HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences, 2005

Thomas Leitner, Bette Korber, Marcus Daniels, Charles Calef, Brian Foley

Los Alamos National Laboratory, Los Alamos, NM 87545
seq-info@t-10.lanl.gov http://hiv.lanl.gov/

Recently there has been intensive activity in the area of discovering new HIV-1 CRFs, although many of the new CRFs have not yet been published (Table 1). In addition, there have been suggestions of new sub-subtypes and potentially a new subtype (10, 11). The classification of new HIV-1 sequences follows the proposed HIV nomenclature guidelines (13, 14). When classifying new sequences, the HIV-1 subtyping reference set is often used. This set has not been updated since 2001, and thus it is time to update this set.
The criteria for updating the reference set were:

1. Four sequences of each HIV-1 group, subtype and sub-subtype, are included, if available.
2. The four sequences should roughly describe the diversity of each class as an effective population.
3. Further selection criteria for reference sequences were:
   a) full length genomes that cover all genes,
   b) no clear sign of recombinant history,
   c) published with a peer reviewed citation,
   d) recent rather than older samples,
   e) covered major geographic distribution,
   f) no sign of hypermutation,
   g) not synthetic, *i.e.,* real sequences from a patient
   h) no extreme indels,
   i) viable and intact as far as known.
4. The CRFs included are now described by one sequence, the prototype of each recombination pattern. Thus the breakpoint pattern is based on the prototype, and will agree with the updated CRF page.

The alignments were based on the 2004 web and compendium alignments, which were constructed using HMMER (with a model calculated on the 2003 alignment) (2). The alignment was further improved using SynchAligns (http://hiv-web.lanl.gov/content/hiv-db/SYNCH_ALIGNS/SynchAligns.html) to add additional sequences, GeneCutter (http://hiv-web.lanl.gov/content/hiv-db/GENE_CUTTER/cutter.html) to correct reading frames, and manual editing using Se-Al (12).

Notes regarding the updated reference set

As in the original subtype reference proposals (1, 6, 9), four sequences per subtype were chosen so that the reference set remains small while allowing the diversity of each subtype to be roughly the same as for all available sequences (similar to an effective population size). In addition, four taxa is the smallest informative unit in an unrooted tree.

Subtype A: Sub-subtype A1 is well established and somewhat more diverse than some other subtypes. A1 is updated with one more recent sequence, and the oldest and divergent reference sequence U455 has been omitted. Sub-subtype A2 is less well established, and as only three full genome sequences have been described there is no choice as to which sequences to include in this sub-subtype. Sub-subtype A3 (10) has not been included at this time because it is less well established and also does not cluster separately from A1 throughout the genome. Similarly, the proposed sub-subtypes A4 and A5

