Bioinformatics Trends : A Journal of Bioinformatics and its Applications
Volume 1, Issue 1, Jan - Mar 2006.

Clustering of HIV-I Subtype: Study of Molecular Diversity using Phylogenetic Analysis.

*Dipankar S. Gupta, Deeptak Verma, Viplav S. Mishra and Pradeep K. Naik**

Department of Bioinformatics, Jaypee University of Information Technology, Waknaghat, Dumehar Bani, Solan 173 215, Himachal Pradesh.
*E-mail: pknaik73@rediffmail.com

Abstract

AIDS is one of the diseases which attract the attention of scientists across the world as we can't find a country which is trying to prevent its people from this deadly disease. Till now 126 strains of HIV-1 have been reported around the world, which have been classified into groups A-J. The major difference among these strains is their genetic composition. The values of genetic variation were ranged form 8903 to 9720. B-subtype was found to be more diverse (mean 9381.5 ± 396.34) in comparison to others subtypes. Phylogenetic analysis revealed perfect clustering of subtypes. Further wide genomic variation between subtypes is due to more polymorphic sites; varies from 6538 to 8913. On an all genomic bases, the transition mutation was found to be from 2195 to 5161 and the transversion mutation was found from 4343 to 7682.

Keyword: HIV-1, Molecular diversity, Phylogenetic analysis.

Introduction

Now, in 2005, we know from bitter experience that AIDS is caused by the virus HIV, and that it can devastate families, communities and whole continents. It has been seen the epidemic knock decades off countries' national development, widen the gulf between rich and poor nations and push already-stigmatized groups closer to the margins of society. We are living in an 'international' society, and HIV has become the first truly 'international' epidemic, easily crossing oceans and international boarders. HIV displays important genetic variability. Differences between HIV-1 and HIV-2 are fairly well documented in terms of transmissibility, pathogenesis, and pattern of spread (Cock *et al.* 1993, Kanki *et al.* 1994). Our knowledge about the biological characteristics and the epidemic spread of the different HIV-1 strains, however, is still patchy. Genetic variability of HIV-1 has been shown among isolates from different geographic locations and different individuals and within individuals during disease progression (Kostrikis *et al.* 1995, Robertson *et al.* 1995). The known universe of HIV-1 viruses is divided into two groups, a major "M" group and an outlier "O" group (Moore *et al.*1996). Although the genomic organization of the HIV-1 group O is similar to group M, the sequences of isolates vary significantly (Myers *et al.* 1990, Myers *et.al.* 1994, Myers *et al.* 1994). Group M consists of subtypes A to J, of which A to E are presently found in the centers of the pandemic (Leitner *et al.* 1995, McCutchan *et al.* 1992).

The subtypes are very unevenly distributed throughout the world. For instance, subtype B is mostly found in America, Japan, Australia, the Caribbean and Europe; subtypes A and D predominate in sub-Saharan Africa; subtype C in South Africa and India; and subtype E in Central African Republic, Thailand and other countries of Southeast Asia. Subtypes F (Brazil and Romania), G and H (Russia and Central Africa), I (Cyprus), and group O (Cameroon) are of very low prevalence. In Africa, most subtypes are found, although subtype B is less prevalent. The major difference is their genetic composition; biological differences observed *in vitro and/or in vivo* may reflect this. It has also been suggested that certain subtypes may be predominantly associated with specific modes of transmission: for example, subtype B with homosexual contact and intravenous drug use (essentially via blood) and subtypes E and C, with heterosexual transmission (via a mucosal route). Before 1992, HIV-1 strains were classified on the basis of their geographic origin into two subgroups, North American and African variants (Myers *et al.*

1990). Since 1992, the *env* coding sequence has been used to classify globally prevalent viruses. The first five sequence subtypes — A, B, C, D and E— thus identified differed from each other by approximately 30% in their *env* coding sequences and 14% in their *gag* coding sequences. Thereafter, five more subtypes were described and labeled F through J (Myers *et al.* 1994, Janssens *et al.* 1994, Kostrikis *et al.* 1995 and Leitner *et al.* 1995), of which I and J viruses have been reported only once. Moreover, several sequences still await classification. These 10 subtypes constitute the group M (major) HIV-1 viruses. They differ from one another by an average nucleotide percentage distance of 27% (range, 21–31%) (Moore *et al.* 1996). Again, within each subtype, a plethora of HIV-1 variants are expressions of minor intra subtype genomic diversity; their average nucleotide distances are approximately 11%.

