The EMBO Journal vol.11 no.9 pp.3405-3412, 1992

Evolution of the primate lentiviruses: evidence from *vpx* and *vpr*

Michael Tristem, Craig Marshall[1], Abraham Karpas and Fergal Hill[2]

Department of Haematology, University of Cambridge Clinical School, Hills Road, Cambridge CB2 2QH, UK

[1]Present address: Department of Biochemistry, University of Otago, Dunedin, New Zealand
[2]Corresponding author

Communicated by A.Klug

The genomes of the four primate lentiviral groups are complex and contain several regulatory or accessory genes. Two of these genes, *vpr* and *vpx*, are found in various combinations within the four groups and encode proteins whose functions have yet to be elucidated. Comparison of the encoded protein sequences suggests that the *vpx* gene within the HIV-2 group arose by the duplication of an ancestral *vpr* gene within this group. Evolutionary distance analysis showed that both genes were well conserved when compared with viral regulatory genes, and indicated that the duplication occurred at approximately the same time as the HIV-2 group and the other primate lentivirus groups diverged from a common ancestor. Furthermore, although the SIVagm vpx proteins are homologous to the HIV-2 group vpx proteins, there are insufficient grounds from sequence analysis for classifying them as vpx proteins. Because of their similarity to the vpr proteins of other groups, we suggest reclassifying the SIVagm *vpx* gene as a *vpr* gene. This creates a simpler and more uniform picture of the genomic organization of the primate lentiviruses and allows the genomic organization of their common precursor to be defined; it probably contained five accessory genes: *tat*, *rev*, *vif*, *nef* and *vpr*.

Key words: gene duplication/genomic organization/immunodeficiency viruses/orthologous/phylogeny

Introduction

The emergence of the acquired immune deficiency syndrome pandemic, and the unexpectedly complex genomic organization of its principal causative agent HIV-1, have created enormous interest in the evolution of lentiviruses. It has become clear that lentiviruses are widely distributed in mammals, having been found in domestic cats and four domesticated ungulate species (sheep, goats, cattle and horses) as well as in a number of primate species (Barré-Sinoussi *et al.*, 1983; Daniel *et al.*, 1985; Sonigo *et al.*, 1985; Clavél *et al.*, 1986; Fultz *et al.*, 1986; Kawakami *et al.*, 1987; Fukasawa *et al.*, 1988; Henderson *et al.*, 1988a; Olmstead *et al.*, 1989; Tsujimoto *et al.*, 1989; Garvey *et al.*, 1990; Huet *et al.*, 1990; Saltarelli *et al.*, 1990). The recent isolation of a lentivirus from a Sykes monkey suggests that further groups remain to be discovered (Emau *et al.*, 1991).

The primate lentiviruses consist of the human immunodeficiency viruses types 1 and 2 (HIVs) and the related simian immunodeficiency viruses (SIVs). They have been divided into four groups on the basis of homology of their *gag* and *pol* gene products (Desrosiers, 1990). These are: (i) the HIV-1 group, which includes an isolate from a chimpanzee (SIVcpz) as well as many isolates from humans, (ii) the HIV-2 group comprising HIV-2 isolates and viruses infecting sooty mangabey monkeys (SIVsmm) and a number of macaque species (SIVmac), (iii) the African green monkey group (SIVagm) and (iv) the mandrill group (SIVmnd) from which only one isolate has been well characterized (Tsujimoto *et al.*, 1989). Several studies have shown that the four groups of primate lentiviruses appear to have diverged from each other at approximately the same time. However, it has been difficult to determine the order in which these groups arose and the exact relationship between them due to the divergence that has since occurred (Sharp and Li, 1988; Dietrich *et al.*, 1989; Doolittle *et al.*, 1989; Tsujimoto *et al.*, 1989; Gojobori *et al.*, 1990; Myers *et al.*, 1990).

The primate lentiviruses have complex genomes with many of the encoded proteins, such as the products of the three major open reading frames gag, pol and env and the regulatory proteins tat, rev, vif and nef, present in all four groups (Wain-Hobson *et al.*, 1985; Chakrabarti *et al.*, 1987; Guyader *et al.*, 1987; Fukasawa *et al.*, 1988; Tsujimoto *et al.*, 1989; Huet *et al.*, 1990). In addition, three of the four groups (HIV-2, HIV-1 and SIVmnd) encode a protein termed vpr (virus protein R) (Wong-Staal *et al.*, 1987), whereas two (HIV-2 and SIVagm), encode a protein termed vpx (virus protein X) (Guyader *et al.*, 1987).

We have previously suggested that the *vpx* gene within the HIV-2 group arose by duplication of the *vpr* gene within this group, and have proposed that the presence of a *vpr* or *vpr*-like gene is a feature shared by all primate lentiviruses, which distinguishes them from other mammalian isolates (Tristem *et al.*, 1990). In this study we present additional evidence for these conclusions; we use genetic distances to calculate the rate of evolution of the *vpr* and *vpx* genes in the HIV-2 group, estimate the date of the duplication, and show that there are no compelling grounds from sequence analysis for classifying an SIVagm gene as a *vpx* rather than a *vpr*. Our results are discussed in relation to the genomic organization and evolution of the primate lentiviruses.