Table 1A Description of subtype reference sequences

Subtype	Sequence	Acc. No.	Year of sampling	Country of sampling (origin)	Reference
A1	92UG037.1	U51190	1992	Uganda	Gao, F. *J Virol* **70**:7013–29(1996)
A1	Q23	AF004885	1994	Kenya	Poss, M. *J Virol* **72**:8240–51(1998)
A1	SE7253	AF069670	1994	Sweden (Somalia)	Carr, J.K. *AIDS* **13**:1819–26(1999)
A1	UG57136	AF484509	1998	Uganda	Harris M.E. *ARHR* **18**:1281–90(2002)
A2	CDKTB48	AF286238	1997	DRC	Gao, F. *ARHR* **17**:675–88(2001)
A2	CY017	AF286237	1994	Cyprus	Gao, F. *ARHR* **17**:675–88(2001)
B	HXB2	K03455	1983	France	Wong-Staal, F. *Nature* **313**:277–284 (1985)
B	BK132	AY173951	1990	Thailand	Hierholzer, J. *ARHR* **18**:1339–50(2002)
B	671	AY423387	2000	Netherlands	Geels, M.J. *J Virol* **77**:12430–40(2003)
B	1058	AY331295	1998	USA	Bernardin, F. *J Virol* **79**:11523–8(2005)
C	ETH2220	U46016	1986	Ethopia	Salminen, M.O. *ARHR* **12**:1329–39(1996)
C	92BR025.8	U52953	1992	Brazil	Gao, F. *J Virol* **70**:1651–57(1996)
C	IN21068	AF067155	1995	India	Lole, K.S. *J Virol* **73**:152–60(1999)
C	SK164B1	AY772699	2004	South Africa	Kiepiela, P. *Nature* **432**:769–75(2004)
D	ELI	K03454	1983	DRC	Alizon, M. *Cell* **46**:63–74(1986)
D	94UG114.1	U88824	1994	Uganda	Gao, F. *J Virol* **72**:5680–98(1998)
D	4412HAL	AY371157	2001	Cameroon	Kijak, G.H. *ARHR* **20**:521–30(2004)
D	A280	AY253311	2001	Tanzania	Herbinger, K-H. *ARHR* **20**:895–901(2004)
F1	93BR020-1	AF005494	1993	Brazil	Gao, F. *J Virol* **72**:5680–98(1998)
F1	VI850	AF077336	1993	Belgium (DRC)	Laukkanen, T. *Virology* **269**:95–104 (2000)
F1	FIN9363	AF075703	1993	Finland	Laukkanen, T. *Virology* **269**:95–104 (2000)
F1	MP411	AJ249238	1996	France	Triques K. *ARHR* **16**:139–151(2000)
F2	MP255	AJ249236	1995	Cameroon	Triques K. *ARHR* **16**:139–151(2000)
F2	MP257	AJ249237	1995	Cameroon	Triques K. *ARHR* **16**:139–151(2000)
F2	0016BBY	AY371158	2002	Cameroon	Kijak, G.H. *ARHR* **20**:521–30(2004)
F2	CM53657	AF377956	1997	Cameroon	Carr, J.K. *Virology* **286**:168–81(2001)
G	SE6165	AF061642	1993	Sweden (DRC)	Carr, J.K. *Virology***247**:22–31 (1998)
G	HH8793.1.1	AF061640	1993	Finland (Kenya)	Carr, J.K. *Virology***247**:22–31 (1998)
G	DRCBL	AF084936	1996	Belgium (DRC)	Debyser Z. *ARHR* **14**:453–9 (1998)
G	NG083	U88826	1992	Nigeria	Gao, F. *J Virol* **72**:5680–98(1998)
H	056.1	AF005496	1990	Cent. Afr. Rep.	Gao, F. *J Virol* **72**:5680–98(1998)
H	VI991	AF190127	1994	Belgium (?DRC)	Janssens, W. *AIDS* **14**:1533–43(2000)
H	VI997	AF190128	1993	Belgium (?DRC)	Janssens, W. *AIDS* **14**:1533–43(2000)
J	SE9280.9	AF082394	1994	Sweden (DRC)	Laukkanen, T. *ARHR* **15**:293–97(1999)
J	SE9173.3	AF082395	1993	Sweden (DRC)	Laukkanen, T. *ARHR* **15**:293–97(1999)
K	EQTB11C	AJ249235	1997	DRC	Triques K. *ARHR* **16**:139-51(2000)
K	MP535	AJ249239	1996	Cameroon	Triques K. *ARHR* **16**:139-51(2000)

are not included because they have yet not been published (11). Finally, the cluster of A1 sequences from the former Soviet Union countries is well defined in all genomic regions of the HIV genome, but at this time it is not assigned a separate sub-subtype to avoid confusion with other potential sub-subtype candidates.

Subtypes B and D: Subtypes B and D are closer to each other than other subtypes. In most genomic regions they behave as sub-subtypes. Subtype D has a large sample from Uganda (27 out of 51 available sequences), which makes it look as if D may be divided into sub-subtypes. But this is not true; it is simply an effect of dense sampling of an Ugandan population (3). The D reference sequences have