So far it has been very difficult to conduct large-scale comparative studies of these subtypes because of the lack of a simple, rapid, and cheap test for the identification of HIV-1 subtypes. Of the two types, HIV-1 and HIV-2, the former type is by far the most widely distributed and also the most studied virus.

The genetic variation of microorganisms and their evolution in time have important implications for the control of infectious diseases. Genetic variation may be reflected in differences in biological characteristics that may determine transmissibility, pathogenesis, and immunogenicity. Genetic variability of microorganisms needs to be taken into account when developing or adapting diagnostic tests and vaccines, and when making projections of the burden of morbidity and mortality. Identification of strains or subtypes has also proved to be an invaluable tool in studying the spread of infectious pathogens. For instance, according to several reports in the recent scientific literature about the use of restriction fragment length polymorphism (RFLP) analysis in investigating outbreaks of tuberculosis, classical epidemiological investigations were found to be too crude or too insensitive.

The objectives of this study were to detect genetic variability and relatedness of 6 subtypes (21 strains) of HIV-1 located from different geographical regions. In this study, many analytical procedures such as neighbor-joining (NJ) phylogenetic tree, bootstrapping, analysis of molecular variance (AMOVA) etc. has been widely used to derive genetic distances among the subtypes and to assess the structure of molecular data in a reduced dimensional space.

Materials and Methods

HIV-1 subtype sequences:

The complete genome sequences of twenty one different strains belonging to different subtypes of HIV-1 were retrieved from the HIV database at http://hiv-web.lanl.gov**.** Table 1 represents the detail information of the subtypes used in this study.

Table 1. Different strains of HIV 1 virus along with their locality and genomic information used in this study.

STRAINS	LOCALITY (COUNTRY)	NCBI-ID	GENOME SIZE(BASES)	%A	%G	%C	%T
A(STRAIN1)	REFERENCE	M62320	9178	35.90	24.06	17.70	22.34
A(STRAIN2)	CYPRUS	AF286237	9060	35.90	24.06	17.70	22.34
A(STRAIN3)	UGANDA	U51190	8999	35.90	24.06	17.70	22.34
A(STRAIN4)	KENYA	AF004885	9160	35.90	24.06	17.70	22.34
B(STRAIN1)	GENEVA	M17451	9128	36.17	24.02	17.58	22.23
B(STRAIN2)	WEAU	U21135	9720	36.17	24.02	17.58	22.23
B(STRAIN3)	BRAZIL	U52953	8959	36.17	24.02	17.58	22.23
B(STRAIN4)	REFERENCE	K03455	9719	36.17	24.02	17.58	22.23
C(STRAIN1)	BOSTWANA	AF110967	9056	26.24	23.92	17.64	22.20
C(STRAIN2)	REFERENCE	U46016	9031	26.24	23.92	17.64	22.20
D(STRAIN1)	BELGIUM	M27323	9143	36.45	23.80	17.65	22.10
D(STRAIN2)	REFERENCE	K03454	9176	36.45	23.80	17.65	22.10
D(STRAIN3)	UGANDA	U88824	8952	36.45	23.80	17.65	22.10
D(STRAIN4)	ZAIRE	U88822	8975	36.45	23.80	17.65	22.10
F(STRAIN1)	BRAZIL	AF00549	8968	35.97	24.00	17.95	22.03
F(STRAIN2)	CONGO	AF077336	8903	35.97	24.00	17.95	22.03
F(STRAIN3)	FINLAND	AF075703	8925	35.97	24.00	17.95	22.03
G(STRAIN1)	CONGO	AF084936	9707	36.27	24.08	17.76	21.89
G(STRAIN2)	FINLAND	AF061641	9047	36.27	24.08	17.76	21.89
G(STRAIN3)	NIGERIA	U88826	8987	36.27	24.08	17.76	21.89
G(STRAIN4)	SWEDEN	AF061642	9074	36.27	24.08	17.76	21.89