Results

Homology between vpr and vpx proteins

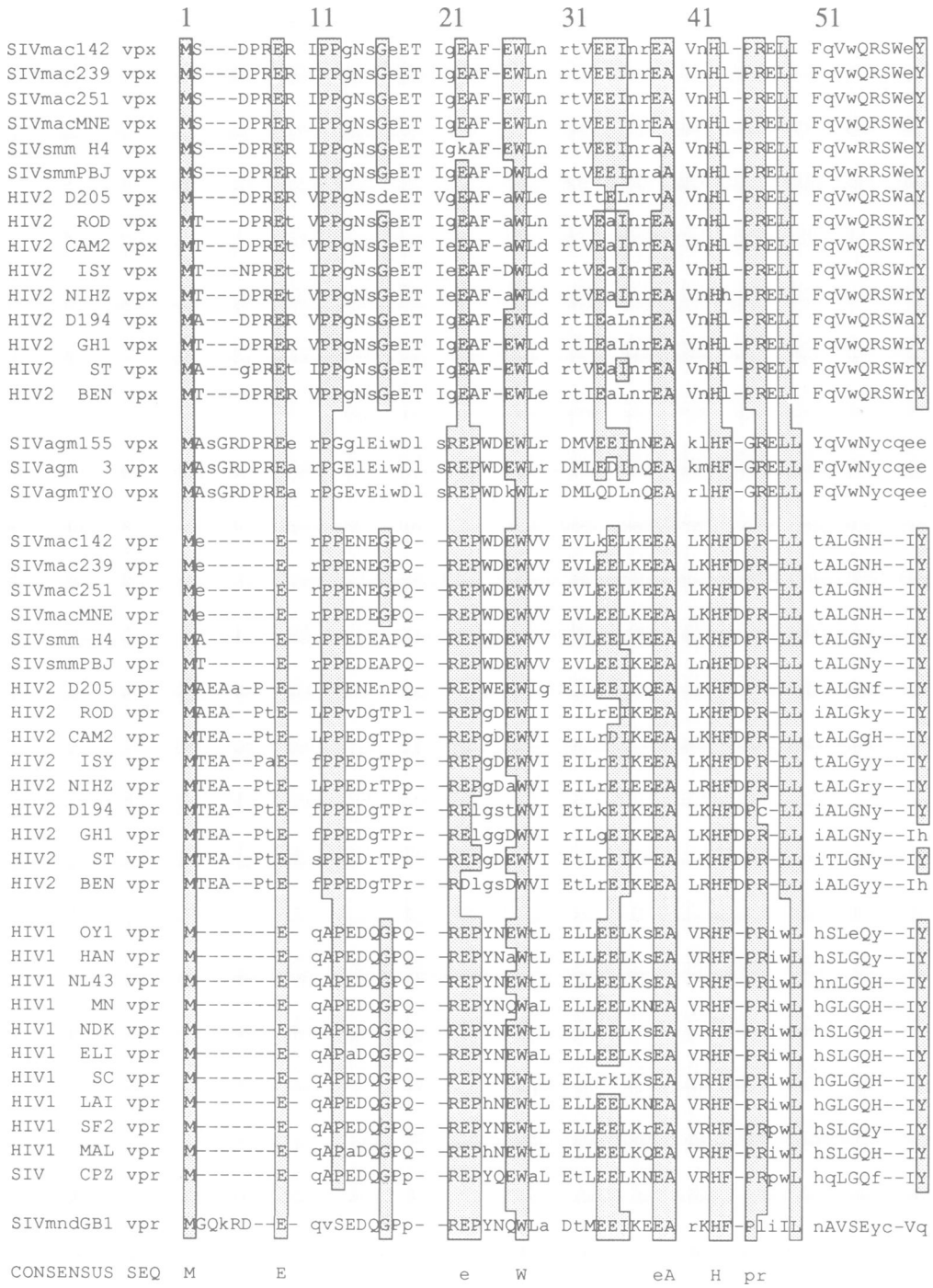
Forty-five vpx and vpr protein sequences from the human retroviruses and AIDS database (Myers *et al.*, 1990) and other sources were aligned as described (Figure 1). In

© Oxford University Press

addition to the initiation codon, five other amino acids are completely conserved between all the proteins studied: E9, W28, A39, H43 and H78. Eight more residues are present in almost all the proteins: E23, E38, P46, R47, Q82, G92 and C93, whereas many other residues are conserved according to the classification of Dayhoff *et al.* (1978). Furthermore, similarities were also apparent when hydrophobicity or hydrophilicity plots (Hopp and Woods, 1981; Kyte and Doolittle, 1982) of vpr or vpx proteins from the HIV-2 group were constructed (M.Tristem and F.Hill, unpublished data). The statistical significance of HIV-2 vpx and vpr protein similarity was investigated using shuffle analysis (Pearson and Lipman, 1988). In all cases similarity was significant to at least 9.5 standard deviations. A search of the PIR and SWISSPROT databases by the programs FASTA (Pearson and Lipman, 1988), PSQ (George and Barker, 1990) and PROFILE (Gribskov *et al.*, 1987) using the regions or residues conserved between vpx and vpr did not show any obvious similarity to other proteins or classes of proteins.

Analysis of SIVagm vpx proteins

The SIVagm vpx proteins align well with the vpr proteins from the other groups. This similarity was measured (Table I) using the program SIP (Staden, 1990) which uses a binary scoring system (allocating 1 for an identical match and 0 for a non-identical match or gap) to calculate the percentage similarity between two proteins. The result shows that SIVagm vpx proteins generally have the highest number of



