FROM: 231. **Hansen CN, Heslop-Harrison JS. 2004** . Sequences and phylogenies of plant pararetroviruses, viruses and transposable elements. *Advances in Botanical Research* **41** : 165-193.

Sequences and Phylogenies of Plant Pararetroviruses, Viruses and Transposable Elements

CELIA HANSEN AND JS HESLOP-HARRISON*
DEPARTMENT OF BIOLOGY
UNIVERSITY OF LEICESTER
LEICESTER LE1 7RH, UK

*AUTHOR FOR CORRESPONDENCE
E-MAIL: PHH4@LE.AC.UK
WEBSITE: WWW.MOLCYT.COM

DNA elements found within cells that include the enzyme reverse transcriptase can be brought together within a unified taxonomic framework. The framework includes retroelements that are components of the nuclear genome, and recognized viruses where nuclear integration is unknown, occurs occasionally or is frequent. The classification probably has a natural basis and reflects aspects of the evolution and phylogeny of the elements. Complete retroelements and retroviruses include two or more open reading frames (ORFs) that encode single proteins or polyproteins. The order of the genes in the elements varies. In recent years, it has been shown that pararetroviruses can be integrated in the plant genome and evidence indicates they can be transcribed to give infectious virus. In this review, we show scale alignments of genomes from the six taxonomic families of reverse-transcribing viruses, gypsy and copia-like retroelements and LINEs, and also the enzyme telomerase, to show the lengths of the elements and the order of genes. We also show amino-acid alignments and key conserved residues or domains within the reverse transcriptase(RT), RNase H (RH), integrase (INT) and aspartic protease (PR) genes and in a conserved cysteine-histidine (CH) zinc-finger-like domain. A unified classification of reverse-transcribing elements is useful for phylogenetic and taxonomic purposes and understanding their contribution to plant genome function and evolution.

I. INTRODUCTION

A. PLANT GENOME ORGANIZATION

It has long been established that genomes contain, in addition to genes and regulatory sequences, various classes of tandemly or dispersed repetitive DNA sequences each with characteristic chromosomal locations (review: Schmidt and Heslop-Harrison, 1998). Some of this repetitive DNA is structural, such as the repeat motifs found at the telomeres, centromeres and secondary constrictions. Other significant components of the genome are the transposable elements present in all species of bacteria, animals, fungi and plants, with a copy number of hundreds to millions in most species (Flavell *et al.*, 1997). Transposable elements can be divided into two major types: firstly, the class (sometimes type) I transposable elements, retroelements including long terminal repeat (LTR) and non-LTR retrotransposons, and secondly, the DNA transposable elements, DNA transposons or class II transposable elements, which have the capacity to excise themselves and reintegrate elsewhere in the genome. The retroelements replicate via an RNA intermediate using reverse transcriptase (RT), leaving the original element in the genome, and have the potential to insert a new copy of the element in the DNA. These properties of mobility and replication allow transposable elements (typically 2 to 15 kb in length) to be one of the most dynamic and rapidly evolving components of the genome, often being dispersed over much of the chromosome's length. These class I transposable elements are closely related to some viruses, many having viral properties. As suggested by Hull (1999a, 2001), the retroelements can be integrated into a common taxonomic system with a basis in evolutionary phylogeny.

This paper does not repeat the comprehensive reviews of Kunze *et al.* (1997) or Kumar and Bennetzen (1999) which describe the diversity and nature of DNA transposons and retroelements. It will focus on the understanding of the relationship between retroelements, their similarities with virus sequences, reviewing key conserved domains and amino acid residues, and also discuss the presence of integrated sequences of the family of plant pararetroviruses, an emerging subject.

B. RETROELEMENTS IN THE GENOME

To many researchers, the proportion of the genome represented by transposable elements, and their recognizable but degenerate derivatives, has come as a surprise. In the human genome sequence (some 3000 Mbp long), transposable elements account for 45% of the genomic DNA (International Human Genome Consortium, 2001), and in the mouse (2,500 Mbp) they account for 38% of the genome (Mouse Genome Sequencing Consortium, 2002). In the small genome plant species sequenced to date; *Arabidopsis thaliana* (145 Mbp) and rice (430 Mbp), the transposable elements account for between 10% (The Arabidopsis Genome initiative, 2000) and 18% (Feng *et al*., 2002; Sasaki *et al*., 2002). These plant species are not representative for plants in general as the proportion of transposable elements is higher in species with larger genomes. Complete genome sequencing projects have difficulties tackling large intergenic or repeat regions where retroelements may be abundant (Brandes *et al.* 1997; Tikhonov *et al*., 1999; Barakat *et al*., 1997; The Arabidopsis Genome Initiative, 2000). In the rice draft genome, assembled contigs (361 Mbp) had 16% transposable elements, while fully masked reads which could not be integrated into the complete sequence (78 Mbp) included 59% transposable elements (Yu *et al*., 2002). It is notable that in the unassembled fully masked reads, 97% of transposable elements are retroelements, while another group of retroelement-related sequences, miniature inverted-repeat tandem elements (MITEs) represented only 1%; in contrast, these repeat classes accounted for 42% and 40% respectively in the assembled contigs. Since much single-copy-rich DNA is in the assembled regions, the difference indicates that retroelements are mostly in intergenic heterochromatic regions and that MITEs insert preferentially near genes (Yu *et al.*, 2002). The same may be the case for the number of different transposable elements found in the *Arabidopsis* genome.

In mammals and plants, the largest portion of the transposable elements in the genome are usually retrotransposons: in mammals, non-LTR LINE elements are most abundant, while in the plants the largest portion is made up of LTR *copia* and *gypsy* elements (International Human Genome Sequencing Consortium, 2001; The Arabidopsis Genome Initiative, 2000; Sasaki *et al.*, 2002; Feng *et al.*, 2002). The genome-integrated retrotransposons have been recognized as a major evolutionary force in the host genome, which can be very diverse organisms from bacteria and yeasts to plants and animals, both because of their abundance and the effects of insertion in or near to genes and regulatory sequences. Nevertheless, most retroelements have no known phenotypic effect on the host, although their insertion into the genome can disrupt gene expression, and transcribed copies in the form of viruses give severe illnesses in mammals, including diseases caused by retroviruses (e.g. *Human immunodeficiency virus*, HIV) and pararetroviruses (e.g. *Hepatitis B virus*, HBV).

C. Reverse transcriptase

Retroelements are characterized by the presence of a gene encoding reverse transcriptase (RT), RNA-directed DNA polymerase, which is capable of using an RNA template to make a complementary DNA molecule, thus allowing their autonomous amplification via an RNA intermediate transcribed from the DNA form using cellular RNA polymerase. RT was discovered by Baltimore (1970) and Temin and Mizutani (1970), and is considered to be an ancient and widespread enzyme. Support for its early origin comes from the similarity of extant RT enzymes across viruses, prokaryotes and eukaryotes. RT has domains in common with the RNA-directed RNA polymerase of RNA viruses, suggesting that they share an ancient common ancestor (Xiong and Eickbush, 1990). The RNA viruses are believed to be at least as old as retroelements as they have a greater diversity and are present in many prokaryotes and eukaryotes. The history of retroelements may well coincide with the origin of a DNA based life form some 3.5 billion years ago. At the earliest stages of life it is likely that RNA genes were converted to DNA which then provided the basis for subsequent evolution. This suggests that an RNA-dependent DNA polymerase – reverse transcriptase – was an early and critical enzyme in the origin of DNA-based organisms (see Heslop-Harrison, 2000). While widespread distribution of retroelements is most likely to be explained by their single origin and evolutionary descent into virtually all modern organisms (both prokaryotes and eukaryotes), it is probable that the cross-species transfer of sequences, either as DNA or RNA (horizontal transfer), and perhaps the convergent evolution of sequences, has contributed to the extant distribution of different retroelement types (Brown, 2003).