Table 1B Description of CRF and other reference sequences

CRF	Mix	Sequence	Acc. No.	Year of sampling	Country of sampling (origin)	Reference
01	AE	CM240	U54771	1990	Thailand	Carr, J.K. *J Virol* **70**:5935–43 (1996)
02	AG	IBNG	L39106	NA	Nigeria	Howard, T.M. *ARHR* **10**:1755–57 (1994)
03	AB	KAL153	AF193276	NA	Russia	Salminen, M.O., *Unpublished* (1998)
04	AGHKU	CY032.3	AF049337	1994	Cyprus (Greece)	Gao, F. *J Virol* **72**:10234–41 (1998)
05	DF	VI310	AF193253	NA	Belgium (?DRC)	Laukkanen, T. *Virology* **269**:95-104 (2000)
06	AGJK	BFP90	AF064699	1996	Australia (Burkina Faso)	Oelrichs, R.B. *ARHR* **14**:1495–500 (1998)
07	BC	CN54	AX149771	1997	China	Guan, Q. *J. Clin. Microbiol.* **42**:4261–7 (2004)
08	BC	GX-6F	AY008715	1997	China	Piyasirisilp, S. *J Virol***74**:11286–95 (2000)
09		96GH2911	AY093605	1996	Ghana	McCutchan, F.E. *ARHR* **20**:819–26 (2004)
10	CD	TZBF061	AF289548	1996	Tanzania	Koulinska, I.N. *ARHR* **17**:423–31 (2001)
11	A01GJ	GR17	AF179368	NA	Greece (?DRC)	Paraskevis, D. *ARHR* **16**:845–55 (2000)
12	BF	ARMA159	AF385936	1991	Argentina	Carr, J.K. *AIDS* **15**:F41–7 (2001)
13	A01GJU	96CM-1849	AF460972	1996	Cameroon	Wilbe, K. *ARHR* **18**:849–56 (2002)
14	BG	X397	AF423756	1999	Spain	Delgado, E. *JAIDS* **29**:536–43 (2002)
15	01B	99TH.MU2079	AF516184	1999	Thailand	Viputtigul, K. *ARHR* **18**:1235–7 (2002)
16	A2D	KISII5009	AF457060	NA	Kenya	Dowling,W.E. *AIDS* **16**:1809–20 (2002)
17	BF1	BF	PSP0096			Jean Carr unpublished
18		CU14	AY586540	NA	Cuba	Thomson, M. *AIDS* **19**:1155–63 (2005)
19		CU38	AY588970	1999	Cuba	Cuevas, M.T. *AIDS* **16**:1643–53 (2005)
20	BG					Michael Thomson unpublished
21	A2D					Francine McCutchan unpublished
22	01A1	3097MN				Jean Carr unpublished
23	01A1					Jean Carr unpublished
24	BG					Michael Thomson unpublished
25	BG					Michael Thomson unpublished
26	AG					Jean Carr unpublished
27						Martine Peeters unpublished
28						Martine Peeters unpublished
29	BF	UFRJ1				Janini unpublished
30	BF	BREMP12313				Janini unpublished
31	0206					Martine Peeters unpublished
32	02A1					Jean Carr unpublished

been updated to reflect the modern diversity of D. Similarly, subtype B has been updated with more recently sampled sequences. Also, since subtype B is involved in several epidemics in Asia, a reference sequence from this region is included in the set.

Subtype C: This subtype is the most well-described subtype as measured by available full-length genomes, and many countries are represented. There is no sign of subdivision within this subtype, although there is limited geographic clustering of subtype sequences from Asia versus those form India. One more recent sequence from South Africa has been added to the current reference set.

Subtypes F, G, H, J and K: Few full genomes exist, and what is available has been used. Sub-subtype F2 now has four sequence representatives, but otherwise there are no changes since the 2001 reference set.

Table 1C Description of N, O, and CPZ group reference sequences

CRF	Sequence	Acc. No.	Year of sampling	Country of sampling (origin)	Reference
N	YBF30	AJ006022	1995	Cameroon	Simon, F. *Nature Medicine* **4**:1032–37 (1998)
N	YBF106	AJ271370	1997	Cameroon	Roques, P. *AIDS* **18**:1371–81 (2004)
N	DJO0131	AY532635	2002	Cameroon	Bodelle, P. *ARHR* **20**:902–8 (2004)
O	MVP5180	L20571	1991	Cameroon	Gurtler, L.G. *J Virol* **68**:1581–85 (1994)
O	ANT70	L20587	1987	Cameroon	Haesevelde, M. *J Virol* **68**:1586–96 (1994)
O	MP1300	AJ302647	1999	Senegal (?Cameroon)	Toure-Kane, C. *ARHR* **17**:1211–6 (2001)
O	CMU2901	AY169812	1998	Cameroon	Yamaguchi, J. *ARHR* **19**:979–88 (2003)
CPZ	GAB	X52154	NA	Gabon	Huet, T. *Nature* **345**:356–59 (1990)
CPZ	ANT	U42720	NA	DRC	Haesevelde, M. *Virol* **221**:346–50 (1996)
CPZ	TAN1	AF447763	2000	Tanzania	Santiago, M.L. *J Virol* **77**:2233–42 (2003)
CPZ	CAM5	AJ271369	1998	Cameroon	Muller-Trutwin, M.C. *J Med Primatol* **29**:166–72 (2000)