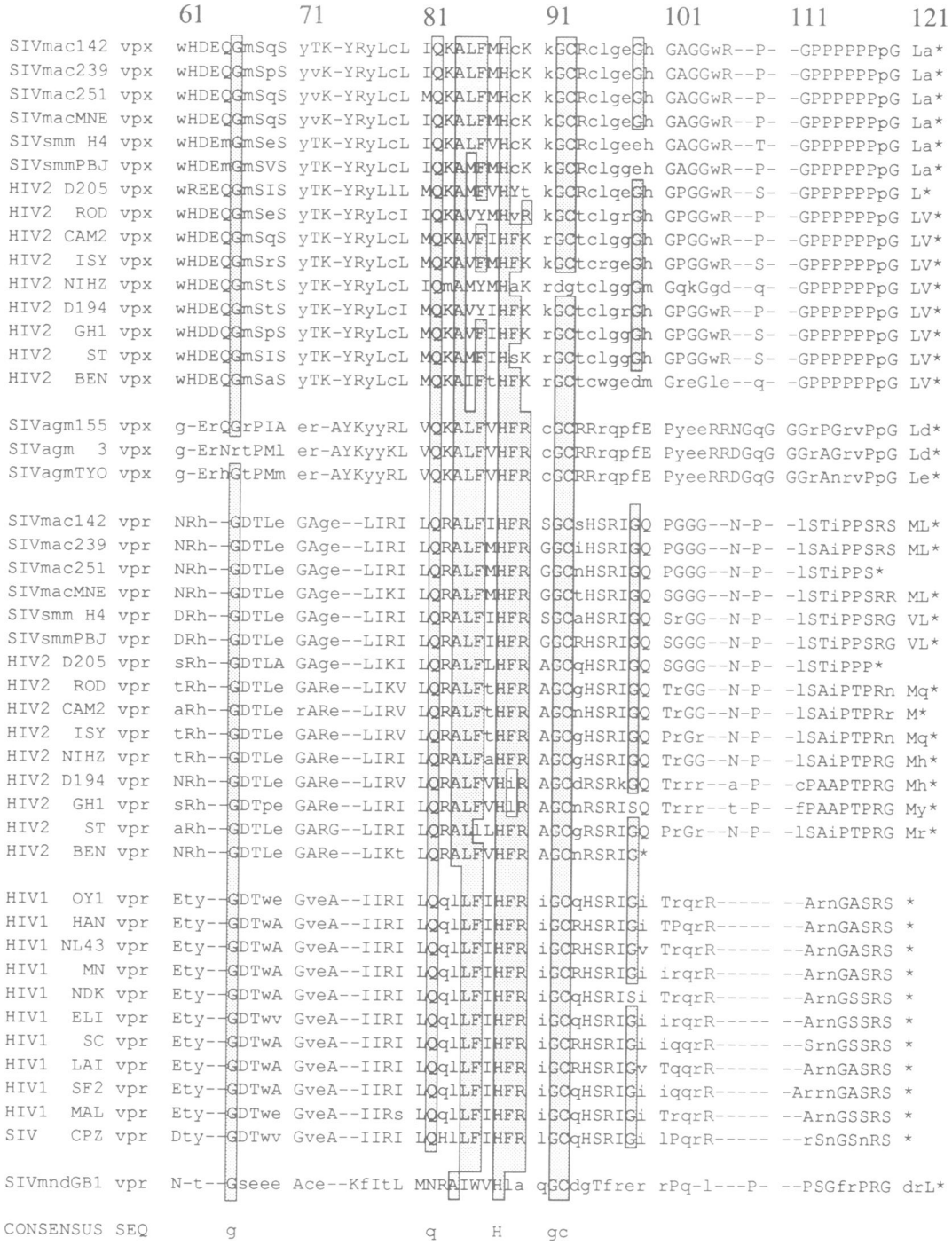


Fig. 1. Multiple alignment of the vpx and vpr proteins from the four primate lentiviral groups. Protein sequences were assembled using the HOMED (Stockwell and Petersen, 1987) program and a pairwise comparison was performed using the program AMPS (Barton and Sternberg, 1987). A multiple alignment of the sequences was obtained by again using the program AMPS with a gap penalty of 8. A small number of adjustments were then made by eye. Shaded boxes represent residues which are well conserved between vpx and vpr. Capital letters represent classes of amino acids conserved between the various groups, as follows: DENQ, ILVM, SPTAG, RKH, FYW, C. The consensus sequence shows invariant residues (upper case) and highly conserved residues (lower case). Dashes represent gaps introduced into the sequences to allow optimal alignment.

matches with the vpx proteins of the HIV-2 group with a high number also being obtained with the vpr proteins of this group. The lowest percentage of matches was obtained with the SIVmnd and HIV-1 groups.

The relationship of SIVagm vpx proteins was further investigated by phylogenetic analysis. To determine whether SIVagm vpx proteins were significantly more similar to the vpx proteins of the HIV-2 group than to any of the prototype vpr proteins, we used the EVOMONY program (Lake, 1987), which uses four-taxon tests to establish the statistical significance of a particular topology. We were unable to establish, using this method, that phylogenetic trees which clustered the SIVagmTYO *vpx* and HIV-2ROD *vpx* genes on the same lineage were significantly more likely than those which clustered the SIVagmTYO *vpx* gene with any of the prototype *vpr* genes (unpublished data).

Phylogenetic trees containing vpr proteins from the HIV-2, HIV-1 and SIVmnd groups and vpx proteins from the

SIVagm group were then constructed using the PROTPARS program from the PHYLIP package (Felsenstein, 1990) and were based on the alignment shown in Figure 1. Multiple runs were performed and the resulting trees were combined into one consensus tree, therefore allowing each node on the tree to be assigned a consistency index (a consistency index of 1 indicating that the node was present in every tree obtained). The resulting tree (Figure 2) shows the branch lengths of all the major lineages are of approximately equal length. Although the topology of our tree was reasonably consistent (we obtained the same topology using a random sequence as the root, and also when the gaps shown in Figure 1 were included) we do not suggest it is correct, as there appears to be insufficient phylogenetic information to make definite statements about the relationships of the four primate lentiviral groups.

Table I. Percentage identity between the vpx proteins of SIVagm and vpx or vpr proteins from the other HIV/SIV groups

Isolate	SIVmndGB1[a]	HIV-1LAI[a]	HIV-2CAM2		SIVsmmH4	
	vpr	vpr	vpr	vpx	vpr	vpx
SIVagmTYO	26.7	27.0	33.6	35.1	36.3	33.3
SIVagm155	28.3	26.4	33.1	38.5	35.5	38.6
SIVagm3	27.5	26.2	31.3	36.6	35.2	36.7

[a]SIVmnd and HIV-1 lack a vpx protein.

Calculating the rate of evolution of vpx and vpr

The rate of evolution of the vpr and vpx genes in the HIV-2 group was estimated using the method of Nei and Gojobori (1986) or DNADIST (Felsenstein, 1990) and pairs of genes from HIV-2ROD/SIVmac142 or HIV-2CAM2/SIVsmmH4. In order to determine the rate of evolution of two genes it is also necessary to know the divergence date between two species which contain both genes. It has been proposed that the lentiviral *pol* genes have a relatively consistent molecular clock; if so, this enables phylogenetic trees to be calibrated on the basis of one time point (Quérat *et al.*, 1990). Thus, although the divergence date of HIV-2 and SIVmac/