D. Viruses

Viruses have well-recognized properties of gene expression and replication in host cells, but there is no concise universally accepted definition of a virus. The International Committee on Viral Taxonomy describes a virus as "an elementary biosystem that possesses some of the properties of living systems such as having a genome and being able to adapt to changing environments. However, viruses cannot capture and store free energy and they are not functionally active outside their host cells" (quoted by Hull, 2001). Until the 1990s, the viruses were named individually using physical and biological properties based on the symptoms they cause, their host range, replication strategy, particle structure and, to some extent, biochemical composition. The nature of the viral genome – whether DNA or RNA, single or double-stranded – has been used to categorize viruses for many years, and both the size of the genomic nucleic acid and its sequence are now important characters in classification which have allowed some grouping of individuals. Now, viral taxonomy is stabilizing with most viruses fitting into larger taxonomic groupings having a natural basis related to phylogeny (Hull, 1999a, 2001, 2002; Buchen-Osmond, 2003; ICTV, 2003), at least at levels that have been named as family and suborder levels. As pointed out by Hull (2001), genomic retroelements fit the ICTV definition of a virus, and based on their common features can be fitted into the structure of a natural phylogeny.

II. RETROELEMENTS

Nuclear DNA elements that include the gene RT can be regarded as 'retroelements', and Hull (1999a, 2001) proposed a unified classification for reverse transcribing elements that includes viruses and transposable elements with RT. The elements can be divided into retroviruses, pararetroviruses and the abundant group of nuclear sequences, the retrotransposons, including the long terminal repeat (LTR) retrotransposons, non-LTR retroposons and group II mitochondrial introns (Hull, 1999a, 2001; Fig. 1). Another important enzyme, telomerase (Blackburn, 1992), which adds the telomere sequences to most eukaryotic chromosomes, also incorporates a RT function (Lingner *et al.*, 1997) and can be aligned with other sequences (Fig. 3).

A. VIRAL RETROELEMENTS – *RETROVIRALES*

The group of "DNA and RNA reverse transcribing viruses" (Pringle, 1999) includes the *Retrovirales* (Hull, 2001) and consists of elements potentially capable of infection such as the retroviruses (RNA genome) and the pararetroviruses (DNA genome) (Fig. 1) and those with no known infectivity such as *copia* and *gypsy* retrotransposons. The vertebrate retroviruses of the suborder *Orthoretrovirineae* have an RNA genome in the infective form and are transcribed into DNA by RT, which is then integrated into the nuclear genome of the host with the assistance of the encoded integrase (Table 1). The suborder *Pararetrovirineae* (pararetroviruses), found in both plant and animal kingdoms, encapsulates a double-stranded (ds) DNA genome and replicate through an RNA intermediate; no integrase function is detected in their genome and integration is not an obligatory part of their replication, infection and transmission cycle (Hull and Covey, 1996). In Hull's (1999a) classification two families of pararetroviruses are given, the animal viruses of the *Hepadnaviridae* (two genera) and the plant viruses of the *Caulimoviridae*, including six genera: *Badnavirus*, *Caulimovirus*, and four genera represented by a small number of individual viruses, *Tungrovirus* (*Rice tungro bacilliform tungrovirus*, RTBV), *Petuvirus* (*Petunia vein clearing petuvirus*, PVCV), *Soymovirus* (*Soybean chlorotic mottle soymovirus*, SbCMV) and *Cavemovirus* (*Cassava vein mosaic cavemovirus*, CsVMV; *Tobacco vein clearing cavemovirus*, TVCV) (ICTV; Pringle, 1999).

No retrovirus *senso stricto* has yet been found in plants although certain characteristic genes, putative *envelope* or transit proteins (see section III A. below), have been identified in some *gypsy*-like and *copia*-like elements such that they have characteristics of retroviruses (Petropoulos, 1997; Kumar, 1998; Wright and Voytas, 2002; Laten *et al.*, 1998). In the classification (Fig. 1), the suborder *Retrotransposineae*, including the *Pseudoviridae* (Ty1-*copia* group) and the *Metaviridae* (Ty3-*gypsy* group), have been placed under *Retrovirales*. These elements can form virus-like particles although have no known viral infectivity. *Retrotransposineae* may have one or two open reading frames (ORFs) containing *gag* and *pol* genes bordered by LTRs, sometimes with a third ORF present. Structurally *copia* and *gypsy* differ in the order of encoded genes (see Table I and Fig. 3). *Retrotransposineae* have a replication strategy similar to that of *Orthoretrovirineae* where integration of a new copy into DNA is an obligatory part of the replication cycle, although they have no encoded features enabling them to move from cell to cell.

B. Non-viral retroelements – *Retrales*

The order *Retrales*, with the suborder of *Retroposineae*, has fewer similarities with infective viruses than the *Retrotransposineae* suborder, although some genes and the organization of the genes have relationships (Figs 1, 2, 3). The *Retroposineae* includes the non-LTR elements LINEs and their truncated derivatives SINEs (Figs 1, 3; Table I): LINEs are simpler structures than *Retrotransposineae* but contain many common functional properties including *gag* and *pol* and an endonuclease function. Included are also the suborder *Retronineae* containing the group II mitochondrial introns.

III. VIRAL AND NON-VIRAL ELEMENTS

A. Between retrotransposon and virus - the *envelope* gene

The *envelope* gene (*env*), or the related coding sequence known as the movement protein (MP) or transit peptide, gives a transcribed DNA element the ability to move with a high frequency between cells and become infective. Although the *envelope* gene is not well conserved in primary sequence, both viral and putative retrotransposon envelope proteins share structural similarities. They are typically translated from spliced mRNAs and the primary product encodes a signal peptide and a transmembrane domain near the C terminus (Wright and Voytas, 1998, 2002; Chavanne *et al*., 1998; Vicient *et al*., 2001).

Malik *et al*. (2000) proposed that a non-viral ancestor to errantiviruses (*Metaviridae*, *Drosophila* specific *gypsy*-like virus) acquired the *envelope* gene from another family of double-stranded DNA insect virus, the *Baculoviridae*, as the *envelope* gene from these two insect viruses was found to share sequence features. Furthermore, baculoviruses were found to harbour inserts of LTR retrotransposons, which could be a step in the acquisition of an *envelope* gene by the latter. There are at least eight cases of *envelope*(-like) gene acquisition in the broad group of retroelements: *Sire1* from the *copia* group; *Athila*, *Cyclops*, *Osvaldo*, *Cer*, *Tas*, and errantiviruses from the *gypsy* group. Vertebrate retroviruses and the family of plant caulimoviruses with *envelope* genes may also have arisen from groups without the gene, perhaps acquiring it by fusion of an LTR-retrotransposable element with a plant virus (Malik *et al*., 2000; Chavanne *et al*., 1998). Alternatively, transposable elements could be remnants of infectious viruses which have lost most of the *envelope* gene: perhaps the gene is less useful in plants compared to animals as cell walls might be an obstacle to membrane-membrane fusions allowing a virus to enter a cell (Bennetzen, 2000).

B. Pararetroviruses in plant genomes

Vertebrate retroviruses have long been known to integrate into the nuclear genome of their host at a stage of their replication cycle (Löwer, 1999; Benit *et al*., 1999; Herniou *et al*., 1998). For example, the mammalian hepatitis B virus (HBV) is spontaneously and illegitimately integrated into nuclei in cancerous and pre-cancerous liver cells (see Pineau *et al*., 1996; Tagieva *et al*., 1995; Wang *et al*., 2001). However, until recently, plant pararetroviruses have been considered as independent particles in the host genome. The ds DNA form of plant pararetroviruses are infective, causing a range of vein chlorosis, ring-spot, mosaic and mottling symptoms. Most pararetroviruses have a narrow host range: CaMV rarely infects species outside the Brassicaceae, and the genera of caulimoviruses in general infect dicotyledons. Caulimoviruses infect most tissues in the host plant, but badnaviruses may be restricted to the vascular tissue (Hohn and Fütterer, 1997). The pararetrovirus particle moves between cells via plasmodesmata and between individual plants by insect transmission: usually *Caulimovirus* by aphids, and *Badnavirus* by mealy bugs (Hull and Covey, 1996).

