JOURNAL OF MOLECULAR EVOLUTION

Fast Analysis of Genomic Homologies: Primate Immunodeficiency Virus

Maurice L.J. Moncany,[1] Pascal R.R. Courtois[1,2]

[1] Laboratory of Cellular and Molecular Biology, University of La Rochelle, Avenue Marillac, 17042, La Rochelle Cedex 1, France
[2] Computer Science Laboratory, Conservatoire National des Arts et Métiers (CNAM), 292 rue Saint-Martin, 75003, Paris, France

Received: 28 September 1995 / Accepted: 26 January 1996

Abstract. We have recently published a new probabilistic algorithm which performs genomic comparisons on a huge scale. In the present paper it was applied to immunodeficiency viral sequences extracted from international gene databanks. During global sequence analysis of human (HIV1 and HIV2) and simian viruses by means of dot-matrix representation, series of homology were obtained which permitted the definition of families of viruses overlapping the species divisions. Sequences of interest were characterized to the lexical base sentence through successive zoomings. Strain-to-strain comparison confirmed subfamily classifications and led, for example, to the identification of divergent LTR sequences. By way of example, we described the application of the algorithm to the ANT70C and MVP5180 HIV1-O viruses, for which the observed differences were shown to correspond to a deletion in the U3 region, situated between the LEF and NF-κB sites. It was of interest to consider these data in a tentative phylogenetic interpretation.

Key words: Primate immunodeficiency virus — Genomic analysis — Genome comparison — HIV1 — HIV2 — SIV — Probabilistic algorithm — Phylogenetic analysis

Introduction

Epidemiological studies revealed extensive divergence between the etiological agents of AIDS HIV1 and HIV2. For example, a subtyping was necessary to distinguish the HIV1s ANT70C (De Leys et al. 1990; Vanden Haesvelde et al. 1994), MVP5180 (Vanden Haesvelde et al. 1994) and VAU (Charneau et al. 1994), which were highly divergent from other HIV1s. More surprisingly, it has been observed that the chimpanzee virus (SIV_{CPZ}) presented genomic sequences which were more closely related to the HIV1s than to the other simian viruses (Vanden Haesvelde et al. 1994; Peeters et al. 1989, 1992). Series of similar data, as reviewed by Myers et al. (1993), have kindled the conception of new viral studies based on the screening of viral genomes to determine the domain(s) which could be of interest in prospective phylogenetic studies. However, the evolution of the HIVs is now generally considered in the perspective of recombination events (Robertson et al. 1995a). This concept of "mosaic virus" made evident that gene-to-gene and domain-to-domain comparisons lead to multiple phylogenetic trees as reported by Gao et al. (1994) for HIV2, Jin et al. (1994) for viruses from West African green monkeys, and Robertson et al. (1995b) for HIV1. The problem of alternative phylogenetic positions justified global genomic analyses.

Genomic comparison is a technique which is now widely used for the detection of homologies between different viral strains. Graphical representations of sequence similarity have been in use for a long time (Fitch 1969; Gibbs and McIntyre 1970; McLachlan 1971; Maizel and Lenk 1981; Steinmetz et al. 1981; Novotny 1982; Harr et al. 1982; Staden 1982; Schwartz et al. 1991; Lefèvre and Ikeda 1994). They allow a fast "intuitive" comprehension even when integrating an increasing amount of computing data. The commonly chosen solution is a dot-matrix representation which tends to form

Correspondence to: M.L.J. Moncany

diagonals when a high degree of similarity on long strands is detected. However, these methods are not only time and memory consuming but also limited in their range of applications, particularly when the treated sequences exceed 50–150 kb, depending on the algorithm.

These problems have led us to design a probabilistic algorithm designated PAGEC which exponentially decreases the computing time (Courtois and Moncany 1995). This has been verified by testing comparisons of up to 4 Mb/4 Mb, which were achieved in 13 min. Furthermore, PAGEC is fast enough to permit an interactive use as it compares genomes as large as 325 kb in 20 s with a single index (Courtois and Moncany 1995). In this paper, PAGEC was applied to the immunodeficiency viruses. The homology domains of the HIV1s, the HIV2s, and the SIVs were rapidly determined by self-comparison of the computer-concatemerized genomes. They were represented by dot-matrix figures which were zoomed to obtain the distinctive lexical sequence. As an example, it appeared that the $\mathrm{HIV1}_{\mathrm{ANT70C}}$ and $\mathrm{HIV1}_{\mathrm{MVP5180}}$ diverged to a high degree from the HIV1 group although they were described as both belonging to the HIV1-O subgroup (Vanden Haesvelde et al. 1994; Gürtler et al. 1994). We made similar striking observations for the SIV group. In our studies, these differences were strongly emphasized when individual divergent strains were compared to members of their immunological group. Cross comparison between the different immunodeficiency virus groups revealed high intragroup divergences and surprising intergroup homology. From these rapid screenings, we confirmed the current definition of families of viral strains which were then considered in greater detail in a phylogenetic interpretation.

Materials and Methods

Probabilistic Algorithm. The PAGEC probabilistic algorithm based on the creation of index tables and large key determination was tested on large genomes as previously described (Courtois and Moncany 1995). The comparison utilizes four parameters which can vary as a function of the length of the compared genomes and the homology rate. These parameters are: the number of indices, the index or key size, the maximum number of mismatches (or transitions), and the sequence length. The two first parameters affect the quality of the data defined by the difference between the theoretical vs the practical results. They were estimated by mathematical calculations and determined by experimentations as reported in the article in which we detailed PAGEC (Courtois and Moncany 1995). In this work, the four parameters were set at 3, 8, 3 and 25. The graphic representation is a dot-matrix picture in which one dot is printed whenever a successful comparison occurs. This happens when the number of detected random mismatches is inferior to the number of mismatches allowed in a sequence of the previously fixed length. Under our conditions, a point is drawn when less than three randomly situated mismatches are found in a sequence of 25 nucleotides.

PAGEC was implemented on an Ultrix system with an interactive system which allows successive zoomings to obtain the base sequence which is considered of interest. The program was written in C programming language.

