ALIGNMENT OF PRIMATE LENTIVIRAL GENOMES

This alignment contains a sampling of HIV-1 genomes (one or two per subtype from the HIV-1 M group, one from the N group and one from the O group), at least one of each type of SIV that has been sequenced (Chimpanzee; Vervet, Grivet, Sabaeus, and Tantalus subspecies of African green monkeys; Sooty mangabeys and macaques infected with Sooty mangabey virus; L'hoest monkey; Mandril) and two sequences of HIV-2 subtypes A and B which are closely related to the Sooty mangabey viruses. Table 3 lists the sequence name, database accession number, country of isolation, first author and publication for each of the sequences in this alignment.

Together, these genomes represent the full breadth of diversity discovered to date in the primate lentivirus lineage. Non-primate lentiviruses such as equine infectious anemia virus (EIAV) and caprine arthritis/encephalitis virus (CAEV) are more distantly related to each other and to the primate lentiviruses, and are not included in this alignment.

The alignment is not guaranteed to be "optimal" by any objective score of sequence identity. It was generated with the help of HMMER models, using iterations of 1) computer alignment; 2) adjust alignment by hand; 3) use adjusted alignment to build new model; repeat steps 1–3. Both the alignment and the HMMER model generated in the last iteration are available on our WWW site. The alignment is not codon-aligned, gaps are often found in the middle of codons, and gaps are not always in multiples of 3 bases. The LTR region was particularly difficult to align, and the 5′ and 3′ LTRs have gaps in different places to illustrate alternative alignments of this region. In the LTRs the NF-κ-B binding sites, the polyadenylation signal and other highly conserved regions were "anchored" and then an attempt was made to align the regions in between. It is of particular interest that the structures of some elements such as the TAR element in the LTR are highly conserved in all primate lentiviruses, while the primary sequences are highly divergent. The fact that the tip of the TAR loop and sequence immediately 3′ from the base of the TAR stem is conserved is also of note.

The alignment is annotated, based on known protein coding regions in HIV-1 and also based on annotations found in SIV sequence database entries. The protein cleavage sites that create Gag p17, Gag p24 and other mature peptides from the Gag and Gag-Pol precursor polyproteins have been experimentally determined for HIV-1 and at least one strain of HIV-2, the study of analogous cleavages in SIV polyproteins have not been published. Three genomes have been translated in all three reading frames; HIV-1 subtype B strain HXB2; HIV-2 subtype A strain BEN; and SIV from a L'hoest monkey. The translations are provided as a visual aid for finding landmarks in the genomes. Because of the way the sequences were output with "-" representing identity to the HIV-1 HXB2 sequence, some of the Cys "C" Thr "T" and Ala "A" amino acid residues for the HIV-2 and L'hoest translations may have become mistranslated as Gly "G". Correct translations are available on the WWW site in other formats.

The HIV-2/SIV-SMM vpx gene is postulated to be a duplication of the vpr gene (Tristem et al. *Nature* **347**:341–342 (1990)) and thus there may be two alternative alignments of this region of the genome, as there are for the duplicated stem-loops of the TAR element.