In the mid-1990s, analysis of the epidemiology of some plant virus diseases revealed an unexplained spread, previously not noticed or explained by asymptomatic and low levels of chronic infection. Bananas (*Musa*) can be infected by *Banana streak badnavirus*, BSV (see Dahal *et al*., 1998; Harper and Hull, 1998; Harper *et al*., 1999), which causes disease throughout tropical regions. However, the appearance of symptoms did not always correlate with the presence of infected plants or insect vectors in the field, and infection was pronounced in plants that had been stressed. In particular, plants from tissue culture of certain varieties (Dahal *et al*., 2000), and plants exposed to low night temperatures showed symptoms. PCR amplification using primers from within the sequence of BSV, *in situ* hybridization to nuclear chromosomes of the *Musa* accessions using BSV fragments as probes (Harper *et al*., 1999), and genomic library screening (Ndowora *et al*., 1999) indicated that there was a sequence homologous to BSV integrated in the nuclear DNA of these *Musa* varieties. *In situ* hybridization to nuclear DNA stretched to its full molecular length showed that the integrated BSV sequence was repeated in two different structures of 150 kb and 50 kb respectively (Harper *et al*., 1999). It is believed that sexual hybridization, tissue culture and other stress can generate episomal viruses by recombination of the integrated sequence. Geering *et al*. (2000, 2001) found that there was variability in the type of BSV-like sequence integrated in the genome of *Musa,* and remnants of other BSV sequences are found in both A and B genome *Musa*. As a consequence of the integration of BSV sequences into the *Musa* genome, consideration and care is needed regarding the safe movement of germplasm and methods of plant breeding and tissue culture (Harper and Hull, 1998).

Evidence from epidemiology and molecular biology suggests that, as in *Musa*, there is a possibility that other plant species also include viral sequences that can be expressed and give rise to episomal viruses and infection. Integrated PVCV, sequences have been detected in *Petunia hybrida* (Richert-Pöggeler *et al*., 1996), and *in situ* hybridization indicates that the sequences are concentrated at relatively few sites (Richert-Pöggeler *et al*., 2003). There is evidence that a complete PVCV genome is present in one *Petunia* cultivar and that at least part of the viral genome is present in many cultivars (Harper *et al*., 2002). The presence of the integrated virus sequence is correlated with the appearance of disease symptoms and virus particles in some *P. hybrida* varieties, again under particular environmental conditions.

The allohexaploid *Nicotiana edwardsonii* was formed by the hybridization between *N. clevelandii* (female, 4x) and *N. glutinosa* (male, 2x). In *N. edwardsonii*, the spontaneous appearance of episomal virus particles (TVCV) was discovered under certain environmental conditions. Southern hybridization of TVCV sequences to genomic DNA of *N. edwardsonii* and *N. glutinosa* showed that TVCV was integrated in the nuclear DNA (Lockhart *et al.,* 2000). It is possible that the expression of episomal TVCV in *N. edwardsonii* was triggered by the rearrangement of otherwise deficient integrants during the interspecific hybridization and the subsequent chromosome doubling.

In a study of DNA flanking transgenes, Jakowitsch *et al.* (1999) sequenced regions of nuclear DNA with high homology to a pararetrovirus from *N. tabacum*. From these fragments it was possible to assemble a hypothetical 7981 bp pararetrovirus-like (PRV-L) genome called here tprv (although named TPV for tobacco pararetrovirus by the authors; ICTV recognizes TPV as the geminivirus Texas Pepper Virus). tprv is most closely related to TVCV (75%) and CsVMV (42%) at the nucleotide level and has the same genomic structure as TVCV. In this case, no expression of a virus has been detected.

Within the same plant family, Solanaceae, pararetrovirus-like sequences with nuclear integration have also been found in tomato and potato. Budiman *et al.* (2000), as part of a tomato genome sequencing project, generated a sequence-tagged connector (STC) framework from a BAC library of *Lycopersicon esculentum*. As might be expected, *copia* and *gypsy*-like retrotransposon sequences were abundant, but PRV-L sequences were also detected in the sequence data. Hansen *et al.* (2003) detected several families of a PRV-L sequence in potato (*Solanum tuberosum*) using PCR primers designed to pararetrovirus consensus sequences, isolating genomic DNA fragments and characterization by sequencing, Southern and *in situ* hybridization.

Genomic sequence data indicate the presence of PRV-L sequences in the rice genome. *Rice tungro bacilliform tungrovirus* (RTBV) is widespread in South East Asia and causes substantial losses in rice production. The genome is about 8000 bp long, with four open reading frames (Hull, 1999b). In the genomic rice sequence (Sasaki *et al.,* 2002), fragments related to all four ORFs are found, with sequences of the RT-RNase H region and part of the movement protein (EMBL sequence number AP000559, Sasaki *et al.*, 1999, unpublished data). Another survey of rice detected three PRV-L fragments related to RTBV (Mao *et al.*, 2000).

IV. RELATIONSHIPS BETWEEN RETROELEMENTS

Whether infectious or not, sequences classified as retroelements have common features which can be used to analyse evolutionary relationships. The characteristic shared RT region, as a defining feature, can be used to align the sequences, while the presence, order and sequence of other conserved regions allows further comparison. Lerat *et al.* (1999) discuss the possible ‘modular’ evolution of the conserved functional blocks with retroelements.

Xiong and Eickbush (1990) analysed the full RT region of a wide selection of retroelements, identifying seven common peptide regions (domains 1-7) containing 178 amino acids with chemically similar residues within the majority of the 82 RT sequences analysed. They rooted their phylogenetic tree with the RNA-directed RNA polymerase from RNA viruses (see Fig. 2). Based on the analysis, it was suggested that the ancestral retrotransposable element had a *gag* gene and a *pol* gene, either as two separate ORFs or one large ORF and no LTRs. Hepadnaviruses and non-LTR retrotransposons become the first branches on the tree (Fig. 2), and branches of the Hepadnaviruses and Caulimoviruses include a fragment of *pol* gene containing the RT-RNase H domain. The retroviruses may represent a retroelement which acquired an *envelope* gene making it possible to be transmitted between cells. For the retroelements of bacteria and organelles they considered the possibility that the RT region was captured by functional bacterial introns, or organelle genomes or plasmids (Xiong and Eickbush, 1990).

Figure 3 illustrates the structures of representative elements from the five families of viruses shown in Figure 1, the suborder *Retroposineae*, and the nuclear-encoded enzyme telomerase, also containing a reverse transcriptase gene. Sequences are drawn to scale and boxes emphasise key coding regions which are conserved between the elements. The elements are aligned through two completely conserved amino acid residues, aspartic acid (DD) in the RT region (Xiong and Eickbush, 1990). Most of the retroelements have their ORFs designated as *gag*, *pol*, and, where present, *envelope*. The *gag* gene is equivalent to the coat protein in viruses, and the *envelope* gene has an equivalent function to the movement protein (MP) of plant pararetroviruses. The C-H motif (see Table I) always precedes the protease which is before the RT and the RNase H is located immediately after the RT. The integrase domain is situated after RNase H in *gypsy* elements and retroviruses, and between the protease and RT in *copia* elements. Pararetroviruses do not have an integrase domain. The *envelope* gene is situated as the last ORF in *gypsy* and retroviruses and before RT in pararetroviruses as a movement domain or function.

A. CONSERVED REGIONS OF RETROELEMENT GENES

In addition to the outline of complete retroelements in Fig. 3, detailed structures of the most conserved domains including the RT, RNase H, C-H motif, integrase and aspartic protease are shown. The RT domain is highly conserved, and Fig. 4 aligns sequences corresponding to domain 3-7 in Xiong and Eickbush (1990).

RNase H, part of the polyprotein ORF, degrades RNA in RNA/DNA hybrids. Malik and Eickbush (2001) aligned RNase H within retroelements and highlighted some single amino acids believed to be important in the catalytic reaction of the protein; D, E, D, D[*]. For the elements shown, the structure can be written as $DX_{27-48}EX_{18-33}DX_{29-54}D$ (where X represents any amino acid and the subscript shows the number between conserved residues; some authors show only the number without X) (Fig. 5). The non-LTR retrotransposons and retroviruses have an H between the last two Ds. A DXS motif can be detected in many of the sequences including several other single or multiple amino acids, many of which are found in the retroelements in Fig. 3.

The integrase is also part of the polyprotein and mediates the integration of an element into nuclear DNA. The integrase domain contains both a well-conserved zinc finger (HHCC) and a $DDX_{35}E$ motif (Khan *et al.*, 1991; Fayet *et al.*, 1990). These two motifs were found in the *copia*, *gypsy* and retrovirus elements in Fig. 3. In these elements the zinc finger motif is $HX_{3-6}HX_{20-33}CX_2C$ (Fig. 6). The two amino acids KD are conserved between *Cyclops*, *Athila*, HERV-K and SFV. The DDX35E motif is some 26-32 amino acids downstream of the last C in the zinc finger. In Fig. 6 the general motif is $DX_{52-64}DX_{32-36}E$, excluding *Cyclops* which has a very long sequence between the two Ds - 111 amino acids. Capy *et al.* (1996) found no similarities to the integrase domain of LTR retrotransposons in LINEs.