Genomes. The complete viral genomes selected for analysis were extracted from the Human Retroviruses and AIDS 1993 data bank (Myers et al. 1993) at the Los Alamos Laboratory. The studied viruses were composed of 21 HIV1s, 10 HIV2s, and 17 SIVs as listed below:

Human viruses

- *HIV1:* BH102, CAM1, ELI, H3BH5, HXB2, JRCSF, LAI, MAL, MN, NDK, NL43, NY5, OYI, PV22, RF, SF2, SG3X, U455A, Z2Z6, ANT70C, MVP5180
- *HIV2:* BEN, CAM2, D194, D205, GH1, NIHZ, ROD, SBL/ISY, ST, UC1

Simian viruses

- *Green Monkeys:* AGM155, AGM3, AGM677A, AGMAA, AGMTYO, AXX
- *Sykes' Monkey:* SYK (COM)
- *Macaque:* MM142, MM239, MM251, MNE, STM
- *Sooty Mangabey:* MMPBJA
- *Mandrill:* MNDGB1
- *Chimpanzee:* CPZ (CIV)
- *D particle forming SIV:* MP, RV1

The last $\mathrm{SIV}_{\mathrm{MP}}$ (Sonigo et al. 1986) and $\mathrm{SIV}_{\mathrm{RV1}}$ (Power et al. 1986) sequences were extracted from the Genebank database.

In order to perform global analyses, the viral genomes were artificially queued by family. These concatemerized genomes, named "allhiv1," "allhiv2," and "allsiv," were assembled following the order indicated above.

Results and Discussion

Self-Comparison of the Viral Groups

The graphic representation of the self-comparison of the artificially concatemerized HIV1s (designated "allhiv1" in the Figs.) displayed a corner-to-corner diagonal line resulting from the self-comparison of each viral strain (Fig. 1A). Continuous lines, which reflected perfect homology, could not be drawn when different strains were compared: Some gaps corresponded to nonhomology zones (area A1), while others were due to missing LTR sequences (area A2). On the other hand, in area A3 we observed parallel lines which corresponded to repeated sequences. A single zooming confirmed that they corresponded to LTR sequences (data not shown). More striking observations were made for the ANT70C (lines L1) and MVP5180 (lines L2) viruses. The weaker dotted lines indicated high divergence between these two subtyped HIV1-Os and the other members of the HIV1 group, as has been proposed previously (Vanden Haesvelde et al. 1994; Gürtler et al. 1994). Dispersed dots corresponded to point homologies between the different viral strains.

The representation (Fig. 1B) of the self-computation of the concatemerized HIV2s (designated "allhiv2" in the figures) is similar to that for HIV1s (Fig. 1A). Of particular interest were the $\mathrm{HIV2}_{\mathrm{D205}}$ and $\mathrm{HIV2}_{\mathrm{UC1}}$ strains previously described as being divergent (Gao et al. 1992; Kreutz et al. 1992), a result which was corroborated on lines L1 and L2 of Fig. 1B, respectively.