Multiple Alignments of the Sequences & Phylogenetic Analysis:

Retrieved sequences were aligned using Clustal X (version 1.81), multiple alignment program using Clustal W (version 1.6) DNA weight matrix and multiple parameters like gap opening 10.0, gap extension 0.20, transition weight 0.50. A phylogenetic tree based on similarity coefficients generated by neighborhood-joining method was performed. The support for clusters

was evaluated by bootstrapping analysis with 1000 permutation time (Felsenstein 1985). The phylogenetic tree was viewed in tree view.

Analysis of Molecular Variance:

The genomic data were used for analysis of molecular variance (AMOVA) to statistically clarify patterns and degree of relatedness, revealed by n-j tree. The analysis was performed using ARLEQUIN 2.00 (Excoffier *et al.* 1992 and Schneider *et al.* 2001). Significant values for the covariance components (between and within subtypes) and the fixation index Φ_{ST} were calculated using 1023 permutations. A measure of genetic distance between subtypes was calculated for all pair wise comparisons. Gene and nucleotide diversity, transition and transversion mutation, number of polymorphic sites, and the molecular diversity indices like $\theta(S)$ and $\theta(\pi)$ between different subtypes of HIV I were calculated using ARLEQUIN software.

Results

Genomic identity and diversity analysis:

A relatively high genomic variation was detected among the 21 subtypes of HIV I (Table 1). The values of genomic variation ranged form 8903 to 9720. However, the B-subtype was found to be more diverse (avg. mean=9381.5, s.d.=396.34), then subtype-G(x=9203.75, s.d.=337.46), subtype-D(avg. mean=9061..5, s.d.=114.35), subtype-A(avg. mean =9099.25, s.d.=73.29), subtype-F(avg. mean =8932, s.d.=33.06) and subtype-C(avg. mean =9043.5, s.d.=17.68). The base composition A was found to be more diverse (26% to 36%) among 21-subtypes, in comparison to G, C and T.

A pairwise difference between subtypes (Table 2) was ranged from 5985 to maximum value of 6943. The pair wise difference was found to be in order of subtype B(6744 to 5985), A(6389 to 6678), D(6708 to 6426), G(6633 to 6461) and F(6545 to 6374).

Table 2: Pair wise differences between HIV-I subtypes

6592	6597	6567	6630	6424	6545	6543	6557	6550	6699	6656	6577	6681	6731	6524	6706	6608	6389	6593	0	**A1**
6628	6357	6590	6618	6550	6572	6553	6518	6500	6644	6738	6666	6611	6634	6559	6640	6649	6678	6626	0		**A2**
6479	6594	6943	6670	6502	6380	6546	6575	6372	6710	6541	6561	6555	6688	6491	6586	6613	6649	0			**A3**
6653	6485	6585	6734	6408	6563	6636	6539	6556	6712	6668	6553	6618	6796	6548	6744	6645	0				**A4**
6584	6600	6583	6759	6502	6412	6540	6134	6450	6627	6565	6556	6601	6721	6578	6744	0					**B1**
6623	6464	6631	6706	6501	6530	6603	6480	6592	6637	6680	6673	6652	5985	6552	0						**B2**
6550	6540	6595	6538	6435	6397	6456	6487	6616	6616	6624	6578	6580	6600	0							**B3**
6555	6529	6542	6690	6585	6534	6539	6503	6530	6690	6603	6602	6641	0								**B4**
6594	6593	6660	6617	6525	6469	6522	6628	6498	6661	6544	6534	0									**C1**
6585	6518	6132	6616	6427	6484	6491	6505	6565	6621	6601	0										**C2**
6573	6597	6634	6659	6473	6458	6490	6541	6541	6708	0											**D1**
6617	6466	6680	6790	6563	6528	6530	6426	6577	0												**D2**
6510	6497	6588	6581	6515	6443	6529	6552	0													**D3**
6484	6478	6526	6446	6534	6521	6542	0														**D4**
6544	6577	6545	6701	6419	6374	0															**F1**
6629	6503	6537	6695	6545	0																**F2**
6511	6464	6610	6501	0																	**F3**
6633	6488	6534	0																		**G1**
6461	6534	0																			**G2**
6611	0																				**G3**
0																					**G4**
G4	G3	G2	G1	F3	F2	F1	D4	D3	D2	D1	C2	C1	B4	B3	B2	B1	A4	A3	A2	A1	