The Cysteine-Histidine motif at the C-terminal of *gag* or in the coat protein is very well conserved (Covey, 1986) and is found in LINE, *copia* and *gypsy* elements, in pararetroviruses and a retrovirus (Fig. 7). The protein may bind genomic RNA or DNA to assist in packaging of virus particles and perhaps other processing. It consists of a short sequence with a characteristic pattern of cysteine and histidine amino acids, making up a zinc finger. The motifs in the LINE, *copia*, *gypsy* and retrovirus elements are very similar, having the amino acid sequence $CX_2CX_{3-4}HX_4C$ while the pararetroviruses have an additional CX with the motif $CXCX_2CX_4HX_4C$. The third LINE C-H motif has longer intervals between the C and H than the two other LINE motifs, $CX_4CX_5HX_6C$. The second C-H motif in BSV is rather different from the others, having six Cs and an H, $CX_2CX_7HX_3CX_2CX_4CX_2C$.

[*] Single letter codes used to designate amino-acid residues: D= aspartate; E= glutamic acid; H= histidine; S= serine; C= cysteine; K= lysine; L= leucine; G= glycine; X= any amino acid

Aspartic protease, also part of the polyprotein, cleaves full length mRNA. The protease region is poorly conserved, the best homology being an $LX_{0-4}DXG$ motif, with a few widely-spaced conserved amino acids (Fig. 8; See also McClure, 1991).

V. INTERACTION BETWEEN THE PLANT GENOME AND RETROELEMENTS

Retroelements represent a major fraction of genomic DNA and their maintenance and replication impacts on the organism where they are present. Insertion of retroelements causes changes in the host genome such as insertional mutation, chromosome breakage, chromosome rearrangement, altered gene regulation and sequence amplification. Even remnants of integrated viruses have been shown to have promoter/enhancer activity in the LTRs, active splice sites, ORFs or RT activity and to have the ability to be retrotransposed by complete elements (Löwer, 1999). Various forms of viruses are integrated in the nuclear genomes of eukaryotes, and some are active in transcription and making of episomal virus particles. They could be under active selection e.g. co-segregation because of integration proximal to a allele. Alternatively the integration could alter the expression of neighbouring host genes in a useful regulatory manner, or give virus resistance by anti-sense expression (Bejarano *et al.*, 1996; Mette *et al.*, 2002) or other silencing mechanisms. The retroviruses and pathogenic pararetroviruses cause disease that will often be detrimental to their host, although cases of cross-protection are know where one infection gives protection against subsequent infection by another virus. Hence retrotransposons and inserted RT viruses may have protective consequences for their 'host'.

A. AMPLIFICATION AND HOST CONTROL

Evolution of copy number in retroelements can be interpreted as showing periods of low and high amplification and insertional activity, with evidence that this is related to the development or environment of the host (Grandbastien, 1998; Kalendar *et al.*, 2000). Investigation of barley centromeres revealed the presence of a family of *gypsy*-like elements (*cereba*), while other *gypsy* elements showed a contrasting distribution (Vershinin *et al.*, 2000). Thus there is control of element location, suggesting host genome-insert interactions are involved.

Unusually high activity or unexpected appearance of retroelements is often found in connection with stress events such as tissue culture and wide hybridization

(Dahal *et al.*, 2000; Lockhart *et al.*, 2000; Mhiri *et al.*, 1997): thus evolution and amplification of retroelements can be suggested to occur in sudden steps (in contrast to operating as a molecular clock), and the periods of activity of retroelements would be difficult to estimate from extant data as they can behave differently from the genic and other DNA. SanMiguel *et al.* (1998) show that retrotransposon activity is recent in maize, with virtually all elements inserting within the last six million years and most in the last three million years. Both *Athila* and *Tat1 gypsy* retrotransposons have high sequence degeneracy in the coding regions whereas they have near sequence identity of their 5' and 3' LTRs (>95%; see Fig. 3). The similarity of LTRs suggests that these elements integrated relatively recently or that transcripts from defective elements were acted upon in *trans* to generate the insertions (Wright and Voytas, 1998).

B. RETROELEMENTS AS MARKERS

Retroelements are important both to evolutionary studies and as tools in molecular studies. Because of their abundance, mode of amplification, and insertion in the genome throughout much of its length, the features of retroelements can be used as a source of polymorphic markers for discrimination of plant species or genotypes. In particular, pairs of outward facing primers from the long terminal repeats of LTR retrotransposons are proving valuable for PCR amplification of DNA lying between retroelements, hence giving inter-retrotransposon amplified polymorphic (IRAP) markers (Kalendar *et al.*, 1999).

C. SEQUENCE MOTIFS AND HORIZONTAL TRANSFER

The motifs in figures 4 to 8 show conservation of key amino-acid residues presumably a consequence of common evolutionary origin. How did sequences come to have their current widespread distribution? In many cases, vertical transmission by descent from a common ancestor can be proposed as the distribution mechanism. However, viruses, including pararetroviruses, spread from cell to cell, and to new organisms, independently of inheritance of nuclear DNA. This horizontal transfer can be proposed for some groups of retroelements (or gene components), with evidence coming from high similarity between elements from distantly related species and inconsistencies between the phylogeny of the element and that of the hosts (Capy *et al.*, 1994). The best example of horizontal transfer is that of the *P* element in *Drosophila* (Daniels *et al.*, 1990). In plants, it can be envisaged that retroelements may be transmitted directly as DNA or RNA, or after packaging with other viral DNA sequences. Sugimoto *et al.* (1994) showed how a virus could package and transfer a transposable element from maize into rice. Genomic DNA sequences unrelated to retroelements (or from other elements) might be transferred by evolutionary mechanisms such as unequal crossing over. Such changes are suggested by the retroviruses in Fig. 3, where the distance between RT and RNase H is larger than for the other elements. Malik and Eickbush (2001) propose that an early lineage of retroviruses replaced their existing RNase H domain with one from a LINE-like element which were placed after the original RNase H: the two share the amino acid H (Fig. 5, arrow second from right).

The integrase component is placed differently in the *copia* group compared to the *gypsy* and retrovirus groups (Fig. 3), while it is missing from the pararetroviruses, showing the flexibility of this motif and changes during evolution. Capy *et al.* (1996, 1997) suggest that the integrase domain with the DDE motif of LTR-retrotransposons and retroviruses originated from the transposases of some DNA transposable elements.

D. Silencing and resistance

In many plant systems, the RNA interference phenomenon leads to small pieces of RNA guiding *de novo* methylation of homologous DNA sequences. Methylation is effectively targeted against the promoters of transposable elements (see Martienssen and Colot, 2001). Apart from this short-term protection, methylation also provides a potential mechanism for long-term protection by driving a C to T mutation of the element sequence (Bestor, 1999). For many years, plant expression of viral coat proteins has been known to confer resistance to viral infection. More recently, Matzke *et al.*, (2001; see also Waterhouse *et al.*, 2001) have discussed how RNA interference might operate, as transcription of retroelement RNA could drive degradation and interfere with replication of viral and other RNA species. Mette *et al.* (2002) investigated whether integrated virus-like sequences exhibit features that would be compatible with a potentially new type of homology dependant virus resistance. It was believed that stably methylated sequences have supplied long-term viral immunity, perhaps accompanied by weakening or extinction of the related exogenous virus.

VII. ACKNOWLEDGEMENTS

We are most grateful to Dr Glyn Harper, John Innes Centre, Norwich UK for help with the manuscript and our work described here. CNH thanks the John Innes Foundation for award of a research studentship. We thank the European Union for support through grant QLRT-2001-02098, “Pararetroviruses: Disease, Integration and Genomes”. Some concepts were developed within the IAEA Coordinated Research Programme “Physical mapping technologies”.

VIII. REFERENCES

Baltimore, D. (1970). RNA-dependant DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209-1211.

Barakat, A., Carels, N. and Bernardi, G. (1997). The distribution of genes in the genomes of Gramineae. *Proceedings of the National Academy of Sciences (USA)* **94**, 6857-6861.

Bejarano, E. R., Khashoggi, A., Witty, M. and Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences (USA)* **93**, 759-764.

Bénit, L., Lallemand, J-B., Casella, J-F., Philippe, H. and Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of Virology* **73**, 3301-3308.

Bennetzen, J. L. (2000). Transposable element contribution to plant genome evolution. *Plant Molecular Biology* **42**, 251-269.