The representation of the self-comparison of the arti-

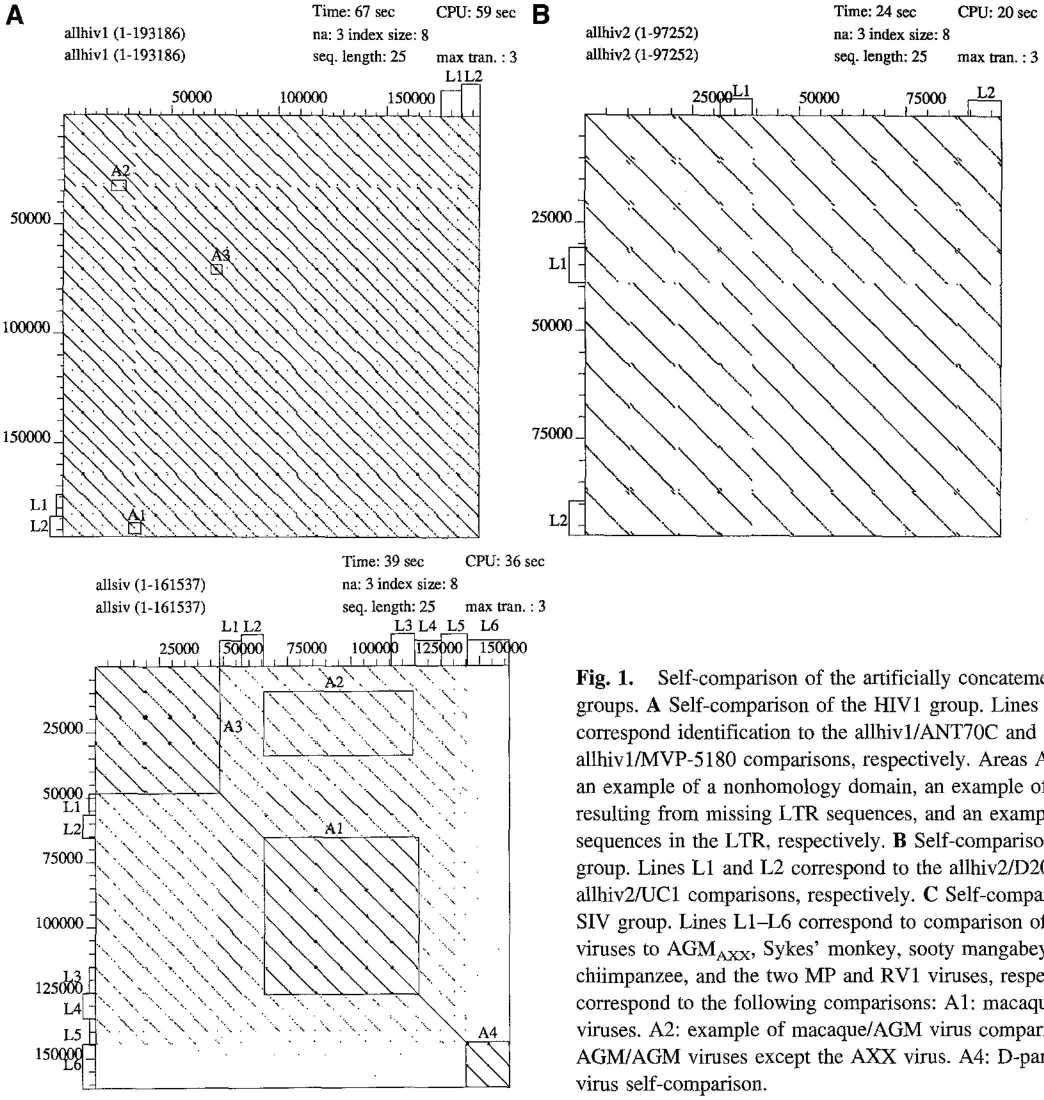

Fig. 1. Self-comparison of the artificially concatemerized viral groups. **A** Self-comparison of the HIV1 group. Lines L1 and L2 correspond identification to the allhiv1/ANT70C and allhiv1/MVP-5180 comparisons, respectively. Areas A1–A3 display an example of a nonhomology domain, an example of a gap resulting from missing LTR sequences, and an example of repeated sequences in the LTR, respectively. **B** Self-comparison of the HIV2s group. Lines L1 and L2 correspond to the allhiv2/D205 and allhiv2/UC1 comparisons, respectively. **C** Self-comparison of the SIV group. Lines L1–L6 correspond to comparison of all the SIV viruses to AGM_{AXX}, Sykes' monkey, sooty mangabey, mandrill, chiimpanzee, and the two MP and RV1 viruses, respectively. Areas correspond to the following comparisons: A1: macaque/macaque viruses. A2: example of macaque/AGM virus comparison. A3: AGM/AGM viruses except the AXX virus. A4: D-particle-forming virus self-comparison.

ficially concatemerized SIV genomes (designated ''allsiv'' in the figures) was particularly informative. In Fig. 1C, area A1 clearly displayed strong homologies between macaque immunodeficiency viruses which were compared in this zone. Area A2, corresponding to the African green monkey virus vs macaque virus comparison, showed that these two subgroups were composed of highly divergent strains. However, one should note the distinctive behavior of the AXX African green monkey virus (Franchini et al. 1987), which behaved like a macaque virus (line L1). Area A3 revealed an intermediate situation with a variable level of homology when African green monkey viruses were compared. The L2 lines indicated the Sykes' virus, which has weak similarity to other viruses (Hirsch et al. 1993). The $SIV_{SMMPBJA}$ isolated from the sooty mangabey (Dewhurst et al. 1990) behaved like macaque viruses, as indicated on lines L3. The L4 lines corresponded to the MNDGB viral genome from the mandrill monkey, which highly diverged from all other simian viruses. MNDGB is known to be a subfamily generator (Tsujimoto et al. 1989). A similar observation was made on lines L5, corresponding to the infected chimpanzee CPZ (Peeters et al. 1989, 1992). A striking divergency was apparent on lines L6, where no homology was noticed outside area A4, in which the homology lines of two strains appeared. The central diagonal line demonstrated perfect homology resulting from self-comparison. The lateral lines indicated the very good homology between the two strains SIV_{MP} and SIV_{RV1} (Sonigo et al. 1986; Power et al. 1986). They were included as controls as they are known to correspond to D-particle-forming SIVs, which never induce a lethal immunodeficiency (Sonigo et al. 1986; Power et al. 1986). Although belonging to the SIV immunological group, they represent an important genomic divergency which could be correlated to their particular etiology.

Cross-Comparison Between the Three Viral Groups

The HIV1/HIV2 cross-comparison (Fig. 2A) showed an equal distribution of the homology domains throughout