The U sequences in the 2001 reference set have been omitted. It is possible that these are representatives of a new subtype, but so are all U sequences. Importantly, the U sequences are not a homogeneous group, but rather a collection of unclassified or at the time of submission unclassifiable sequences.

In previous reference alignments each of the CRFs were described by four representatives. With the large increase of reported CRFs the reference alignment would increase to an extent that it would cause problems in some analyses if four sequences were included for each CRF. Thus, we have limited the CRF section to one sequence per CRF. Except for the E part of CRF01 (and other CRFs that contain subtype E), all subtypes that build up the CRFs are already part of the subtype section in the alignment. The sequence selected for each subtype is now intended to show how it is composed of the included subtypes. If more subtype E sequences are needed in an analysis, one can either refer to the 2001 reference selection or retrieve all E sequences from the HIV sequence database search interface (http://hiv-web.lanl.gov/components/hiv-db/combined_search_s_tree/search.html).

Groups O and N: Group N now has three full genome sequences available, and all are included in the reference set. At this time it is unclear whether group O should be divided into subtypes because only 22 full genomes are available which do not describe the full spectrum of group O diversity that is suggested through analysis of partial genome sequences (15). Four sequences, with one change compared to the 2001 set, are included in the references set.

CPZ sequences are included for outgrouping purposes of HIV clusters, and two from each of the *Pan troglodytes* subspecies *troglodytes* and *schweinfurthii* are included in the alignments. The selection is not meant to be representative for the larger PLV evolution. For that purpose we refer to the complete PLV alignment, which has representatives for all major lineages in the PLV tree. See discussion in the PLV section of the *HIV Sequence Compendium 2003* (8).

Reconstructed phylogenetic trees displaying the subtype divergence

Given enough sequence information, the phylogenetic clades that define HIV-1 groups and subtypes can be reconstructed from any part of the HIV-1 genome. As a rule of thumb, enough sequence information to reconstruct the subtype clades is achieved when the alignment is at least 300-500 characters long. In some regions fewer characters are needed, e.g., *env* V3 region, while other regions under slower evolutionary change, such as *pol* RT, need more characters to give reliable results (5, 6). Also, essentially all phylogenetic reconstruction methods are capable to infer the subtype clades (5). Beyond subtyping, however, for more critical phylogentic analyses of transmission patterns more characters