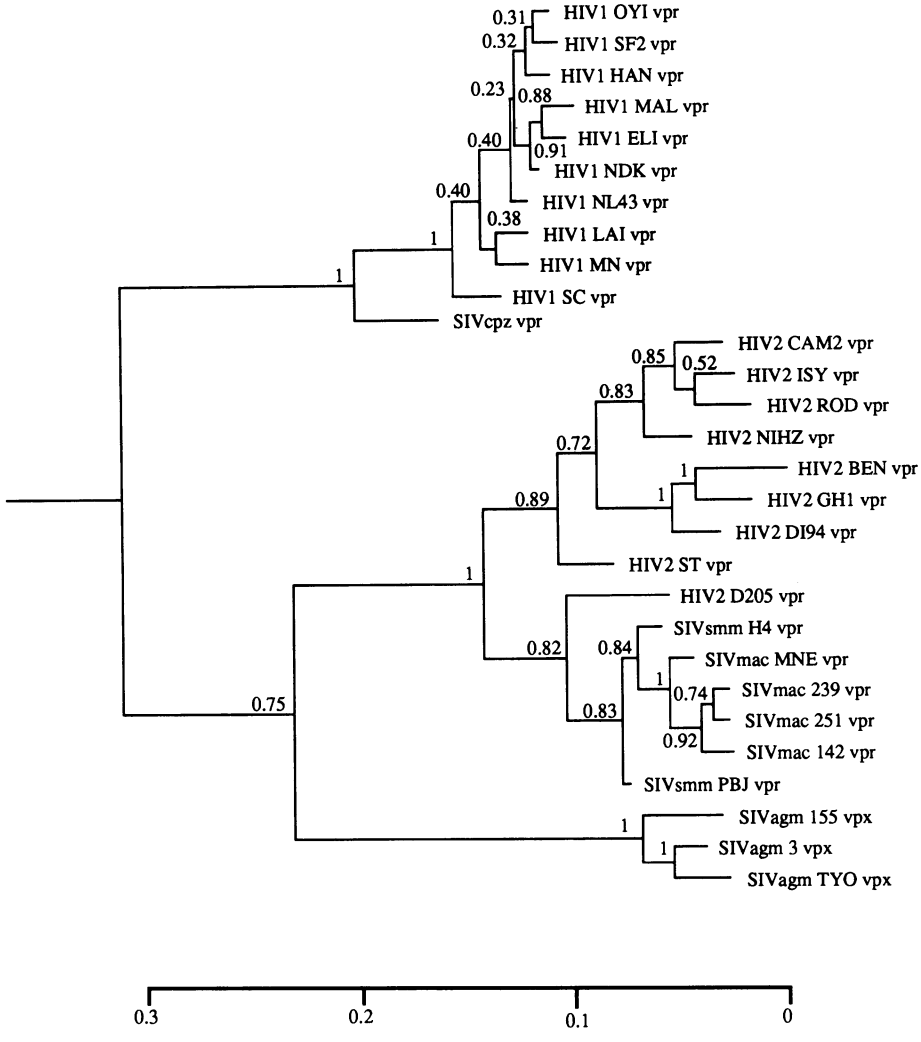


Fig. 2. Phylogenetic tree showing the relationship between the vpr proteins from the HIV-1, HIV-2 and SIVmnd groups and the vpx proteins from the SIVagm group. Trees were constructed using the program PROTPARS from the PHYLIP package (Felsenstein, 1990). The tree was rooted on the SIVmnd vpr sequence. Multiple runs were performed and the resulting trees were combined into one using the program CONSENSE (Felsenstein, 1990). Figures present at each node represent the consistency index for that node. The lengths of the horizontal lines are proportional to the minimum number of nucleotide changes required to generate the observed variation. The lengths of the vertical lines are for clarity only. The bar scale represents the fraction of nucleotide substitutions per site.

SIVsmm is not known directly it can still be inferred from the known divergence dates of other lentiviruses. It has been estimated that isolates of HIV-2 and SIVmac diverged ~50 years ago (Quérat *et al.*, 1990), a figure based on the known minimum divergence times of two closely related ovine lentiviruses (visna and SA-OMVV). This figure was used to estimate the rate of evolution of the *vpx* and *vpr* genes within the HIV-2 group (Table II).

We obtained average rates of 1.6×10^{-3} non-synonymous substitutions per site per year for the *vpx* gene and 2.9×10^{-3} non-synonymous substitutions per site per year for the *vpr* gene. The rates of evolution are similar to those calculated for lentiviral *pol* genes of 0.80×10^{-3} (HIV-1) and 0.96×10^{-3} (SA-OMVV/visna) non-synonymous substitutions per site per year (Sharp and Li, 1988; Quérat *et al.*, 1990). They also agree with the observed variation in protein sequences, which show the vpx proteins within the HIV-2 group to be almost as well conserved as the structural proteins gag and pol, whereas vpr proteins are more variable (Franchini *et al.*, 1989; Hirsch *et al.*, 1989; Myers *et al.*, 1990; Tristem *et al.*, 1991). Furthermore, when we constructed a phylogenetic tree of the vpx proteins from the HIV-2 group the branch lengths were shorter when compared to the corresponding vpr proteins in Figure 2, indicating that *vpx* genes are evolving at a lower rate (unpublished data).

Estimating the date of the duplication

The average rate of evolution of the two genes was then used to estimate the duplication date of the *vpr* gene in the HIV-2 group. The date was obtained by using evolutionary distances calculated between pairs of parologous and orthologous *vpr* and *vpx* genes using DNADIST (see Table III). We estimate the duplication to have occurred ~200 years ago. This date is obviously only approximate as it relies on a divergence date of HIV-2 and SIVmac which is itself only an estimate. However, the figures do agree well with the proposed divergence dates of SIVagm, HIV-1 and HIV-2. Most reports have indicated that divergence occurred in the region of 150–280 years ago (Sharp and Li, 1988; Yokoyama *et al.*, 1988; Quérat *et al.*, 1990). In fact Quérat *et al.* (1990) from which the divergence date of HIV-2/SIVmac was obtained suggested that HIV-2 diverged from the other primate lentiviral groups ~200 years ago.

Evolutionary distances between pairs of parologous genes were ~4-fold higher than those between orthologous genes. Thus, whatever the actual date of the divergence of HIV-2 and SIVmac isolates the duplication event occurred a considerable time before this, by these calculations approximately four times as long, and at about the time the HIV-2 group was diverging from the other three groups.

Discussion

We have previously shown that all prototype primate lentiviral vpr and vpx proteins share regions of similarity with each other and proposed this as evidence for the duplication of the *vpr* gene within the HIV-2 group (Tristem *et al.*, 1990). These similarities are more obvious when more sequences are compared, with a significant proportion of the residues being conserved in all, or the vast majority of these proteins. Some of the HIV-2 group vpr proteins share as much as 30% amino acid identity with the vpx proteins from the same isolate.