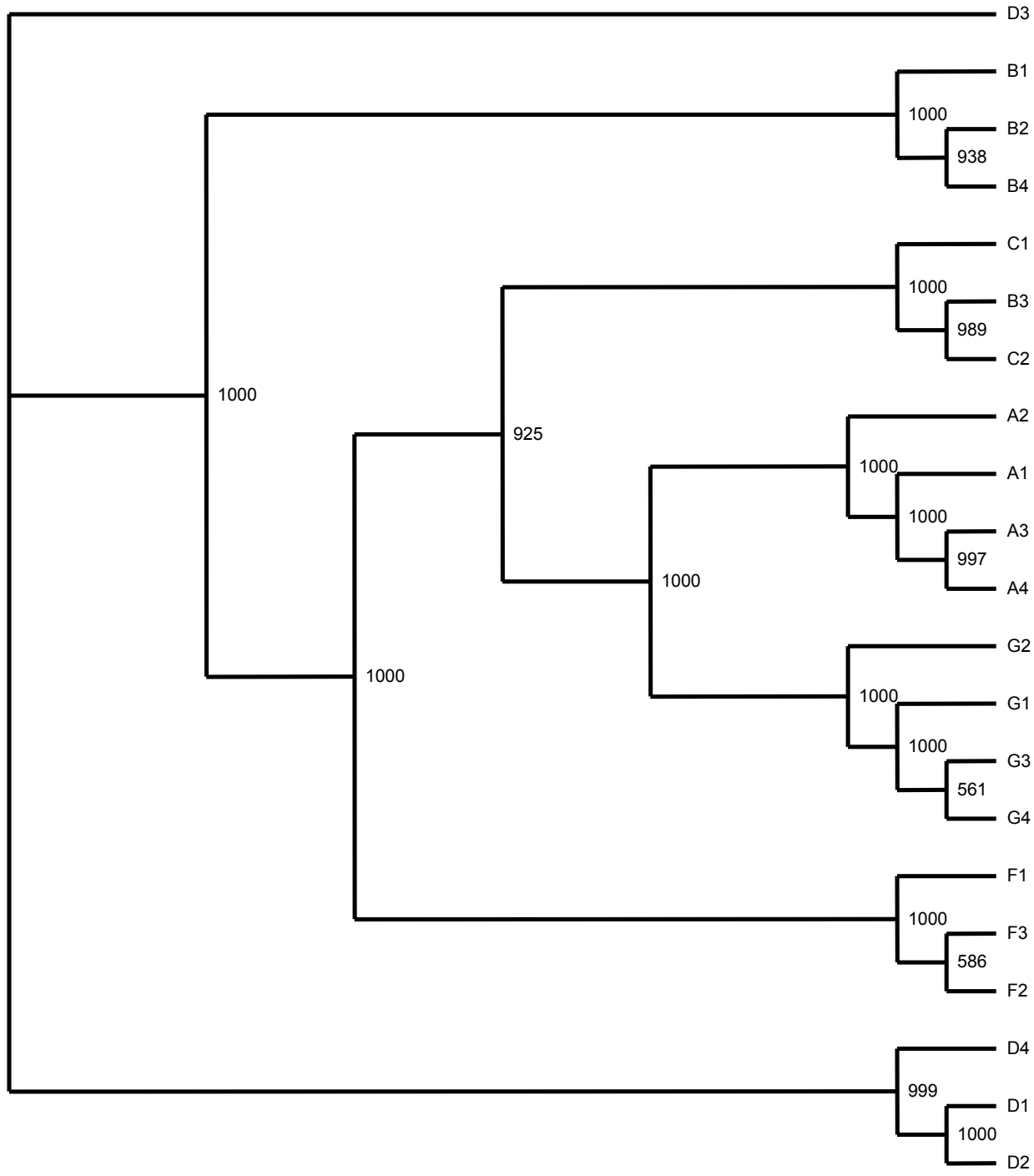

Fig: 1 Phylogenetic pattern among HIV-1 subtypes using n-j method. The value in the clades indicates bootstrap value.