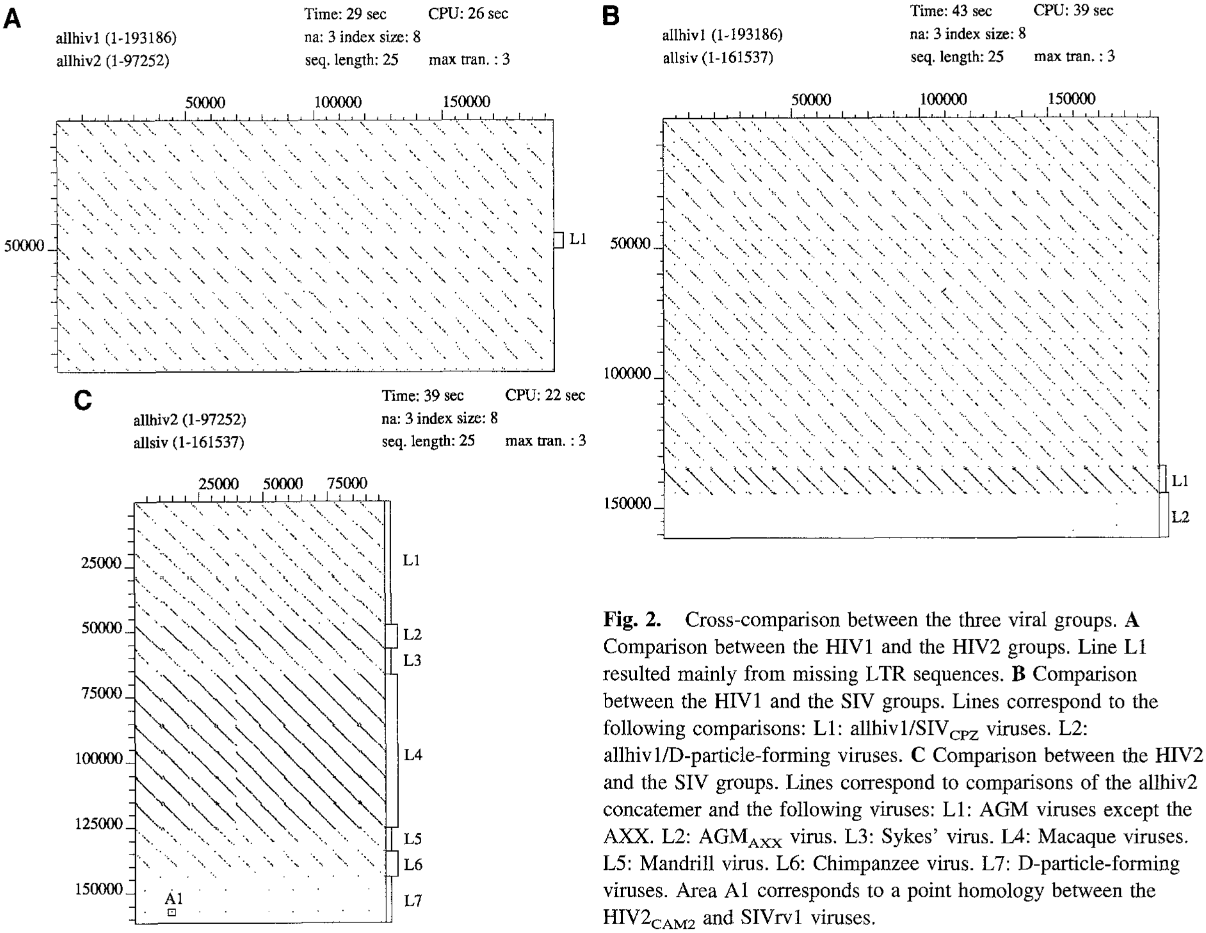

Fig. 2. Cross-comparison between the three viral groups. **A** Comparison between the HIV1 and the HIV2 groups. Line L1 resulted mainly from missing LTR sequences. **B** Comparison between the HIV1 and the SIV groups. Lines correspond to the following comparisons: L1: allhiv1/SIV_{CPZ} viruses. L2: allhiv1/D-particle-forming viruses. **C** Comparison between the HIV2 and the SIV groups. Lines correspond to comparisons of the allhiv2 concatemer and the following viruses: L1: AGM viruses except the AXX. L2: AGM_{AXX} virus. L3: Sykes' virus. L4: Macaque viruses. L5: Mandrill virus. L6: Chimpanzee virus. L7: D-particle-forming viruses. Area A1 corresponds to a point homology between the $HIV2_{CAM2}$ and SIVrv1 viruses.

the viral genomes. Even for the highly divergent HIV1s ANT70C and MVP5180, no important differences were evident. This rapid survey was sufficient to reveal empty lines (L1, as an example), which mainly corresponded to missing LTR sequences in the files retrieved from the databases. Similar observations could be made for the HIV1/SIV comparison (Fig. 2B) with, in addition, an empty zone corresponding to the highly different D-particle-forming SIVs (line L2). These viruses diverge entirely from both other SIVs and the HIV1s. Line L1, which resulted from comparing the simian chimpanzee immunodeficient virus (CPZ) to all the human HIV1s, clearly showed a closer relationship between the SIV_{CPZ} and the HIV1s than between all the other SIVs and the HIV1s. The juxtaposition of line L4 in Fig. 1C and line L1 in Fig. 2B revealed that SIV_{CPZ} was more closely related to the HIV1s than the SIVs. It is noteworthy that SIV_{CPZ} presented a similar degree of homology with divergent HIV1s ANT70C and MVP5180 and with the other HIV1 genomes.

The cross-study of HIV2s and SIVs (Fig. 2C) was highly informative when compared to Fig. 1C. The upper horizontal zone (Fig. 2C, line L1) indicated a high rate of divergence between the African green monkey viruses and HIV2s while the macaque and the sooty mangabey viruses were more closely related to HIV2s (line L4). The close relationship between the SIV_{AXX} African green monkey virus and macaque viruses (line L2) and the distinctiveness of the Sykes' virus (line L3) confirmed the results obtained in Fig. 1C. An unusual observation was made for the wild-caught African mandrill virus (MNDGB), which behaved like African green monkey viruses (line L5) and presented a similar degree of divergence when compared to HIV1s, HIV2s, and SIVs. This reinforced the putative status of the MNDGB virus, which has been proposed as a subfamily ancestor (Tsujimoto et al. 1989). In contrast to the observation made while cross-studying SIV_{CPZ} and HIV1s, the chimpanzee virus diverged greatly from the HIV2s (line L6). The poor homology between the D-particle-forming viruses and HIV2s is shown on line L7. For instance, the zooming of a dot (area A1 in Fig. 2C) corresponding to a point homology between these two families of viruses allowed the determination of related base sequences. The TATCATTCAGTACATGGATGATATCTTAATAGCTAGTGACAGGACAGACT and the TATTATACATTACATGGATGATATCCTCATAGCTGGTAAAGATGGACAAC sequences are situated at positions 14,029 and 157,146, respectively, in the concatenated genomes, corresponding to the $HIV2_{CAM2}$ (Tristem et al. 1991) and SIV_{RV1} viruses (data not shown).

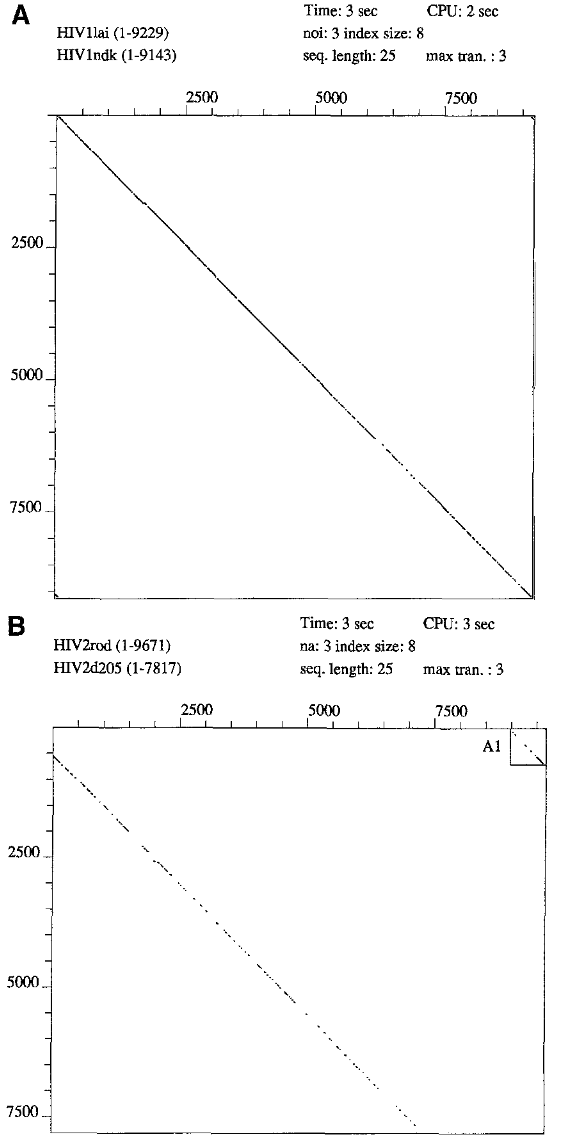

Fig. 3. Specific analysis of viral strains. **A** Comparison between the HIV1 strains LAI and NDK. **B** Comparison between the HIV2 strains ROD and D205. Area A1 shows homologies between LTR sequences.