Figure 1A Non-recombinant sequences

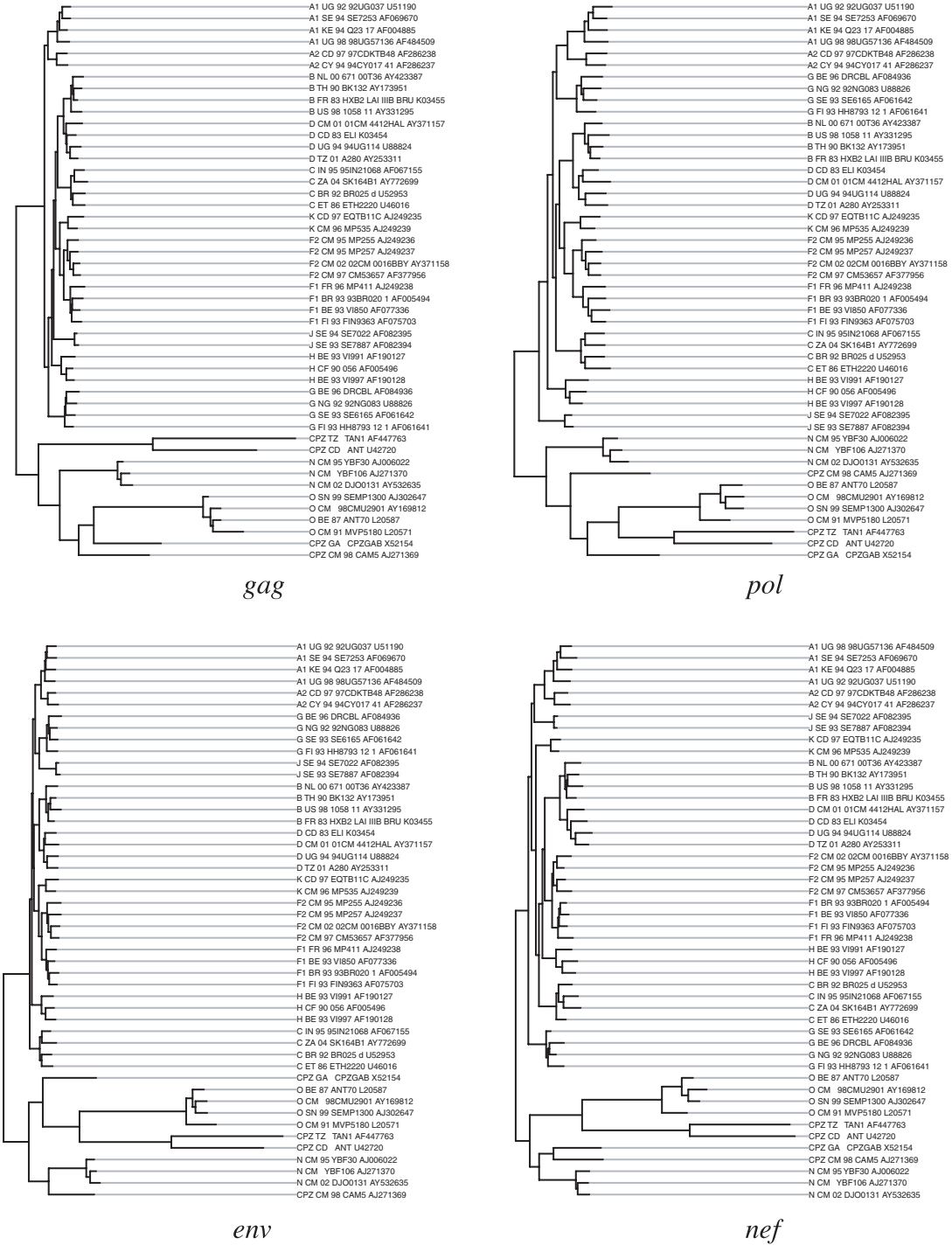
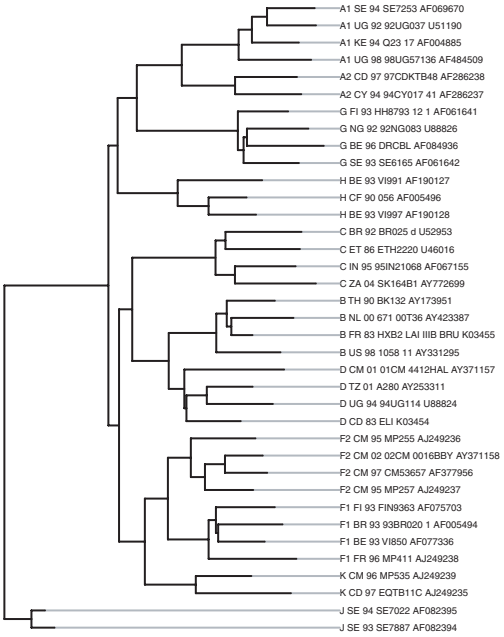
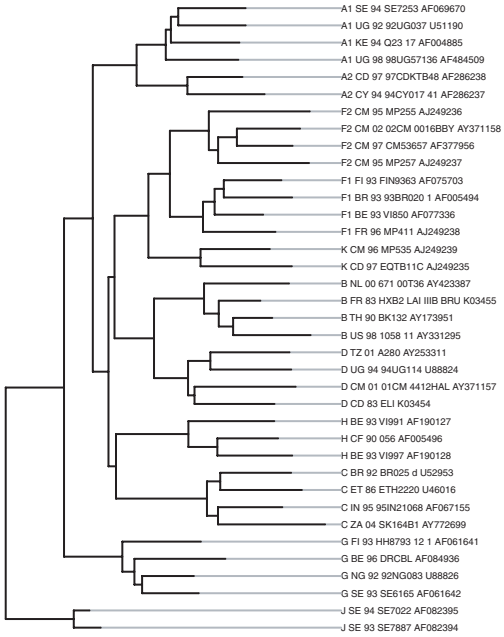