In addition to the homology shown in Figure 1, there is further evidence supporting the duplication as vpx and vpr share many properties. Both are well conserved within and between the primate lentiviral groups, in contrast to the viral regulatory proteins which are more divergent (Franchini *et al.*, 1989; Hirsch *et al.*, 1989; Myers *et al.*, 1990; Tristem *et al.*, 1991). For example, the rate of evolution of the *vpx* gene in the HIV-2 group shows that it is almost as well conserved as the pol polyprotein. Furthermore, vpr and vpx have been shown to be present in virions, whereas other regulatory gene products such as tat, rev, vif and nef are not (Franchini *et al.*, 1988; Henderson *et al.*, 1988b; Yu *et al.*, 1988, 1990; Kappes *et al.*, 1989; Cohen *et al.*, 1990; Yuan *et al.*, 1990). Consistent with this, it has been

Table II. Rates of nucleotide substitution/site/year for *vpx* and *vpr* genes

	Synonymous		Non-synonymous		Total		Total (DNADIST)	
	Substitutions/ site	Rate of evolution[a]	Substitutions/ site	Rate of evolution[a]	Substitutions/ site	Rate of evolution[a]	Substitutions/ site	Rate of evolution[a]
HIV-2ROD vpx/SIVmac142 vpx	1.31	25×10^{-3}	0.07	1.3×10^{-3}	0.21	4.0×10^{-3}	0.23	4.4×10^{-3}
HIV-2CAM2 vpx/SIVsmmH4 vpx	0.78	15×10^{-3}	0.09	1.8×10^{-3}	0.20	3.9×10^{-3}	0.21	4.0×10^{-3}
HIV-2ROD vpr/SIVmac142 vpr	0.63	12×10^{-3}	0.16	3.1×10^{-3}	0.26	5.0×10^{-3}	0.29	5.5×10^{-3}
HIV-2CAM2 vpr/SIVsmmH4 vpr	0.68	13×10^{-3}	0.14	2.7×10^{-3}	0.24	4.6×10^{-3}	0.27	5.2×10^{-3}

[a]Rates (in substitutions/site/year) were calculated using the divergence date of Quérat *et al.* (1990).

Table III. Evolutionary distances between *vpr* and *vpx* genes and the estimated time since their duplication

	Average evolutionary distance between orthologous genes (total substitutions/site)	Average evolutionary distance between parologous genes (total substitutions/site)	Average rate of evolution[a] (substitutions/site/year)	Time since duplication[a] (years)
HIV-2ROD/SIVmac142	0.26	1.00	5.0×10^{-3}	200
HIV-2ROD/SIVsmmH4	0.24	0.99	4.6×10^{-3}	214
HIV-2CAM2/SIVmac142	0.25	0.98	4.8×10^{-3}	204
HIV-2CAM2/SIVsmmH4	0.24	0.96	4.6×10^{-3}	208

[a]Calculations performed using the divergence date of Quérat *et al.* (1990).

suggested that both vpr and vpx function at an early stage in the viral life cycle (Kappes *et al.*, 1991). Packaging probably occurs independently as vpx protein is detectable in vpr mutant virus and vice versa (Yu *et al.*, 1990), but a common packaging mechanism may underlie some of the amino acid conservation. Neither protein is essential for replication in cultured cells (Dedera *et al.*, 1989; Guyader *et al.*, 1989; Ogawa *et al.*, 1989; Cohen *et al.*, 1990; Hattori *et al.*, 1990; Shibata *et al.*, 1990) although phenotypes have been attributed to mutations in these genes. HIV-2 that is deficient in vpx grows normally on lymphocyte cell lines but has been shown to have a severe defect in its ability to propagate in primary peripheral blood lymphocytes, an effect that is enhanced at lower multiplicities of infection (Guyader *et al.*, 1989). Recent reports have confirmed these effects and have suggested that they also occur in peripheral blood mononuclear cells and primary macrophages (Kappes *et al.*, 1991; Yu *et al.*, 1991). The vpr proteins of both HIV-1 and HIV-2 are dispensable for growth in cultured and primary lymphocytes (Ogawa *et al.*, 1989; Cohen *et al.*, 1990; Hattori *et al.*, 1990). However, HIV-2 vpr deficient virus, like HIV-2 vpx deficient virus, is severely defective in growth on primary macrophages but appears to grow normally on primary lymphocytes (Hattori *et al.*, 1990). The replication kinetics of vpr mutant virus, when compared with those of wild type virus, are also delayed at low input multiplicities (Ogawa *et al.*, 1989). Hence it appears that the vpx and vpr proteins of HIV-2 group may be required for viral growth in specific types of primary blood cells, perhaps by performing similar functions in these cell types.

The SIVagm group was originally reported to encode a vpx protein by Fukasawa *et al.* (1988) on the basis of homology with the vpx proteins of the HIV-2 group. However, the similarity of all vpx and vpr proteins and their similar properties, which make them difficult to distinguish on functional grounds, calls this classification into doubt and we believe that the vpx proteins encoded by the SIVagm group should be reclassified as vpr proteins. It is obvious from Figure 1 that the vpx proteins encoded by SIVagm isolates are homologous to the HIV-2 vpx proteins but they also have similar amounts of amino acid identity to the vpr proteins of the other groups, and occasionally have a greater similarity to the vpr protein of a particular isolate than to its respective vpx protein. Thus simply scoring identical amino acids leaves the classification unclear.

To determine whether SIVagm vpx proteins were phylogenetically more similar to HIV-2 vpx proteins than to any of the prototype vpr proteins, we used the four-taxon test (Lake, 1987). This test measures the relative phylogenetic similarity of a particular gene to each of three other genes. Thus if two genes are clearly more closely related to each other than to the two other genes included in the test, then the topology clustering these genes will have a much higher score than either of the other two topologies. This method also allows the statistical likelihood of a particular topology to be calculated. This test failed to show that the SIVagmTYO *vpx* gene was significantly more likely to cluster with the prototype HIV-2 *vpx* gene when any two prototype *vpr* genes were also included. A phylogenetic tree of SIVagm vpx protein and vpr proteins was then constructed. The branch lengths of the major lineages were approximately equal which indicates similar rates of change in all lineages. Thus our phylogenetic studies indicate it is reasonable to consider the SIVagm group vpx protein as a vpr protein. All of these arguments undermine the original classification of the SIVagm *vpx* gene.

Reclassification of this gene as a *vpr* gene creates a simpler and more uniform picture for the genomic organization of the four primate lentiviral groups. All four groups would then contain the same basic set of genes, including *vpr*, which is not present in other mammalian lentiviruses. This allows us to define the genomic organization of the hypothetical primate lentiviral precursor; it probably contained five accessory genes—*tat*, *rev*, *vif*, *nef* and *vpr*. The SIVagm and SIVmnd groups still contain only this basic set. The HIV-1 group including SIVcpz has in addition a *vpu* gene. Finally, the HIV-2 group would then be unique in having a *vpx* gene, which arose by duplication of the *vpr* gene. Assuming on the grounds of parsimony that no other group gained and subsequently lost one of these genes, the duplication event must have occurred after the divergence of the HIV-2 group from the other primate lentiviruses. We calculated that the duplication event occurred at approximately the same time as the HIV-2 group diverged from the other primate lentiviral groups. This calculation was based on a divergence time of HIV-2 and SIVmac reported by Quérat *et al.* (1990). However, this conclusion still holds if data from other groups are used. For example, Sharp and Li (1988) estimated the HIV-2/SIVmac divergence time to be 30 years and using this figure we obtained a duplication date of ~130 years. This compares with an HIV-1/HIV-2 divergence date estimate of these authors of 150 years. This proposed timing of the duplication event explains the similarity between SIVagm vpx proteins and the vpx and vpr proteins from the HIV-2 group, as it appears that all were derived from a single precursor within a short period of time.