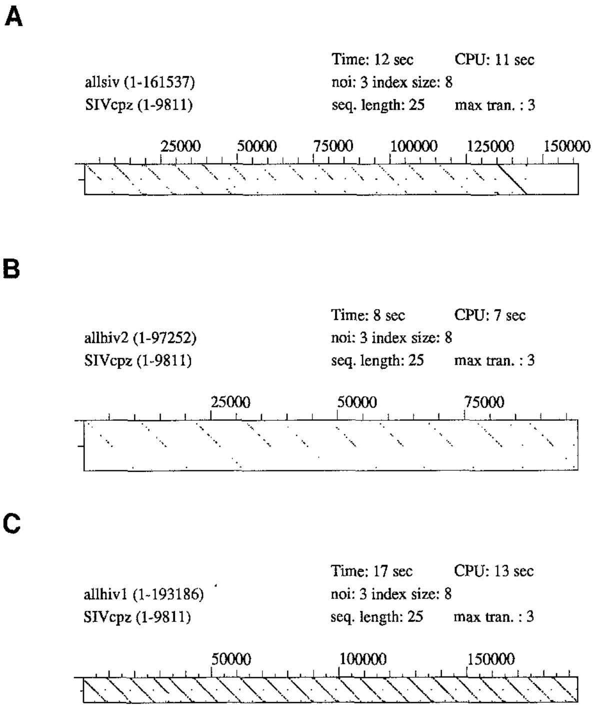

Fig. 4. Comparison between the SIV$_{\mathrm{CPZ}}$ virus and the three viral groups. **A** Comparison with the SIV group. **B** Comparison with the HIV2 group. **C** Comparison with the HIV1 group.

Specific Analysis

The genome analysis was refined by specific check pair comparisons. In a first computation series, we examined reference strains and confirmed the good concordance between, for example, the two HIV1 LAI and ELI (Alizon et al. 1986) genomes (data not shown). A similar degree of homology was observed between the LAI and NDK HIV1s (Fig. 3A), although the latter has been described as a more infectious strain (Wain-Hobson et al. 1985; Spire et al. 1989) and might have been expected to show a higher divergence. A similar check study made on HIV2$_{\mathrm{ROD}}$ (Clavel et al. 1986) and HIV2$_{\mathrm{BEN}}$ (Kirchhoff et al. 1990) strains (data not shown) gave similar results. However, the D205 strain (Gao et al. 1992), which was detected as being less homologous than other HIV2s on the general screening "allhiv2/allhiv2" reported in Fig. 1B line L1, was clearly confirmed in Fig. 3B as a putative subfamily ancestor because of its greater divergence. The lines in area A1 (Fig. 3B) corresponded to homologies between LTR sequences. For viruses which induced shifted positions for these lines, authors had reported only single LTR copies at different extremities of the genome.

This study was extended to some divergent strains belonging to the HIV2 and the SIV groups. They differed from the HIV1s and from the divergent strains HIV2$_{\mathrm{D205}}$ and SIV$_{\mathrm{CPZ}}$ (Peeters et al. 1989, 1992) to a similar extent (data not shown), but from these observations no particular connection could be established between the divergent strains belonging to the three groups. To clarify the distinctive behavior of the SIV$_{\mathrm{CPZ}}$ chimpanzee virus, additional analyses were performed (Fig. 4). Figures 4A and B showed an equal divergence between the SIV$_{\mathrm{CPZ}}$ and other simian viruses and HIV2s, while Fig. 4C clearly indicated a close relationship with HIV1s. These conclusions were reinforced when the SIV$_{\mathrm{CPV}}$ was compared to a selected African green monkey virus (AGM677A) (Fomsgaard et al. 1991), an HIV2 (ROD) (Clavel et al. 1986), and an HIV1 (LAI) (Wain-Hobson et al. 1985) (data not shown). Control comparisons performed pairwise between HIV1$_{\mathrm{LAI}}$, HIV2$_{\mathrm{ROD}}$, SIV$_{\mathrm{AGM677A}}$, and SIV$_{\mathrm{MM251}}$ (Kestler et al. 1988) corroborated the known divergence between these viruses (data not shown).

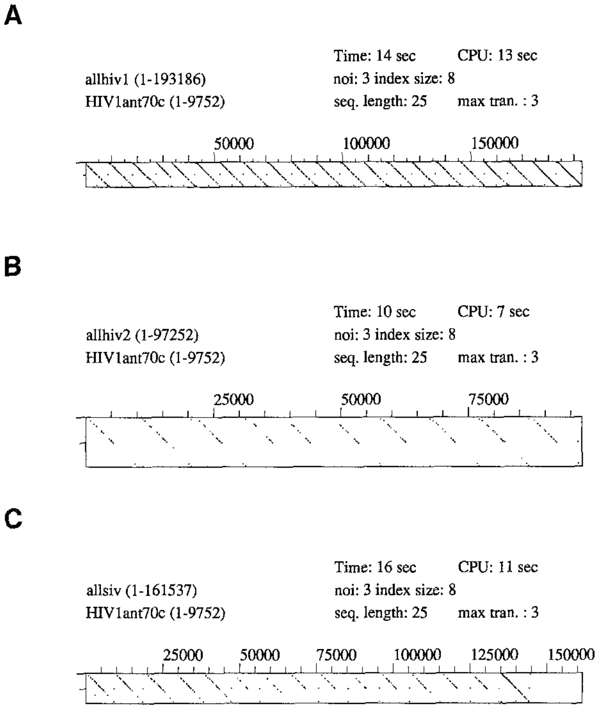

Fig. 5. Comparison between the HIV1-O_{ANT70C} virus and the three viral groups. **A** Comparison with the HIV1 group. **B** Comparison with the HIV2 group. **C** Comparison with the SIV group.

Example of Detailed Analysis

A detailed pair analysis was applied to the well-known divergent HIV1-O_{ANT70C} (De Leys et al. 1990, Vanden Haesvelde et al. 1994) and MVP5180 (Gürtler et al. 1994) HIV1 strains, which clearly appeared highly divergent from other HIV1s in Fig. 1A lines L1 and L2. This was confirmed by comparisons of artificial strain concatemers and single virus comparisons. As an illustration, the comparisons between ANT70C and the allhiv1, allhiv2, and allsiv concatemers were reported in Fig. 5A–C, respectively. More detailed pair comparisons exhibited the distinctive behavior of these to HIV1-O subtyped viruses. The LAI/ANT70C comparison (data not shown) confirmed the divergence described for these two strains, whereas ANT70C and MVP5180 were closely related, as reported in Fig. 6A. In area A1 the lines represented the comparisons of the repetitions of the LTR sequences at the extremities of the genomes. One should note the gap situated in the central part of these lines as well as in the portion of the central diagonal line framed in area A2. The framed gap corresponded to the U3 sequences of the LTR. Successive zoomings were performed to delimit the endpoints of nonhomologous domains from area A2. The two extremities of the gap were displayed together in Fig. 6B. They were designated as areas A1 and A2 and described in Fig. 6C and D, respectively. As shown in these figures, the gap was situated between nucleotides 145 and 382 for $HIV1_{ANT70C}$ and between nucleotides 145 and 353 for $HIV1_{MVP5180}$. The detection of divergent LTR sequences was the consequence of a 29-base-pair insertion in the $HIV1_{ANT70C}$ genome. Notably, a similar observation can be made for HIV2 LTRs. A comparison between the HIV1 and HIV2 LTRs indicated that about 100 bases are lacking in the HIV1 LTR immediately upstream of the LEF binding site.

