HIV data base 1997. Each of the alignments are available from the HIV database at http://hiv-web.lanl.gov. All aligments were globally gapstripped for the generation of the trees to 1398 sites for the gag gene; 2994 sites for the pol gene; 2250 sites for the env gene; 282 sites for the nef gene; 426 sites for the p17gag fragment; and 213 sites for the env V3 fragment. All trees, except the nef tree, were constructed using the program DNAML under the F84 substitution model [5] where nucleotide frequencies were derived from the datasets and the transition/transversion parameter was set to 3.0 for the gag gene, 2.0 for the pol gene, 1.5 for the env gene, 3.0 for the p17 fragment, and 1.42 for the V3 fragment. The nef gene tree was calculated with the programs fastDNAml and DNArates [15], to allow for different substitution rates across sites. This proved to be important for the topology of this gene tree in resolving subtypes B and D. Athough G subtype sequences in the trees shown cluster separate from subtype A, they cluster in subtype A in some subgene regions [6]. Subtypes B and D are generally close to each other in all analyses, but with the nef gene it may be difficult to completely resolve them from each other. All trees are drawn to the same scale, thereby indicating the relative information density in the different regions.