Phylogenetic Analysis

The phylogenetic tree obtained using nj-clustering (Figure 1) showed two main clusters (M and N). The cluster N had five sub-clusters (N_1, N_2, N_3, N_4 and N_5), having 3, 3, 4, 4 and 3 subtypes. All the subtypes of subtype-D formed a single cluster M and similar results were obtained in case of F-subtype (N), G-subtype (N), A-subtype (N) and C-subtype (N). It was found that one subtype (N) of B-subtype virus to be included in C-subtype cluster. Almost all the clusters and sub-clusters were supported with more than 500 bootstrap values.

Genomic Differentiation across the Subtypes

Wide genomic variation between subtypes of the HIV I virus was evident from large differences in number of polymorphic sites, number of transition and transversion mutation (Table 3). Number of polymorphic sites varies form 6538 (subtype-C) to 8913 (subtype-G). On an all genomic bases, the transition mutation was found to be from 2195(subtype-C) to 5161(subtype-A) and the transversion mutation was from 4343(subtype-C) to 7682(subtype-A). Similarly gene diversity values were recorded as 1.00+/-0.1768 to 1.00+/-0.5. Nucleotide diversity was recorded in between 0.713+/-0.465 to 0.721+/-0.721. Among the different subtypes, subtype-C was found to be highly diverse with high standard deviation for both gene and nucleotide diversity. The same order of genomic heterogeneity was discerned through the molecular diversity indices $\theta(S)$, which varies from 4851.272+/-2604.891 to 6538+/-4623.417 and $\theta(\pi)$ which varies from 6446+/-4805.012 to 6578.166+/-4293.534 . Further, the mean number of pair wise difference across the subtypes found to be in between 6446+/-3582.58 to 6578.166+/-3595.559, presents the estimation of genomic variability among the subtypes. AMOVA analysis using ARLEQUIN program enabled a partitioning of the overall genomic variation between subtypes covariance components. AMOVA analysis did not reveal any significant difference between subtypes (0.47%, d.o.f. =5); all of the diversity (99.53%, d.o.f. =15) was attributable to variation within the subtypes.

Table 3 Genomic variation between different strains of HIV-1 subtypes.

Subtypes	No. of polymorphic sites	Gene Diversity	No. of polymorphic Sites	No. of observed sites with transitions	No. of observed sites with tranversions	Mean no. of pairwise difference	Nucleotide Diversity	Theta(s) (+/-S.D.)	Theta(pi) (+/-S.D.)
A	8912	1.00+/-0.1768	8912	5161	7682	6578.166 +/-3595.559	0.716 +/-0.467	4861.090 +/-2610.162	6578.166 +/-4293.534
B	8894	1.00+/-0.1768	8894	5004	7549	6520.666 +/-3564.133	0.714 +/-0.466	4851.272 +/-2604.890	6520.666 +/-4256.008
C	6538	1.00+/-0.5	6538	2195	4343	6534.000 +/-4620.589	0.721 +/-0.721	6538.000 +/-4623.417	6534.000 +/-6534.500
D	8901	1.00+/-0.1768	8901	5129	7653	6555.000 +/-3582.897	0.717 +/-0.467	4855.090 +/-2606.940	6555.000 +/-4278.415
F	8180	1.00+/-0.2722	8180	3931	6315	6446.000 +/-3582.580	0.718 +/-0.535	5453.333 +/-3259.3486	6446.000 +/-4805.012
G	8913	1.00+/-0.1768	8913	5069	7667	6543.500 +/-3576.613	0.713 +/-0.465	4861.636 +/-2610.455	6543.500 +/-4270.910

Discussion

Human Immunodeficiency virus type I (HIV I) has been recognized as one of the causative agents of AIDS. Genetic variability of HIV I have been shown among isolates from different geographic locations and different individuals during disease progression. HIV I group O is similar to group M, the sequences of isolates vary significantly. Group M consists of subtypes A to E are presently found in the centers of the pandemic.