In summary, our studies indicate that it is reasonable to regard the SIVagm *vpx* gene as orthologous to the *vpr* genes of the other lentiviral groups but parologous to the HIV-2 group *vpx* genes. Calculation of the rates of evolution of *vpr* and *vpx* genes underscores their conservation relative to other lentiviral accessory genes and suggests the duplication of *vpr* in the HIV-2 group occurred close to the time that this set of viruses was diverging from the other groups.

Materials and methods

Sequence sources

Sequences were obtained from the AIDS research and human retroviruses database (Myers *et al.*, 1990) with the original sources as follows: SIVmac142 (Chakrabarti *et al.*, 1987); SIVmac239 (Regier and Desrosiers, 1990); SIVmac251 (Franchini *et al.*, 1987); SIVmacMNE (Henderson *et al.*, 1988); SIVsmmH4 (Hirsch *et al.*, 1989); SIVsmmPBJ (Dewhurst *et al.*, 1990); HIV-2D205 (Dietrich *et al.*, 1989); HIV-2ROD (Clavél *et al.*, 1986); HIV-2CAM2 (Tristem *et al.*, 1991); HIV-2ISY (Franchini *et al.*, 1989); HIV-2NIH-Z (Zagury *et al.*, 1988); HIV-2D194 (Kühnel *et al.*, 1989); HIV-2GH1 (Hasagawa *et al.*, 1989); HIV-2ST (Kumar *et al.*, 1990); HIV-2BEN (Klemm *et al.*, 1988); SIVagm155 (Johnson *et al.*, 1990); SIVagm3 (Baier *et al.*, 1989); SIVagmTYO (Fukasawa *et al.*, 1988); HIV-1OYI (Huet *et al.*, 1989); HIV-1HAN (Sauermann *et al.*, 1990); HIV-1NL43 (Adachi *et al.*, 1986); HIV-1MN and HIV-1SC (Gurgo *et al.*, 1988); HIV-1NDK (Spire *et al.*, 1989); HIV-1LAI (Wain-Hobson *et al.*, 19885); HIV-1SF2 (Sanchez-Pescador *et al.*, 1985); HIV-1MAL and HIV-1ELI (Alizon *et al.*, 1986); SIVcpz (Huet *et al.*, 1990); SIVmndGB1 (Tsujimoto *et al.*, 1989).

Amino acid alignments and sequence analysis

Amino acid sequences of vpx and vpr proteins were assembled using a homologous sequence editor (HOMED) database (Stockwell and Petersen, 1987). A pairwise comparison of the sequences was then performed with

a multiple alignment program (AMPS) utilizing the Dayhoff mutation matrix (Dayhoff *et al.*, 1978; Barton and Sternberg, 1987). The sequences were subjected to multiple alignment using the following parameters: gap penalty, 8; constant, 8. The resulting alignment was then studied by eye and a small number of changes were made to improve it (thereby reducing the total number of steps required to generate the observed variation). Sequence similarity was calculated using the program SIP from the STADEN package (Staden, 1990) with a gap opening penalty of 10 and a gap extension penalty of 9 or 10.

The statistical significance of amino acid similarity detected by the multiple alignment was determined using the RDF2 and RSS programs from the FASTA package (Pearson and Lipman, 1988). Pairs of vpr and vpx proteins were aligned and one of the sequences was then subjected to at least 100 shuffles. Both a global and a local shuffle (with a window of 10) were performed.

Protein and nucleic acid databases were searched for the presence of related sequences using FASTA, tFASTA, PSQ and by PROFILE analysis (Gribskov *et al.*, 1987; Pearson and Lipman, 1988; George and Barker, 1990). PSQ allows certain residues to be specified as invariant (protein sequences are therefore scanned for the presence of specific residues at defined positions), whereas PROFILE analysis uses the position dependent scoring of amino acids based on their presence in a representative set of aligned sequences.

Phylogenetic analysis

The program EVOMONY (Lake, 1987) was used to determine the most probable topologies of trees of the prototype *vpr* and *vpx* genes from each of the four primate lentiviral groups. Phylogenetic trees constructed from four sequences have three possible unrooted topologies and EVOMONY can be used to determine the statistical likelihood of a particular topology being correct. Sequences were aligned, back translated into DNA and were then subjected to four-taxon tests using transversion parsimony and evolutionary parsimony.

Phylogenetic trees were also constructed from multiple sequences using the alignment generated by AMPS and the protein sequence parsimony program (PROTPARS) from the PHYLIP package (Felsenstein, 1990). Sequences were first shortened by deleting any position that contained a gap, whereas those adjudged to have premature termination codons (HIV-2 BEN vpr, HIV-2 D205 vpr and SIVmac251 vpr) were padded (these residues are then ignored by the program and hence do not affect the branch orders of the phylogenetic trees). Multiple runs were performed in which sequences were added to the tree in a different order. The SIVmndGB1 sequence, or a random sequence was used as the root. The trees produced were combined using the program CONSENSE (Felsenstein, 1990), which selects the most frequently occurring branching patterns for inclusion into one consensus tree. Branch lengths were calculated using the output from CONSENSE and the program PARCOUNT to correct for the shortened branch lengths produced by this method.

Gene divergence rates

The rates of evolution of the *vpr* and *vpx* genes within the HIV-2 group were calculated using the method of Nei and Gojobori (1986) and by using the DNADIST program from the PHYLIP package. The first method enables the number of non-synonymous and synonymous substitutions to be calculated separately and therefore allows comparison with rates obtained previously with other lentiviral genes. The nucleic acid sequences corresponding to several *vpr* and *vpx* genes from the HIV-2 group were aligned using the output from the amino acid sequence alignment generated by AMPS. Positions containing gaps in one or more of the sequences were deleted and the total number of synonymous and non-synonymous sites for each sequence were calculated. Pairs of *vpx* or *vpr* sequences were then examined codon by codon and the numbers and types (synonymous or non-synonymous) of substitution between them assessed. The proportion of substitutions per site was then corrected for multiple and back mutations before the rate of evolution for each pair of genes was calculated.