Bestor, T. H. (1999). Sex brings transposons and genomes into conflict. *Genetica* **107**, 289-295.

Blackburn, E. H. (1992). Telomerases. *Annual Review of Biochemistry* **61**, 113-129.

Boeke, J. D., Eichinger, D., Castrillon, D. and Fink, G. R. (1988). The *Saccharomyces cerevisiae* genome contains functional and nonfunctional copies of transposon Ty1. *Molecular and Cellular Biology* **8**, 1432-1442.

Brandes, A., Heslop-Harrison, J. S., Kamm, A., Kubis, S., Doudrick, R. L. and Schmidt, T. (1997). Comparative analysis of the chromosomal and genomic organization of Ty1-*copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. *Plant Molecular Biology* **33**, 11-21.

Brown, J. R. (2003). Ancient horizontal gene transfer. *Nature Reviews Genetics* **4**, 121-132.

Buchen-Osmond, C. (2003). The universal virus database ICTVdB. *Computing in Science and Engineering* **5**, 16-25.

Budiman, M. A., Mao, L., Wood, T. C. and Wing, R. A. (2000). A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Research* **10**, 129-136.

Capy, P., Anxolabéhère, D. and Langin, T. (1994). The strange phylogenies of transposable elements: are horizontal transfer the only explanation?*Trends in Genetics* **10**, 7-12.

Capy, P., Italis, R., Langin, T., Higuet, D. and Bazin, C. (1996). Relationship between transposable elements based upon the integrase-transposase domains: Is there a common ancestor?*Journal of Molecular Evolution* **42**, 359-368.

Capy, P., Langin, T., Higuet, D., Maurer, P. and Bazin, C. (1997). Do the interase of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica* **100**, 63-72.

Chavanne, F., Zhang, D-X., Liaud, M-F. and Cerff, R. (1998). Structure and evolution of *Cyclops*: a novel giant retrotransposon of the *Ty3*/*Gypsy* family highly amplified in pea and other legume species. *Plant Molecular Biology* **37**, 363-375.

Covey, S. N. (1986). Amino acids sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of *Cauliflower mosaic virus*. *Nucleic Acids Research* **14**, 623-633.

Dahal, G., Hughes, J. d'A. and Thottappilly, G. (1998). Effect of temperature on symptom expression and reliability of Banana streak badnavirus detection in naturally infected plantain and banana (*Musa* sp.). *Plant Disease* **82**, 16-21.

Dahal, G., Ortiz, R., Tenkouano, A., Hughes, J. d'A., Thottappilly, G., Vuylsteke, D. and Lockhart, B. E. L. (2000). Relationship between natural occurrence of banana streak badnavirus and symptom expression, relative concentration of viral antigen, and yield characteristics of some micropropagated *Musa* spp. *Plant Pathology* **49**, 68-79.

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G. and Chovnick, A. (1990). Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. *Genetics* **124**, 339-355.

Emori, Y., Shiba, T., Kanaya, S., Inouye, S., Yuki, S. and Saigo, K. (1985). The nucleotide sequence of *copia* and *copia*-related RNA in *Drosophila* virus-like particles. *Nature* **315**, 773-776.

Fayet, O., Ramond, P., Polard, P., Prère, M. F. and Chandler, M. (1990). Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences?*Molecular Microbiology* **4**, 1771-1777.

Feng, Q., Zhang, Y. J., Hao, P., Wang, S. Y., Fu, G., Huang, Y. C., Li, Y., Zhu, J. J., Liu, Y. L., Hu, X. *et al.* (2002). Sequence and analysis of rice chromosome 4. *Nature* **420**, 316-320.

Flavell, A. J., Pearce, S. R., Heslop-Harrison, P. and Kumar, A. (1997). The evolution of Ty1-*copia* group retrotransposons in eukaryote genomes. *Genetica* **100**, 185-195.

Franck, A., Guilley, H., Jonard, G., Richards, K. and Hirth, L. (1980). Nucleotide sequence of cauliflower mosaic virus DNA. *Cell* **21**, 285-294.

Geering, A. D. W., McMichael, L. A., Dietzgen, R. G. and Thomas, J. E. (2000). Genetic diversity among *Banana streak virus* isolates from Australia. *Phytopathology* **90**, 921-927.

Geering, A. D. W., Olszewski, N. E., Dahal, G., Thomas, J. E. and Lockhart, B. E. L. (2001). Analysis of the distribution and structure of integrated *Banana streak virus* in a range of *Musa* cultivars. *Molecular Plant Pathology* **2**, 207-213.

Grandbastien, M-A. (1998). Activation of plant retrotransposons under stress conditions. *Trends in Plant Science* **3**, 181-187.

Hansen, C., Harper, G. and Heslop-Harrison, J. S. (2003). Isolation and characterization of pararetrovirus-like sequences from the genome of potato (*Solanum tuberosum*). *Cytogenetics and Genome Research*

Harper, G. and Hull, R. (1998). Cloning and sequence analysis of banana streak virus DNA. *Virus Genes* **17**, 271-278.

Harper, G., Osuji, J. O., Heslop-Harrison, J. S. and Hull, R. (1999). Integration of banana streak badnavirus into the *Musa* genome: Molecular and cytogenetic evidence. *Virology* **255**, 207-213.

Harper, G., Hull, R., Lockhart, B. and Olszewski, N. (2002). Viral sequences integrated into plant genomes. *Annual Review of Phytopathology* **40**, 119-136.

Herniou, E., Martin, J., Miller, K., Cook, J., Wilkinson, M. and Tristem, M. (1998). retroviral diversity and distribution in vertebrates. *Journal of Virology* **72**, 5955-5966.

Heslop-Harrison, J. S. (2000). RNA, genes, genomes and chromosomes: Repetitive DNA sequences in plants. *Chromosomes Today* **13,** 45-56.

Hohn, T. and Fütterer, J. (1997). The proteins and functions of plant pararetroviruses: knowns and unknowns. *Critical Review of Plant Science* **16**, 133-161.

Hull, R. (1999a). Classification of reverse transcribing elements: a discussion document. *Archives of Virology* **144**, 209-214.

Hull, R. (1999b). Plant Pararetroviruses, Rice tungro bacilliform virus. *In* "Encyclopedia of Virology" (A. Granoff and R. G. Webster, eds). Academic Press ,London, UK.

Hull, R. (2001). Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. *Archives of Virology* **146**, 2255-2261.

Hull, R. (2002). Plant virology. Academic Press, London, UK.

Hull, R. and Covey, S. N. (1996). Retroelements: Propagation and adaptation. *Virus Genes* **11**, 105-118.

ICTV (2003). International Committee on Viral Taxonomy. www.ncbi.nlm.nih.gov/ICVT/ Last accessed 9/2003.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921.

Jakowitsch, J., Mette, M. F., van der Winden, J., Matzke, M. A. and Matzke, A. J. M. (1999). Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proceedings of the National Academy of Sciences (USA)* **96**, 13241-13246.

Kalendar, R., Grob, T., Regina, M., Suoniema, A. and Schulman, A. H. (1999). IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. *Theoretical and Applied Genetics* **98**: 704-711.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A. H. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences (USA)* **97**, 6603-6607.

Khan, E., Mack, J. P. G., Katz, R. A., Kulkosky, J. and Skalka, A. M. (1991). retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Research* **19**, 851-860.

de Kochko, A., Verdaguer, B., Taylor, N., Carcamo, R., Beachy, R. N. and Fauquet, C. (1998). Cassava vein mosaic virus (CsVMV), type species for a new genus of plant double stranded DNA viruses?*Archives of Virology* **143**, 945-962.

Kumar, A. (1998). The evolution of plant retroviruses: moving to green pastures. *Trends in Plant Science* **3**, 371-374.

Kumar, A. and Bennetzen, J. L. (1999). Plant retrotransposons. *Annual Review of Genetics* **33**, 479-532.

Kunze, R., Saedler, H. and Lönning, W-E. (1997). Plant transposable elements. *Advances in Botanical Research* **27**, 331-470.

Laten, H. M., Majumdar, A. and Gaucher, E. A. (1998). *SIRE-1*, a *copia*/*Ty1*-like retroelemant from soybean, encodes a retroviral envelope-like protein. *Proceedings of the National Academy of Sciences (USA)* **95**, 6897-6902.

Lecellier, C-H. and Saïb, A. (2000). Foamy viruses: Between retroviruses and pararetroviruses. *Virology* **271**, 1-8.