Figure 1. Phylogenetic trees of HIV-1 reference sequences. (A) All non-recombinant sequences. (B) Only group M sequences. The trees shown here are based on non-gapstripped alignments. As with the shorter genes, gapstripped nef alignments did not produce reliable trees. See text for details on how the trees were calculated.

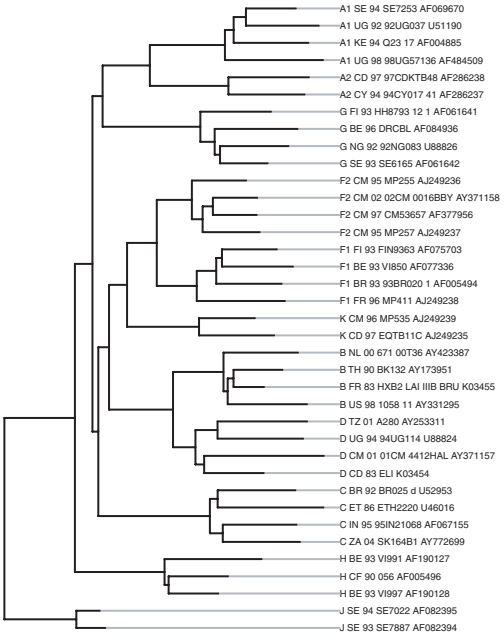
Figure 1B Group M sequences only



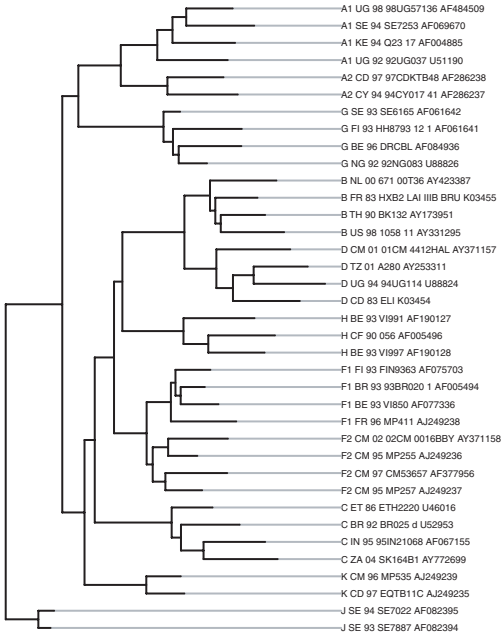
gag



pol



env



nef

than the minimum above and more advanced reconstruction methods such as maximum likelihood should be used (4, 5, 7).

As part of the revision of the reference alignments presented in the 2005 version, many trees were constructed. These trees were created using enhanced and parallelized versions of Gary Olsen's fastDNAml maximum likelihood tree fitting (RevML) and site rate estimation codes (RevRates). This code was written by Tanmoy Bhattacharya of LANL, and fits a general time reversible model (4).

The trees were created as follows: A candidate tree topology was created assuming uniform site rates and an initial random estimate of nucleotide frequencies and transition rates. RevML proceeds in a heuristic and piecewise way, starting from a small set of sequences and building up the tree topology and branch lengths while making placement decisions that maximize the tree likelihood score. The resulting tree then constrains per-site rate optimization of tree likelihood as a function of global estimates of baseline nucleotide frequency and transition rates. These estimates are fit using the conjugate gradient algorithm in the RevRates program. A second RevML run was then performed using these estimates and in turn another rate estimation procedure refined from the second tree. A final tree was estimated using the twice-refined global and local site rates. Each of the trees in the refinement procedure was independently estimated from the global and site local rate parameters.