Dating duplication events

The elapsed time since a duplication event is calculated from the average evolutionary distance between pairs of parologous genes and the average rate of evolution of the orthologous genes (Li and Graur, 1990). Evolutionary distances between pairs of parologous (vpr/vpx) genes were much higher than those between pairs of orthologous (vpr/vpr, vpx/vpx) genes. Thus evolutionary distances were calculated using the program DNADIST rather than by a manual method.

Acknowledgements

The authors wish to thank G.Myers for sending a copy of the AIDS research and human retroviruses database and its updates; P.Sharp and E.Holmes for reviewing an earlier version of the manuscript and J.Felsenstein and A.Lesk for some of the software used. We also thank M.Gribskov, J.Lake, W.Pearson, J.Petrik and J.Sinsheimer for advice. This work was supported in part by a Medical Research Council AIDS directed programme grant. M.T. thanks Amersham International for a research studentship. C.M. was the recipient of an MRC(NZ)–Wellcome Overseas Fellowship.

References

Adachi,A., Gendelman,H.E., Koenig,S., Folks,T., Willey,R., Rabson,A. and Martin,M.A. (1986) *J. Virol.*, **59**, 284–291.

Alizon,M., Wain-Hobson,S., Montagnier,L. and Sonigo,P. (1986) *Cell*, **46**, 63–74.

Baier,M., Werner,A., Cichutek,K., Garber,C., Muller,C., Kraus,G., Ferdinand,F.J., Hartung,S., Papas,T.S. and Kurth,R. (1989) *J. Virol.*, **63**, 5119–5123.

Barré-Sinoussi,F., Chermann,J.C., Rey,F., Nugeyre,M.T., Chamaret,S., Gruest,J., Dauguet,C., Axler-Blin,C., Brun-Vézinet,F., Rouzioux,C., Rozenbaum,W. and Montagnier,L. (1983) *Science*, **220**, 868–870.

Barton,G.J. and Sternberg,M.J.E. (1987) *J. Mol. Biol.*, **198**, 327–337.

Chakrabarti,L., Guyader,M., Alizon,M., Daniel,M.D., Desrosiers,R.C., Tiollais,P. and Sonigo,P. (1987) *Nature*, **328**, 543–547.

Clavél,F., Guyader,M., Guetard,D., Salle,M., Montagnier,L. and Alizon,M. (1986) *Nature*, **324**, 691–695.

Cohen,E.A., Dehni,G., Sodroski,J.G. and Haseltine,W.A. (1990) *J. Virol.*, **64**, 3097–3099.

Daniel,M.D., Letvin,N.L., King,N.W., Kannagi,M., Sehgal,P.K., Hunt,R.D., Kanki,P.J., Essex,M. and Desrosiers,R.C. (1985) *Science*, **228**, 1201–1204.

Dayhoff,M.O., Schwartz,R.M. and Oncutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, Vol. 5, Suppl. 3, pp. 345–352.

Dedera,D., Hu,W., Vander Heyden,N. and Ratner,L. (1989) *J. Virol.*, **63**, 3205–3208.

Desrosiers,R.C. (1990) *Nature*, **345**, 288–289.

Dewhurst,S., Embretson,J.E., Anderson,D.C., Mullins,J.I. and Fultz,P.N. (1990) *Nature*, **345**, 636–640.

Dietrich,U., Adamski,M., Kreutz,R., Seipp,A., Kühnel,H. and Rübsamen-Waigmann,H. (1989) *Nature*, **342**, 948–950.

Doolittle,R.F., Feng,D.F., Johnson,M.S. and McClure,M.A. (1989) *Q. Rev. Biol.*, **64**, 1–30.

Emau,P., McClure,H.M., Isahakia,M., Else,J.G. and Fultz,P.N. (1991) *J. Virol.*, **65**, 2135–2140.

Felsenstein,J.A. (1990) *PHYLIP Manual Version 3.3*. University Herbarium, University of California, Berkeley, CA.

Franchini,G., Gurgo,C., Guo,H.-G., Gallo,R.C., Collati,E., Fargnoli,K.A., Hall,L.F., Wong-Staal,F. and Reitz,M.S., Jr (1987) *Nature*, **328**, 539–543.

Franchini,G., Rusche,J.R., O'Keeffe,T.J. and Wong-Staal,F. (1988) *AIDS Res. Hum. Retroviruses*, **4**, 243–250.

Franchini,G., Fargnoli,K.A., Giombini,F., Jagodzinski,L., De Rossi,A., Bosch,M., Biberfeld,G., Fenyo,E.M., Albert,J., Gallo,R.C. and Wong-Staal,F. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2433–2437.

Fukasawa,M., Muira,T., Hasegawa,A., Morikawa,S., Tsujimoto,H., Miki,K., Kitamura,A.T. and Hayami,M. (1988) *Nature*, **333**, 457–461.

Fultz,P.N., McClure,H.M., Anderson,D.C., Brent-Swenson,R., Anand,R. and Srinivasan,A. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 5286–5290.

Garvey,K.J., Oberste,M.S., Elser,J.E., Braun,M.J. and Gonda,M.A. (1990) *Virology*, **175**, 391–409.

George,D.G. and Barker,W.C. (1990) *Users Guide. Protein Identification Resource*. National Biomedical Research Foundation, Washington DC 20007, USA.

Gojobori,T., Moriyama,E.N., Ina,Y., Ikeo,K., Muira,T., Tsujimoto,H., Hayami,M. and Yokoyama,S. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4108–4111.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.

Gurgo,C., Guo,H.-G., Franchini,G., Aldovini,A., Collalti,E., Farrell,K., Wong-Staal,F., Gallo,R.C. and Reitz,M.S., Jr (1988) *Virology*, **164**, 531–536.

Guyader,M., Emerman,M., Sonigo,P., Clavél,F., Montagnier,L. and Alizon,M. (1987) *Nature*, **326**, 662–669.