Lerat, E., Brunet, F., Bazin, C. and Capy, P. (1999). Is the evolution of transposable elements modular? *Genetica* **107**, 15-25.

Lingner, J., Hughes, T. R., Shevchenko, A., Mann, M., Lundblad, V. and Cech, T. R. (1997). Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276**, 561-567.

Lockhart, B. E., Menke, J., Dahal, G. and Olszewski, N. E. (2000). Characterization and genomic analysis of tobacco vein-clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *Journal of General Virology* **81**, 1579-1585.

Löwer, R. (1999). The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends in Microbiology* **7**, 350-356.

Löwer, R., Löwer, J. and Kurth, R. (1996). The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences (USA)* **93**, 5177-5184.

Malik, H. S. and Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Research* **11**, 1187-1197.

Malik, H. S., Henikoff, S. and Eickbush, T. H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Research* **10**, 1307-1318.

Manninen, I. and Schulman, A. H. (1993). *BARE-1*, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). *Plant Molecular Biology* **22**, 829-846.

Mao, L., Wood, T. C., Yu, Y., Budiman, M. A., Tomkins, J., Woo, S-S., Sasinowski, M., Presting, G., Frisch, D., Goff, S. *et al.* (2000). Rice transposable elements: a survey of 73. 000 sequence-tagged-connectors. *Genome Research* **10**, 982-990.

Martienssen, R.A. and Colot, V. (2001). DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**, 1070-1074.

Matzke, M., Matzke, A. J. M. and Kooter, J. M. (2001). RNA: Guiding gene silencing. *Science* **293**, 1080-1083.

McClure, M. A. (1991). Evolution of retrotransposons by acquisition or deletion of retrovirus-like genes. *Molecular Biology and Evolution* **8**, 835-856.

Mette, M. F., Kanno, T., Aufsatz, W., Jakowitsch, J., van der Winden, J., Matzke, M. A. and Matzke, A. J. M. (2002). Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *EMBO Journal* **21**, 461-469.

Mhiri, C., Morel, J-B., Vernhettes, S., Casacuberta, J. M., Lucas, H. and Grandbastien, M-A. (1997). The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Molecular Biology* **33**, 257-266.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Nakamura, T. M., Morin, G. B., Chapman, K. B., Weinrich, S. L., Andrews, W. H., Lingner, J., Harley, C. B. and Cech, T. R. (1997). Telomerase catalytic subunit homologs from fission yeast and human. *Science* **277**, 955-959.

Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olszewski, N. E., Lockhart, B. (1999). Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. *Virology* **255**, 214-220.

Ono, M. (1986). Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. *Journal of Virology* **58**, 937-944.

Panstruga, R., Büschges, R., Piffanelli, P. and Schulze-Lefert, P. (1998). A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucleic Acids Research* **26**, 1056-1062.

Petropoulos, C. J. (1997). Appendix 2: Retroviral taxonomy, protein structure, sequences, and genetic maps. *In* “Retroviruses” (J. M. Coffin, S. H. Hughes, H. E. Varmus, eds), pp. 757; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY, USA.

Pineau, P., Marchio, A., Terris, B., Mattei, M-G., Tu, Z-X., Tiollais, P., Dejean, A. (1996). A t(3;8) chromosomal translocation associated with Hepatitis B virus integration involves the carboxypeptidase N locus. *Journal of Virology* **70**, 7280-7284.

Pringle, C. R. (1999). Virus taxonomy. *Archives of Virology* **144**, 421-429.

Renne, R., Friedl, E., Schweizer, M., Fleps, U., Turek, R. and Neumann-Haefelin, D. (1992). Genomic organization and expression of simian foamy virus type 3 (SFV-3). *Virology* **186,** 597-608.

Richert-Pöggeler, K. R., Shepard, R. J. and Caspar, R. (1996). Petunia vein clearing virus, a pararetrovirus that exists as a retroelement in the chromosomes of its host. Abstracts of Xth International Congress of Virology, Jerusalem, Israel, W05-1.

Richert-Pöggeler, K.R., Schwarzacher, T., Harper, G. and Hohn, T. (2003). *EMBO Journal* *11

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. and Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nature Genetics* **20**, 43-45.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, JZ., Niimura, Y., Cheng, Z. K., Nagamura, Y. *et al*., (2002). The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312-316.

Schmidt, T. and Heslop-Harrison, J. S. (1998). Genomes, genes and junk: the large scale organization of plant chromosomes. *Trends in Plant Science* **3**, 195-199.

Seeger, C. (1999). Molecular Biology (of Hepatitis B). *In* “Encyclopedia of Virology, 2nd edition” (A. Granoff and R. G. Webster eds). Academic Press, London, UK.

Sugimoto, K., Otsuki, Y., Saji, S. and Hirochika, H. (1994). Transposition of the maize Ds element from a viral vector to the rice genome. *Plant Journal* **5**, 863-871.

Tagieva, N. E., Gizatullin, R. Z., Zakharyev, V. M. and Kisselev, L. L. (1995). A genome-integrated hepatitis B virus DNA in human neuroblastoma. *Gene* **152**, 277-278.

Takahashi, K., Akahane, Y., Hino, K., Otha, Y. and Mishiro, S. (1998). Hepatitis B virus genomic sequence in the circulation of hepatocellular carcinoma patients: comparative analysis of 40 full-length isolates. *Archives of Virolology* **143**, 2313-2326.

Temin, H. M. and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226,** 1211-1213.

The Arabidobsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815.

Tikhonov, A. P., SanMiguel, P. J., Nakajima, Y., Gorenstein, N. M., Bennetzen, J. L. and Avramova, Z. (1999). Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proceedings of the National Academy of Sciences (USA)* **96**, 7409-7414.

Vershinin, A. V., Druka, A., Alkhimova, A. G., Kleinhofs, A. and Heslop-Harrison, J. S. (2002). LINEs and *gypsy*-like retrotransposons in *Hordeum* species. *Plant Molecular Biology* **49**, 1-14.

Vicient, C. M., Kalendar, R. and Schulman, A. H. (2001). *Envelope*-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Research* **11**, 2041-2049.

Wang, P-C., Hui, E. K-W., Chiu, J-H. and Lo, S. J. (2001). Analysis of integrated hepatitis B virus DNA and flanking cellular sequence by inverse polymerase chain reaction. *Journal of Virological Methods* **92**, 83-90.

Waterhouse, P. M., Wang, M-B. and Lough, T. (2001). Gene silencing as an adaptive defence against viruses. *Nature* **411**, 834-842.

Wright, D. A. and Voytas, D. F. (1998). Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana Ty3*/*gypsy* retrotransposons that encode envelope-like proteins. *Genetics* **149**, 703-715.

Wright, D. A. and Voytas, D. F. (2002). *Athila4* of *Arabidopsis* and *Calypso* of Soybean define a linage of endogenous plant retroviruses. *Genome Research* **12**, 122-131.

Xia, J., Peng, Y., Mian, I. S. and Lue, N. F. (2000). Identification of functionally important domains in the N-terminal region of telomerase reverse transcriptase. *Molecular and Cellular Biology* **20**, 5196-5207.

Xiong, Y. and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO Journal* **9**, 3353-3362.

Yu, J., Hu, S. N., Wang, J., Wong, G. K. S., Li, S. G., Liu, B., Deng, Y. J., Dai, L., Zhou, Y., Zhang, X. Q. *et al*. (2002). A draft of the Rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79-92.

LEGENDS TO FIGURES

NOTE TO FIGURES: Figures 1-3 are created in Adobe Illustrator Version 9 (infinite resolution). Low-resolution versions are placed in this file. Figures 4-8 are shown with key residues in colour but should be reproduced in black and white with grey boxes.

Figure 1. A classification of retroelements and related viruses (after Hull, 1999a, 2001; ICTV, 2003). Abbreviations: CaMV, *Cauliflower mosaic caulimovirus*; BSV, *Banana streak badnavirus*; SbCMV, *Soybean chlorotic mottle soymovirus*; PVCV, *Petunia vein clearing petuvirus*; CsVMV, *Cassava vein mosaic cavemovirus*; RTBV, *Rice tungro bacilliform tungrovirus*; TVCV, *Tobacco vein clearing cavemovirus*. The taxonomic endings follow the ICTV nomenclature: order - *virales*; suborder (after Hull) - *ineae*; family - *viridae*; subfamily (not shown) - *virinae*; genus - *virus*.