The subtypes A, B C, D, F and G taken into consideration in this study were clearly defined than others subtypes. Each subtype taken for study contains at least one full length, apparently non-recombinant (there are no identified conflicting subtype associations in different regions of the sequence) genome available as a reference sequences as well as multiple additional full length *env* and *gag* sequences. The study reveals that the HIV-1 subtype B is widely diverse then that of subtype G, D, A, F and C. Phylogenetic analysis of full length genomes among the subtypes also reveals the same result. One of the subgroup of subtype B is clustered into subtype – c cluster. This may be due to the process of recombination between B and C subtype.

The observed high proportion of polymorphic sites, number of observations and number of transversion mutation suggest that there is a profound intra-subtype variation existing among the HIV I viruses. AMOVA analysis also collaborate this finding. However the variation between the subtypes is not significant (0.47 %). This result is supported by Janssens *et al.* (1997), Robertson *et al* (1995). Further the higher level of theta(s) and theta (pi) observed in this study probably were associated with the process of recombination (exchange of genetic material between the subtypes). This study indicates that phylogenetic and diversity analysis provides a more reliable method to address the puzzle existing in the classification of HIV I viruses.

Conclusion

There are many questions on the biological character of different HIV I subtypes that remain to be answered. Differences in biological character could have consequences for the spread of HIV I, the detection of infection, the treatment of infected persons and the development of a vaccine. More work needs to be done on documenting and monitoring distribution patterns of HIV I subtypes.

References

De Cock, K.M., Adjorlolo, G,, and Ekpini, E. 1993.: Epidemiology and transmission of HIV-2: why there is no HIV-2 pandemic. *JAMA*. **270**: 2083-2086.

Excoffier, L., Smouse, P.E. and Quattro, J. M. 1992: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. **131**: 479-491.

Felsenstein, J. 1985: Confidence limits on phylogenies: an approach using the boot strap. *Evolution* **39:** 783-791.

Janssens, W., Heyndrickx, L., and Fransen, K. 1994.: Genetic and phylogenetic analysis of env subtypes G and H in central Africa. AIDS Res Hum Retrovirus. **10**: 277-279.

Kanki, P.J. and De Cock, K.M. 1994: Epidemiology and natural history of HIV-2. AIDS. **8** (suppl 1):S85-S93.

Kostrikis, L.G., Bagdades, E., Cao, Y., Zhang, L. and Dimitrious, D. H. D.D 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.

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

McCutchan, F.E., Hegerich, P.A. and Grennan, T.P. 1992.: Genetic variation of HIV-1 in Thailand. *AIDS Res Hum Retrovirus*. **8**:1887-1895.

Moore, J.P., Cao, Y., Leu, J., Qin, L. and Korber, B. H.D. 1996: Inter and intraclade neutralization of human immunodeficiency virus type I,: genetic clades do not correspond to neutralization serotypes but partially correspond to gp120 antigenic serotypes. *J. Virol*, **8**:427-444.

Myers, G., Rabson, R,B,, Berzofsky, J.A., Smith, R.F. and Wong-Staal, F. (Eds) 1990: Human Retroviruses and AIDS 1990. Los Alamos National Laboratory.

Myers, G., Korber, B., Berzofsky, J.A., Smith, R.F. and Pavlakis, G.N. (Eds) 1992: Human Retroviruses and AIDS 1992. Los Alamos National Laboratory.

Myers, G., Korber, B., Wain-Hobson, S., Jeang, K.T., Henderson, L. and Pavlakis, G.N. (Eds) 1994: Human Retroviruses and AIDS 1994. Los Alamos National Laboratory.

Myers, G. 1994: Tenth anniversary perspective on AIDS. HIV: between past and future. *AIDS Res Hum Retrovirus*. *10*:1317-1324.

Robertson. D.L., Sharp, P.M., McCutchan, F.E. and Hahn, B.H. 1995: Recombination in HIV-1. *Nature*. **374**:805-810.

Schneider. S., Roessli. D. and Excoffier. L. 2001: Arlequin: Software for population geneticsdata analysis, Version 2001. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva.