Guyader,M., Emerman,M., Montagnier,L. and Pedan,K. (1989) *EMBO J.*, **8**, 1169–1175.
Hasagawa,A., Tsujimoto,H., Maki,N., Ishikawa,K.-I., Muira,T., Fukasawa,M., Miki,K. and Hayami,M. (1989) *AIDS Res. Hum. Retroviruses*, **5**, 593–604.
Hattori,N., Michaels,F., Fargnoli,K., Marcon,L., Gallo,R.C. and Franchini,G. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 8080–8084.
Henderson,L.E., Benveniste,R.E., Sowder,R., Copeland,T.D., Schultz,A.M. and Oroszlan,S. (1988a) *J. Virol.*, **62**, 2587–2595.
Henderson,L.E., Sowder,R.C., Copeland,T.D., Benveniste,R.E. and Oroszlan,S. (1988b) *Science*, **241**, 199–201.
Hirsch,V.M., Olmstead,R.A., Murphey-Corb,M., Purcell,R.H. and Johnson,P.R. (1989) *Nature*, **339**, 389–391.
Hopp,T.P. and Woods,K.R. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.
Huet,T., Dazza,M.-C., Brun-Vézinet,F., Roelants,G.E. and Wain-Hobson,S. (1989) *AIDS*, **3**, 707–715.
Huet,T., Cheynier,R., Meyerhans,A., Roelants,G. and Wain-Hobson,S. (1990) *Nature*, **345**, 356–359.
Johnson,P.R., Fomsgaard,A., Allan,J., Gravell,M., London,W.T., Olmstead,R.A. and Hirsch,V.M. (1990) *J. Virol.*, **64**, 1086–1092.
Kappes,J.C., Morrow,C.D., Lee,S.-W., Jameson,B.A., Kent,S.B.H., Hood,L.E., Shaw,G.M. and Hahn,B.H. (1989) *J. Virol.*, **62**, 3501–3505.
Kappes,J.C., Conway,J.A., Lee,S.-W., Shaw,G.M. and Hahn,B.H. (1991) *Virology*, **184**, 197–209.
Kawakami,T., Sherman,L., Dahlberg,J., Gazit,A., Yaniv,A., Tronick,S.R. and Aaronson,S.A. (1987) *Virology*, **158**, 300–312.
Klemm,E., Schneweis,K.E., Horn,R., Tackmann,W., Schulze,G. and Schnieder,J. (1988) *J. Neurol.*, **235**, 304–307.
Kühnel,H. *et al.* (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2383–2387.
Kumar,P., Hui,H., Kappes,J.C., Haggarty,B.S., Hoxie,J.A., Arya,S.K., Shaw,G.M. and Hahn,B. (1990) *J. Virol.*, **64**, 890–901.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lake,J.A. (1987) *Mol. Biol. Evol.*, **4**, 167–191.
Li,W.-H. and Graur,D. (1990) *Fundamentals of Molecular Evolution*. Sinauer Associates, MA, pp. 149–151.
Myers,G., Berzofsky,J.A., Rabson,A.B., Smith,T.F. and Wong-Staal,F. (1990) *Human Retroviruses and AIDS*. Los Alamos Laboratory, Los Alamos, USA.
Nei,M. and Gojobori,T. (1986) *Mol. Biol. Evol.*, **3**, 418–426.
Ogawa,K., Shibata,R., Kiyomasu,T., Higuchi,I., Kishida,Y., Ishimoto,A. and Adachi,A. (1989) *J. Virol.*, **63**, 4110–4114.
Olmstead,R.A., Hirsch,V.M., Purcell,R.H. and Johnson,P.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8088–8092.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.
Quérat,G., Audoly,G., Sonigo,P. and Vigne,R. (1990) *Virology*, **175**, 434–447.
Regier,D.A. and Desrosiers,R.C. (1990) *AIDS Res. Hum. Retroviruses*, **6**, 1221–1231.
Saltarelli,M., Quérat,G., Konigs,D.A., Vigne,R. and Clements,J.E. (1990) *Virology*, **179**, 347–364.
Sanchez-Pescador,R. *et al.* (1985) *Science*, **227**, 484–492.
Sauermann,U., Schnieder,J., Mous,J., Brunckhorst,U., Schedel,I., Jentsch,K.D. and Hunsmann,G. (1990) *AIDS Res. Hum. Retroviruses*, **6**, 813–823.
Sharp,P.M. and Li,W.-H. (1988) *Nature*, **336**, 315.
Shibata,R., Muira,T., Hayami,M., Ogawa,K., Sakai,H., Kiyomasu,T., Ishimoto,A. and Adachi,A. (1990) *J. Virol.*, **64**, 742–747.
Sonigo,P., Alizon,M., Staskus,K., Klatzmann,D., Cole,S., Danos,O., Retzel,E., Tiollais,P., Haase,A. and Wain-Hobson,S. (1985) *Cell*, **42**, 369–382.
Spire,B., Sire,J., Zachar,V., Rey,F., Barré-Sinoussi,F., Galibert,F., Hampe,A. and Chermann,J.-C. (1989) *Gene*, **81**, 275–284.
Staden,R. (1990) *Comput. Appl. Biosci.*, **6**, 387–393.
Stockwell,P.A. and Petersen,G.B. (1987) *Comput. Appl. Biosci.*, **3**, 37–43.
Tristem,M., Marshall,C., Karpas,A., Petrik,J. and Hill,F. (1990) *Nature*, **347**, 341–342.
Tristem,M., Hill,F. and Karpas,A. (1991) *J. Gen. Virol.*, **72**, 721–724.
Tsujimoto,H., Hasagawa,A., Maki,N., Fukasawa,M., Muira,T., Spiedel,S., Cooper,R.W., Moriyama,E.N., Gojobori,T. and Hayami,M. (1989) *Nature*, **341**, 539–541.
Wain-Hobson,S., Sonigo,P., Danos,O., Cole,S. and Alizon,M. (1985) *Cell*, **40**, 9–17.
Wong-Staal,F., Chanda,P.K. and Ghrayeb,J. (1987) *AIDS Res. Hum. Retroviruses*, **3**, 33–39.
Yokoyama,S., Chung,L. and Gojobori,T. (1988) *Mol. Biol. Evol.*, **5**, 237–251.
Yu,X.-F., Ito,S., Essex,M. and Lee,T.-H. (1988) *Nature*, **335**, 262–265.
Yu,X.-F., Matsuda,M., Essex,M. and Lee,T.-H. (1990) *J. Virol.*, **64**, 5688–5693.
Yu,X.-F., Yu,Q.-C., Essex,M. and Lee,T.-H. (1991) *J. Virol.*, **65**, 5088–5091.
Yuan,X.-F., Matsuda,Z., Essex,M and Lee,T.-H. (1990) *AIDS Res. Hum. Retroviruses*, **6**, 1265–1271.
Zagury,J.F. *et al.* (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 5941–5945.

Received on May 14, 1992