Figure 2. The relationship and origin of retroelements based on alignment of the reverse transcriptase (RT) region (after Xiong and Eickbush, 1990).

Fig. 3. Alignment of retroelements including a LINE, *copia* and *gypsy* elements, pararetroviruses and retroviruses, with telomerase, another enzyme with reverse transcriptase activity. A scale in base pairs is shown. The alignment is manually optimized around the amino acids DD, key aspartate residues at the active site of the reverse transcriptase (RT). For the abbreviations of genes and other components see Table I. Colour code: orange, DD site of RT; purple, RNase domain (RH); yellow, integrase domain (INT); blue, cysteine-histidine motif (C-H); green protease domain (PR); pink, envelope/movement protein domain (ENV/MP). See figures 4 to 8 for aligments of the genes and other components. References are given in descriptions of the individual elements.

Fig. 4. Alignment of the conserved reverse transcriptase region (RT) of the retroelements in Fig. 3. The sequences cover domain 3-7 from Xiong and Eickbush (1990). The telomerase (Eap123), BLIN and HBV have longer sequences than the others, and a group of amino acids has been removed and replaced with the

corresponding number, 103 for Eap123, 27 for BLIN and 58 for HBV, all at the start of the alignment. Each dot (·) represents three amino acids from the sequences.

Fig. 5. Alignment of the RNase H region (RH) of the retroelements in Fig. 3. The arrows above the alignment point to amino acid residues believed to be important for the catalytic mechanisms of RNase H; D, E, D, (H), D. Each dot (·) represents three amino acids from the sequences.

Fig. 6. Alignment of conserved regions from the integrase region (INT) of ten sequences from Fig. 3. The first motif (H-H-C-C) is a zinc finger; the next motif is D-D-E. Part of the Cyclops sequence has been replaced with the number (71) of amino acids removed to show alignment. Each dot (·) represents three amino acids from the sequences.

Fig. 7. Alignment of the Cysteine-Histidine motif (C-H) with ten sequences from Fig. 3. The motif includes a zinc finger with the amino acids C-C-H-C.

Fig. 8. Alignment of the aspartic protease region (PR) of retroelements in Fig. 3. Of conserved amino acids are P-L-D-G-G-G.

FOR WEB PUBLICATION

APPENDIX A. DESCRIPTION OF THE RETROELEMENTS AND TELOMERASE IN FIGURE 3.

Telomerase:

Eap123; the subunit of telomerase containing the RT motif from the ciliate *Euplotes aediculatus*, 3095bp. Upstream of the RT are four conserved domains, QG, CP, QFP and T. The subunit has a GC content of 32% (Lingner *et al*., 1997; EMBL U95964; Xia *et al*., 2000; Nakamura *et al*., 1997).

LINE retroposons:

BLIN; a LINE element from barley (*Hordeum vulgare*). It is too degenerate to give border sites for the two possible ORFs. It has three C-H motifs, the two first are in reading frame one corresponding to ORF1, followed by a possible protease motif; the third is in reading frame two, corresponding to ORF2, between RT and RNase H. A poly-A motif is present at the end of the element. It has a GC content of 62% and is represented by 40-50 copies per genome (Vershinin *et al*., 2002; EMBL AJ270056; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

Pseudoviridae (Ty1-*copia*) elements:

Ty1 element from yeast (*Saccharomyces cerevisiae*). There are 5' and 3' LTRs with 97% identity. It contains two ORFs, TyA and TyB. TyB contains protease, integrase, RT and RNase domains. The GC content is 37% (Boeke *et al*., 1988; EMBL M18706; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

copia element from *Drosophila melanogaster*. The element is bordered by LTRs having 97% identity. There is only one ORF. A C-H and a protease motif in the first third of the sequence is followed by an integrase motif, with the RT and RNase domains in the last third. The GC content is 33% (Emori *et al*., 1985; EMBL X02599; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

BARE-1 element from barley (*Hordeum vulgare*). *BARE-1* is bordered by LTRs with 96% identity. PBS is complementary to the 3' end of the wheat initiator methionyl-tRNA. PPT is present after *pol*. There are one or two ORFs followed by an insert of unknown origin and function. The *gag* has a C-H and a protease motif, *pol* has integrase, RT and RNase H. The whole element has a GC content of 47% (Manninen and Schulman, 1993; EMBL Z17327)

Metaviridae (Ty3-*gypsy*) elements:

gypsy element from *Drosophila melanogaster*. It is bordered by short LTRs, 49% identical. There are three ORFs, *gag*, *pol* and *envelope*. *Pol* contains protease, RT, integrase and RNase H. The GC content is 46% (Petropoulos, 1997; NCBI AF033821; Xiong and Eickbush, 1990).

BAGY-1 *gypsy* element from barley (*Hordeum vulgare*). The bordering LTRs are 94% identical. The PBS next to the 5' LTR is complementary to the 3' end of a methionine initiator tRNA from wheat. A PPT is present just upstream of the 3' LTR. One *gag-pol* ORF was designated containing C-H motif, protease, RT, RNase H and integrase. The GC content is 46% (Panstruga *et al*., 1998; EMBL Y14573; Xiong and Eickbush, 1990; Wright and Voytas, 2002).

Cyclops-2 element from pea (*Pisum sativum*). It is bordered by LTRs that are 95% identical. The PBS is probably complementary to tRNA-glu from pea. The PPT is next to the 3'LTR. There are three ORFs, *gag*, *pol* and one of unknown function. *Gag* has the C-H motif, *pol* has protease, RT, RNase H and integrase. The unknown ORF has no homology with known *envelope* genes of other retroelements and is surrounded by non-coding regions. The element has a GC content of 42% and is present with about 500 copies (Chavanne *et al*., 1998; EMBL AJ000640; Wright and Voytas, 2002).

Athila4-1 element from *Arabidopsis thaliana*. The LTRs are 94% identical with PBS and PPT next to them. There is no C-H motif in the *gag* region. *Pol* has protease, RT, RNase H and integrase. There is a putative *envelope* gene surrounded by non-coding regions. Three transmembrane domains are found in the *envelope*-like ORF including a second PPT. The GC content is 43% (Wright and Voytas, 2002; EMBL AC007209; Malik and Eickbush, 2001).

Pararetroviruses and pararetrovirus-like sequences (PRV-L):

CaMV (*Cauliflower mosaic virus*); a *Caulimovirus*. There are six ORFs and an intergenic region. Hull (2002) gives two additional small ORFs, ORF7 before ORF1 and ORF8 within ORF4. The movement protein is located in ORF1, the coat protein in ORF4, and protease, RT and RNase H are located in ORF5. As with other caulimoviruses and badnaviruses, the numbering of the sequence begins at the putative 5' minus-strand priming site, conserved tRNA-met. The GC content is 40% (Franck *et al*., 1980; EMBL J02048; de Kochko *et al*., 1998; Harper and Hull, 1998; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

Tprv, a reconstructed tobacco pararetrovirus-like sequence from tobacco (*Nicotiana tabacum*). It has four ORFs plus a repeat and an intergenic region. ORF1 contains the coat protein, ORF2 contains the movement protein and ORF3 encodes protease, RT and RNase H. ORF4 has a transactivator (TAV). The GC content is 28% (Jakowitsch *et al*., 1999; EMBL NTA238747; de Kochko *et al*., 1998; Harper and Hull, 1998; Xiong and Eickbush, 1990).

BSV (*Banana streak badnavirus*); a *Badnavirus*. There are three ORFs of which the third is large (5.5 kb) and contains all the genes: movement protein, coat protein, protease, RT and RNase H. The function of ORF1 and ORF2 is unknown. All members of the badnavirus genus have two different C-H motifs in the CP region. The GC content is 41% (Harper and Hull, 1998; EMBL AJ002234; de Kochko *et al*., 1998; Hull, 2002; Malik and Eickbush, 2001).