It must be emphasized that 25 extra bases in the $HIV1_{ANT70C}$ LTR were situated between the LEF and the first NF-κB fixation sites and five extra bases were located between the two NF-κB binding sites (Jones and Petterlin 1994). For the $HIV1_{ANT70C}$ virus, the LEF site was situated between nucleotides 328 and 334, the first NF-κB site between nucleotides 368 and 377, and the second NF-κB site between nucleotides 396 and 410. For the $HIV1_{MVP5180}$ virus, the LEF site was located between nucleotides 328 and 335, the first NF-κB site between nucleotides 344 and 353, and the second NF-κB site between nucleotides 367 and 381, the latter corresponding to the position of the first NF-κB site in $HIV1_{ANT70C}$. On the other hand, one base was inserted in the LEF fixation site for the $HIV1_{MVP5180}$, lowering the difference to 29 bases. These 30 inserted bases contained a quadruplicate CTG motif organized as a tandem repeat and two separate single motifs. These data are in good agreement with the finding made by Vanden Haesevelde et al. (1994), who reported a triplication of the CTG motif upstream from the NF-κB sites. It is also noteworthy that the two NF-κB sites were not organized exactly like the HIV1 tandem profile described by Jones and Peterlin (1994) since 15 and 20 extra bases are inserted between the sites of the MVP5180 and the ANT70C viruses, respectively. These inserted sequences contained a triplicate CTG motif organized as a tandem repeat and a single motif separated by the ACA sequence **CTGCTGACACTG** instead of the single CT motif usually encountered in the HIV1s.

Conclusion

The present work investigated the computed homologies between different groups of immunodeficiency viruses, the HIV1s, HIV2s, and SIVs. The use of the PAGEC program, which we have described previously (Courtois and Moncany 1995), allowed rapid comparison between the different genomes. Its flexibility made different types of comparison possible, from cross-comparison of huge genomes to strain-to-strain comparison. All these analyses permitted identification of the domain of interest and, ultimately, the lexical base sequence. In our study, huge genomes corresponded to artificially concatemerized viral genomes belonging to defined groups. They were labeled ''allhiv1,'' ''allhiv2,'' and ''allsiv,'' and their self-computation time required 30–85 s. Computation time dropped to about 10 s when comparing concatemers