Trees were inferred from each gene (*env, gag, nef, pol, rev, tat, vif, vpr* and *vpu*) on alignments with all non-recombinant sequences, only group M sequences and all sequences on both globally gap stripped and non-gap stripped data. As expected, alignments with fewer than 400 characters generated trees with some problems. In general, the problems consisted of a lack in resolution among sub-subtypes (mixing of A1 and A2, mixing of F1 and F2 and sometimes K, mixing of B and D). In addition, placement of CPZ sequences was not consistent among short genes. Thus, this reiterates the fact that too short an alignment will not give good tree reconstructions. Trees based on full-length *env, pol* and *gag* showed full subtype and sub-subtype resolution (Figure 1). The subtype classifications were clear whether only group M sequences or all sequences were used. Gap stripping already short alignments made the results even worse, while on long alignments it had no effect on subtype associations.

References

1. Carr, J. K., B. Foley, T. Leitner, M. O. Salminen, B. Korber, and F. McCutchan. 1999. Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic, p. III-10–19. In B. Korber, C. Kuiken, B. Foley, B. Hahn, F. McCutchan, J. Mellors, and J. Sodroski (ed.), *Human Retroviruses and AIDS 1998*. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM.
2. Eddy, S. 1995. *HMMER Hidden Markov Models of Protein and DNA Sequence*, 1.8 ed. Washington University School of Medicine, St. Louis, MO.
3. Harris, M., D. Serwadda, N. Sewankambo, B. Kim, G. Kigozi, N. Kiwanuka, J. Phillips, F. Wabwire, M. Meehen, T. Lutalo, J. Lane, R. Merling, R. Gray, M. Wawer, D. Birx, M. Robb, and F. McCutchan. 2002. Among 46 near full length HIV type 1 genome sequences from Rakai district, Uganda, subtype D and AD recombinants predominate. *AIDS Research and Human Retroviruses* **18**:1281–1290.
4. Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789–1796.
5. Kuiken, C. L., and T. Leitner. 2001. HIV-1 subtyping, p. 27-53. In A. Rodrigues and G. Learn (ed.), *Computational analysis of HIV molecular sequences*. Kluwer Academic Publishers.
6. Leitner, T. 1997. Genetic subtypes of HIV-1, p. III-28–40. 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.

7. Leitner, T., D. Escanilla, C. Franzén, M. Uhlén, and J. Albert. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences USA* **93**:10864–10869.
8. Leitner, T., B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber. 2004. *HIV Sequence Compendium*. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM.
9. Leitner, T., B. Korber, D. Robertson, F. Gao, and B. Hahn. 1998. Updated proposal of reference sequences of HIV-1 genetic subtypes. In B. Korber, B. Foley, T. Leitner, J. W. Mellors, F. McCutchan, B. Hahn, G. Myers, and C. Kuiken (ed.), *Human Retroviruses and AIDS 1997*: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, NM.
10. Meloni, S., B. Kim, J. Sankale, D. Hamel, S. Tovanabutra, S. Mboup, F. McCutchan, and P. Kanki. 2004. Distinct human immunodeficiency virus type 1 subtype A virus corculating in West Africa: sub-subtype A3. *Journal of Virology* **78**:12438–12445.
11. Peeters, M. 2005. Personal communication.
12. Rambaut, A. 1996-2002. *Sequence Alignment (Se-Al) Program*, 2.0a11 ed. Department of Zoology, University of Oxford, Oxford.
13. Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, F. Gao, B. H. Hahn, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, S. Wolinsky, and B. Korber. 2000. HIV-1 nomenclature proposal. *Science* **288**:55.
14. Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, R. K. Funkhouser, F. Gao, B. H. Hahn, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, S. Wolinsky, and B. Korber. 2000. HIV-1 nomenclature proposal: a reference guide to HIV-1 classification. In B. Korber and et al (ed.), *Human Retroviruses and AIDS 1999*: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, NM.
15. Yamaguchi, J., P. Bodelle, L. Kaptue, L. Zekeng, L. Gurtler, S. Devare, and C. Brennan. 2003. Near full-length genomes of 15 HIV type 1 group O isolates. *AIDS Research and Human Retroviruses* **19**:979–988.