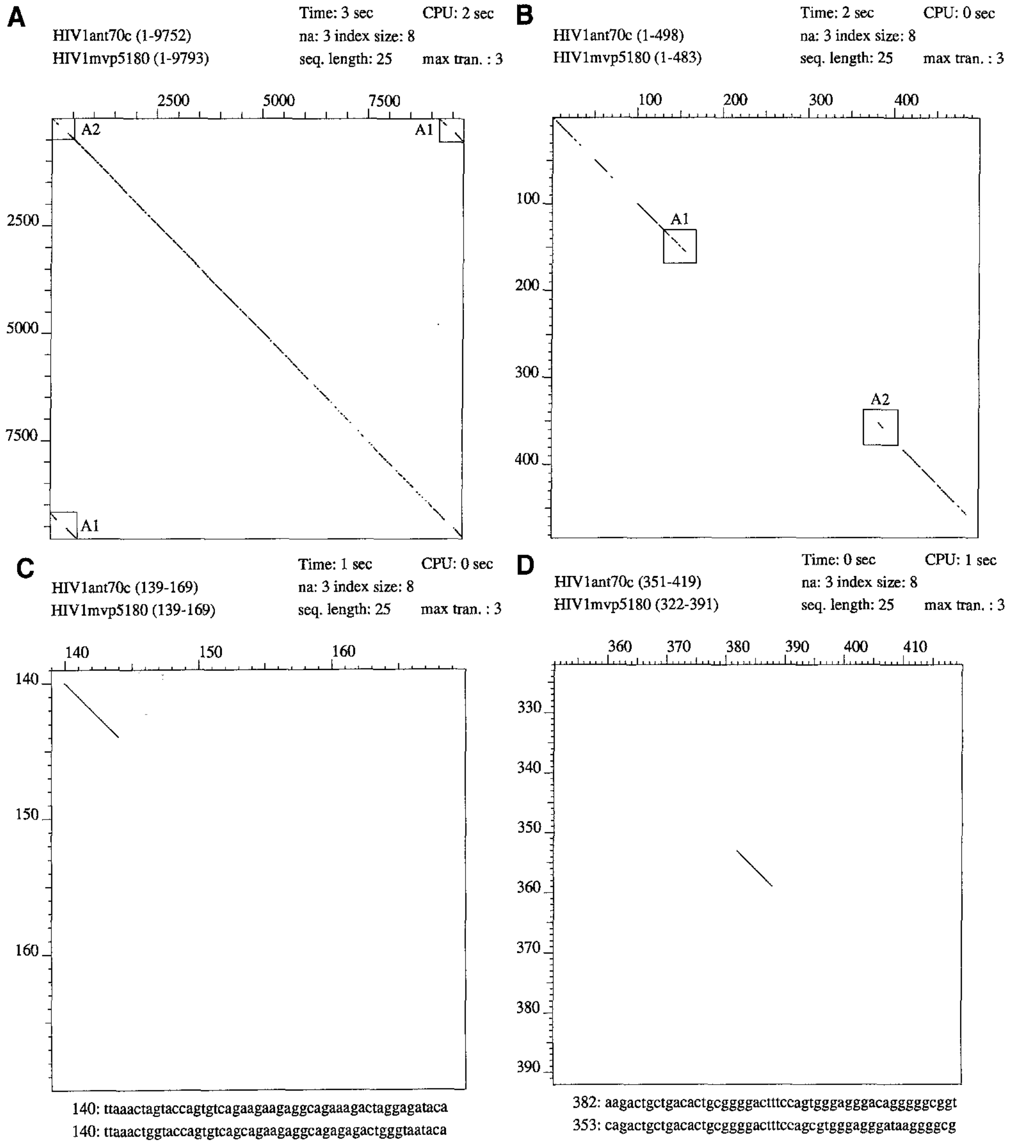

Fig. 6. Specific analysis of the HIV1-O viruses. **A** Comparison between the two HIV1-O ANT70C and MVP5180 strains. Areas A1 and A2 showed divergence in the LTR sequences. **B** Zoom of area A2 indicated in **A**. The magnification of the gap in the LTR sequences allowed the visualization of the limits of the nonhomology domains framed in areas A1 and A2. **C** Zoom of area A1 (left end of the gap) framed in **B**. The base sequences reported under the graph detail the homology domain visualized by the diagonal line. The *number preceding each of the sequences* indicates the position of the starting nucleotide in the complete viral genome. The sequences belong, from top to bottom, to the HIV1-Os ANT70C and MVP5180. **D** Zoom of area A2 (right limit of the gap) framed in **B**. The base sequences reported under the figure are defined as in **C**.

to a single genome, reaching only a few seconds for pair comparisons.

The concatemer comparisons were considered to be a rapid screening technique allowing the differentiation of homologous and nonhomologous domains, such as those resulting from omitted LTR sequences and redundancies in these sequences and, more strikingly, strain divergence (Fig. 1). This was possible for the HIV1s ANT70C and MVP5180 (Fig. 1A), the HIV2$_{D205}$ (Fig. 1B) and the SIV$_{CPZ}$ (Fig. 1C). The flexibility of the system also allowed the direct detection of remarkable homologies in the LTR sequences.

The variation observed in the extent of homology led to artificial concatemer/individual strain comparison and to cross-comparisons between viral strains. Such comparisons confirmed the divergence previously observed for particular viral genomes and allowed tentative phylogenetic interpretation. For instance, these comparisons clearly demonstrated that the SIV chimpanzee virus is more closely related to HIV1s than to SIVs and HIV2s (Figs. 2B and 4) (Vanden Haesvelde et al. 1994; Gürtler et al. 1994; Peeters et al. 1989, 1992; Huet et al. 1990). However, viruses infecting macaque monkeys, as well as the single African green monkey AXX virus (Franchini et al. 1987), had a tendency to be organized like the HIV2s (Fig. 2C). Another striking observation concerned MP and RV1 viruses, which have been described as D-particle-forming viruses which do not induce lethal immunodeficiency (Sonigo et al. 1986; Power et al. 1986). All the comparisons we made confirmed their specific identities as D particles possessing only point homologies with other viruses. These homologous domains were also localized to the LTR region.

The methodology used in this study (which led to confirming homology observations), when followed by more refined sequence analysis, is essentially a rapid means of deriving phylogenetic trees. Moreover, this type of survey is highly useful in the selection of domains which should be studied in greater detail. Figure 6A–D presented an example of such an analysis. From the allhiv1/allhiv1 comparison, we verified the specific subtyping of the two HIV1-Os ANT70C and MVP5180 (De Leys et al. 1990; Vanden Haesvelde et al. 1994; Gürtler et al. 1994). The pairwise comparison of these strains (Fig. 6A) revealed a gap in the LTR sequences, which usually correspond to well-conserved regions (area A2). Zooming for each extremity of the gap allowed the identification of the divergent sequence (Fig. 6B–D). The gap was shown to correspond to an extra 29-nucleotide sequence, present only in the ANT70C LTR sequence and positioned between the LEF and the first NF-κB fixation site. Our sequence study pointed out that this domain contained a quadruplicate of the CTG motif described by Vanden Haesvelde et al. (1994).

In conclusion, by using PAGEC and the concatemerization method, a global analysis accomplished in a few seconds permitted the detailed identification of divergent viral strains and led to the characterization of some distinctive lexical base sequences. The data provided an immediate illustration of viral families and the unique behavior of some strains such as the SIV_{CPZ} and SIV_{AXX} viruses. They also led to a fast intuitive assessment of the studied sequences and to the selection of more-or-less homologous domains. The study of the organization of a subregion of the HIV1-O LTR—from a genomic comparison to the identification of a short base sequence—demonstrates the efficiency of the method for molecular genetic investigations.

Acknowledgments. The authors thank Mr. Jean-Pierre Cros (director of the C.S.L./C.N.A.M.) for his logistic support, and Prof. André Klier and Drs. Uriel Hazan and Jennifer Richardson for fruitful discussions and critical review of the manuscript.

References

Alizon M, Wain-Hobson S, Gluckman J-C, Sonigo P (1986) Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell 46:63–74

Charneau P, Borman AM, Kuillent C, Guétard D, Chamaret S, Cohen J, Remy G, Montagnier L, Clavel F (1994) Isolation and envelope sequence of a highly divergent HIV1 isolate: definition of a new HIV1 group. Virology 205:247–253

Clavel F, Guyader M, Guetard D, Salle M, Gluckman J-C, Alizon M (1986) Molecular cloning and polymorphism of the human immunodeficiency virus type 2. Nature (London) 324:691–695

Courtois PRR, Moncany MLJ (1995) A probabilistic algorithm for interactive huge genome comparison. Comput Appl Biol Sci 11: 657–665

De Leys R, Vanderborght B, Vanden Haesvelde M, Heyndrickx I, van Geel A, Wauters C, Bernaerts R, Saman E, Nijs P, Willems B, Taelman H, van der Groen G, Piot P, Tersmette T, Huisman JG, van Heuverswyn H (1990) Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of West-Central African origin. J Virol 64:1207–1216

Dewhurst S, Embretson JE, Anderson DC, Mullins JI, Fultz PN (1990) Sequence analysis and acute pathogenicity of molecular cloned SIV. Nature (London) 345:636–640

Fitch WM (1969) Locating gaps in amino acid sequences to optimize homology between proteins. Biochem Genet 3:99–108

Fomsgaard A, Hirsch VM, Allan JS, Johnson PR (1991) A highly divergent proviral DNA clone of SIV from a distinct species of african green monkey. Virology 182:397–402

Franchini G, Gurgo C, Guo HG, Gallo RC, Collati E, Fargnoli KA, Hall LF, Wong-Staal F, Reitz MS Jr (1987) Sequence of simian immunodeficiency virus and its relationship to the human immunodeficiency viruses. Nature (London) 328:539–543

Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, Greene BM, Sharp PM, Shaw GM, Hahn BH (1992) Human infection by genetically diverse SIVsm-related HIV-2 in West Africa. Nature (London) 358:495–499

Gao F, Yue L, Robertson DL, Hill SC, Hui H, Biggar RJ, Neequaye AE, Whelan TM, Ho DD, Shaw GM, Sharp PM, Hahn BH (1994) Genetic diversity of Human Immunodeficiency Virus type 2: evidence for distinct sequence subtypes with differences in virus biology. J Virol 68:7433–7447

Gibbs AJ, McIntyre GA (1970) The diagram, a method for comparing sequences, its use with amino acid and nucleotide sequences. Eur J Biochem 16:1–11

Gürtler LG, Hauser PH, Eberle J, von Brunn A, Knapp S, Zekeng L, Tsague JM, Kaptue L (1994) A new subtype of Human Immunodeficiency Virus type 1 (MVP-5180) from Cameroon. J Virol 68: 1581–1585

Harr R, Hagblom P, Gustafsson P (1982) Two dimensional graphic analysis of DNA sequence homologies. Nucleic Acids Res 10:365–374

Hirsch VM, Dapoliti GA, Goldstein S, McClure H, Emau P, Fultz PN, Isahakia M, Lenroot R, Myers G, Johnson PR (1993) A distinct African lentivirus from Sykes' monkeys. J Virol 63:1517–1528

Huet T, Cheynier R, Meyerhans A, Roelants G, Wain-Hobson S (1990) Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature (London) 345:356–359

Jin MJ, Hui H, Robertson DL, Müller MC, Barré-Sinoussi F, Hirsch VM, Allan JS, Shaw GM, Sharp PM, Hahn BH (1994) Mosaic genome structure of Simian immunodeficiency virus from West African green monkeys. EMBO J 13:2935–2947

Jones K, Petterlin T (1994) Control of RNA initiation and elongation at the HIV-1 promoter. Annu Rev Biochem 63:717–743

Kestler HW, Li Y, Naidu YM, Naidu YM, Butler CV, Ochs MF, Jaenel G, King NW, Daniel MD, Desrosiers RC (1988) Comparison of Simian immunodeficiency virus isolates. Nature (London) 331: 619–622

Kirchhoff F, Jentsch K, Bachmann B, Stuke A, Laloux C, Lueke W, Stahl-Henning C, Schneider J, Nieselt K, Eigen M, Hunsmann G (1990) A novel proviral clone of HIV-2: biological and phylogenetic relationship to other primate immunodeficiency viruses. Virology 177:305–311

Kreutz R, Dietrich U, Kuehnel H, Nieselt-Struwe K, Eigen M, Ruebsamen-Waigmann H (1992) Analysis of the envelope region of the highly divergent HIV-2alt isolate extends the known range of variability within the primate immunodeficiency viruses. AIDS Res Hum Retroviruses 8:1619–1629

Lefêvre C, Ikeda J-E (1994) A fast word search algorithm for the representation of sequence similarity in genomic DNA. Nucleic Acids Res 22:404–411

McLachlan AD (1971) Tests for comparing related amino acid sequences. Cytochrome c and cytochrome c_{551}. J Mol Biol 61:409–424

Maizel JV Jr, Lenk RP (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc Natl Acad Sci USA 78: 7665–7669

Myers G, Korber B, Wain-Hobson S, Smith RF, Pavlakis GN (1993) Human retroviruses and AIDS 1993. A compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, NM

Novotny J (1982) Matrix program to analyze primary structure homology. Nucleic Acids Res 10:127–131

Peeters MC, Honoré T, Huet L, Bedjabaga L, Ossari S, Bussi P, Cooper RW, Delaporte E (1989) Isolation and partial characterization of an HIV-related virus occurring naturally in chimpanzees in Gabon. AIDS 3:625–630

Peeters MC, Fransen K, Delaporte E, Vanden Haesvelde M, Gershy-Damet GM, Kestens L, van der Groen G, Piot P (1992) Isolation and characterization of a new chimpanzee lentivirus (simian immunodeficiency virus isolate cpz-ant) from a wild captured chimpanzee. AIDS 6:447–451

Power MD, Marx PA, Bryant ML, Gardner MB, Barr PJ, Luciw PA (1986) Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome virus. Science 231:1567–1572

Robertson DL, Hahn BH, Sharp PM (1995a) Recombination in AIDS viruses. J Mol Evol 40:249–259

Robertson DL, Sharp PM, McCutchan FE, Hahn BH (1995b) Recombination in HIV-1. Nature (London) 374:124–126

Schwartz S, Miller W, Yang C-M, Hardison RC (1991) Software tools for analysing pairwise alignments of long sequences. Nucleic Acids Res 17:4663–4667

Sonigo P, Barker CS, Hunter E, Wain-Hobson S (1986) Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. Cell 45:375–385

Spire B, Sire J, Zachar V, Rey F, Klatzmann D, Galibert F, Hampe A, Chermann J-C (1989) Nucleotide sequence of HIV1-NDK, a highly cytopathic strain of the human immunodeficiency virus HIV1. Gene 81:275–284

Staden R (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res 10:2951–2961

Steinmetz M, Frelinger JG, Fischer D, Hunkapillar T, Periera D, Weissman SM, Uehara H, Natheson S, Hood L (1981) Three cDNA clones encoding mouse transplantation antigens: homology to immunoglobulin genes. Cell 24:125–134

Tristem M, Hill F, Karpas A (1991) Nucleotide sequence of a Guinea-Bissau-derived human immunodeficiency virus type 2 proviral clone (HIV-2CAM2). J Gen Virol 72:721–724

Tsujimoto H, Hasegawa A, Maki N, Fukasawa M, Miura T, Ohta Y, Ishibawa KI, Nakai M, Frost E, Roelants GE, Roffi J, Hayami M (1989) Sequence of a novel simian immunodeficiency virus from a wild-caught African Mandrill. Nature (London) 341:539–541

Vanden Haesvelde M, Decourt J-L, De Leys R, Vanderborght B, van der Groen G, van Heuverswijn H, Saman E (1994) Genomic cloning and complete sequence analysis of a high divergent african Human Immunodeficiency Virus isolate. J Virol 68:1586–1596

Wain-Hobson S, Sonigo P, Danos O, Cole S, Alizon M (1985) Nucleotide sequence of the AIDS virus, LAV. Cell 40:9–17