HBV (*Hepatitis B virus*); a *Hepadnavirus*. In HBV nt 1 is set to be at an *Eco*RI restriction site. To align HBV to the other sequences, a start point was placed at the site for initiation of viral DNA synthesis at nt 1611 which then becomes nt 1. There are four ORFs. The core is equivalent to the coat protein of other pararetroviruses. The *envelope* ORF encodes three polypeptides possibly with transmembrane function. The *pol* contains RT and RNase H. The function of ORF x is unknown but it is able to activate many viral and cellular promoters as well as several signal transduction pathways. The GC content is about 48%. Hepadnaviruses do not encode protease (Takahashi *et al*., 1998; EMBL AB014360; Seeger, 1999; Hull, 1999a; Xiong and Eickbush, 1990; Malik and Eickbush, 2001)

Orthoretorvirineae (retroviruses):

HERV-K10(+); a human endogenous retrovirus. The LTRs are 99% identical with adjacent PBS and PPT. The PBS is complementary to tRNAlys. The element has five ORFs, first two *gag* where the second has two C-H motifs. The third ORF is designated protease. *Pol* has RT, RNase H and integrase. The *envelope* ORF has three transmembrane domains SP, OM and TM. HERV-K provirus (integrated form) is present with about 50 copies per haploid human genome. HERV-K10(+) is a prototype HERV-K genome as it is a construct of HERV-K10 plus a 290 bp fragment from HERV-K8 which is deleted from HERV-K10. Although defective in *gag* and *envelope* this virus still serves as a useful standard for sequence comparison. The GC content is 42% (Ono *et al*., 1986; EMBL M14123; Manninen and Schulman, 1993; Löwer *et al*., 1996; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

SFV-3 (*Simian foamy virus*); a spumavirus isolated from an African green monkey. The element is bordered by LTRs 100% identical with adjacent PBS and PPT. There are three large and three small ORFs. The classical tripartite retroviral division of *gag* does not exist in spumaviruses (see Table I *gag*) and instead *gag* contains some GR-boxes (glysine/argenine) complementary to the C-H motif seen in other retroelements. *Pol* contains a potential protease, RT and integrase. The *envelope* gene has three transmembrane domains SP, SU/OM and TM including and internal promoter (IP) used as a second site of initiation of the plus (+) strand during reverse transcription. After the *envelope* is an ORF containing a putative TAV followed by two small ORFs of unknown function. The GC content is 38% (Renne *et al*., 1992; EMBL M74895; Lecellier and Saïb, 2000; Xiong and Eickbush, 1990; Malik and Eickbush, 2001).

Table I. Genes and other components of retroelements, the abbreviations used in the text, the full name, their position in the element and the function are listed.

Gene or component	Full name	Position	Function	References
ORF	Open reading frame		Sequence capable of translation into a protein	
LTR	Long terminal repeat	Flanking *retrotrans-posineae*	Regions of several hundred base pairs (250-4000) containing regulatory sequences for gene expression: Enhancer, promoter, transcription initiation (capping), transcription terminator and polyadenylation signal. The 3' LTR is not normally functional as a promoter, although it has exactly the same sequence arrangement as the 5' LTR. Instead, the 3' LTR acts in transcription termination and polyadenylation. As a consequence of the replication mechanism of the elements the two LTRs are identical at the time of integration.	Petropoulos, 1997
PBS	Primer binding site	About 18 nt at the end of the 5'LTR	Binding site for a specific tRNA that functions as the primer for reverse transcriptase to initiate synthesis of the minus (-) strand of viral DNA	Petropoulos, 1997
Gag	Group-specific antigen	Usually one of the first ORFs	The gag precursor is cleaved by the viral protease (encoded by pol) into three mature products: the matrix (MA), the capsid (CA), and the nucleocapsid (NC) together forming the "capsid" which surrounds the genome – this complex is the virus core. Equivalent to the coat or transit protein.	Lecellier and Saïb, 2000.
CP	Coat protein		Equivalent to gag	
Cys-His or C-H	Cysteine-histidine repeat motif	C-terminal of gag	RNA or DNA binding site of the coat protein or gag	de Kochko *et al*., 1998
GR box		C-terminal of gag in certain retroelements	Contains three glysine/arginine basic sequences – functionally equivalent to C-H?	Lecellier and Saïb, 2000
Pol	Polyprotein		Contains aspartic protease, reverse transcriptase and RNase H and in some cases integrase	
PR	Aspartic protease	pol	Cleaves the full length mRNA. PR has a significant role in the processing of the polyprotein precursor into the mature form.	Ono *et al*., 1986

RT	Reverse transcriptase	pol	RNA dependant DNA polymerase – translates RNA to DNA	
RH	Ribonuclease H/ RNase H	pol	RNase H is an enzyme that specifically degrades RNA hybridized to DNA.	Petropoulos, 1997
INT	Integrase	pol	Enzyme responsible for removing two bases from the end of the LTR and inserting of the linear double stranded DNA copy of the retroelement genome into the host cell DNA	Petropoulos, 1997
Env	Envelope gene	After pol, but not in pararetrovirus if MP=env	Envelope genes mediate the binding of virus particles to their cellular receptors enabling virus entry, the first step in a new replication cycle. Thus the envelope genes give retroelements the ability to spread between cells and individuals - infectivity. Contain the proteins SU (surface) and TM (transmembrane).	Löwer *et al*., 1996
MP	Movement protein		Cell to cell movement, maybe equivalent to env	Hull, 2002
TAV	Transactivator		Regulating translation of the polycistronic mRNA	de Kochko *et al*., 1998
PPT	Polypurine tract	7-18 nt just upstream of the 3’LTR	The ppt produce the RNA primer for the synthesis of the plus (+) strand of viral DNA	Petropoulos, 1997

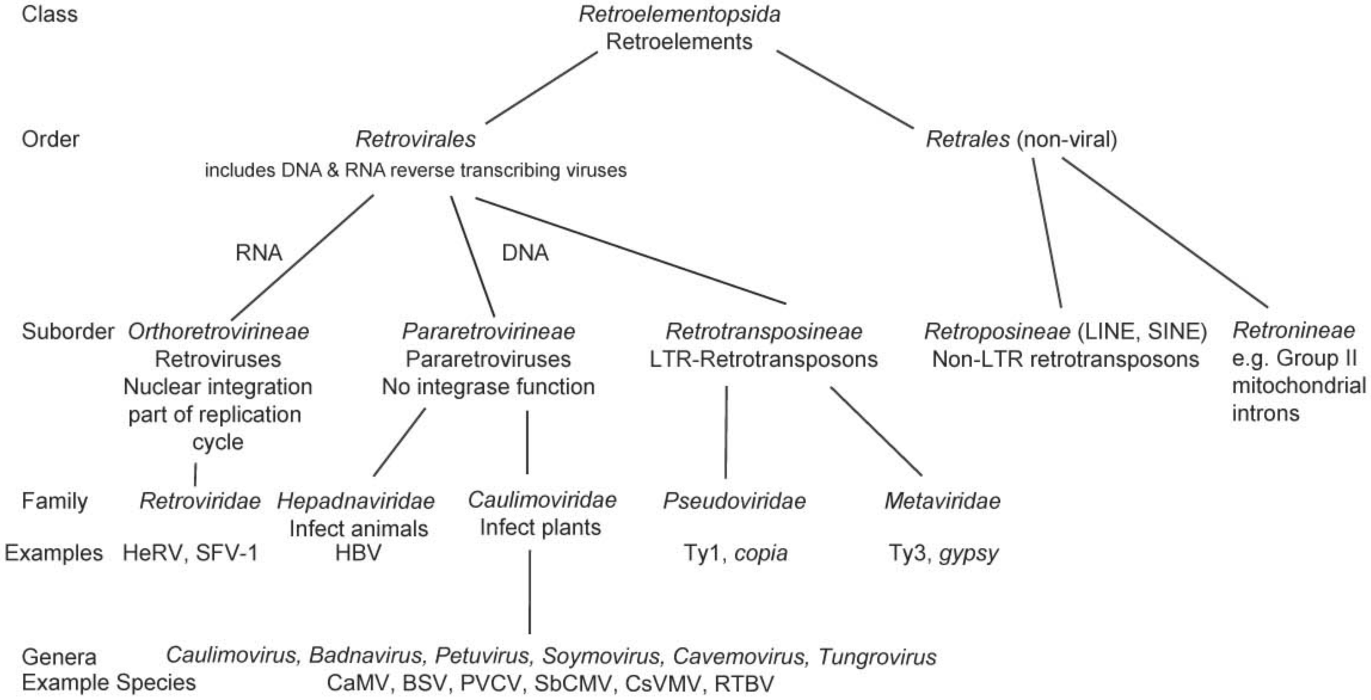


Figure 1. Hansen and Heslop-Harrison.

LOW RESOLUTION FROM VASCTOR GRAPHICS IN ADOBE ILLUSTRATOR

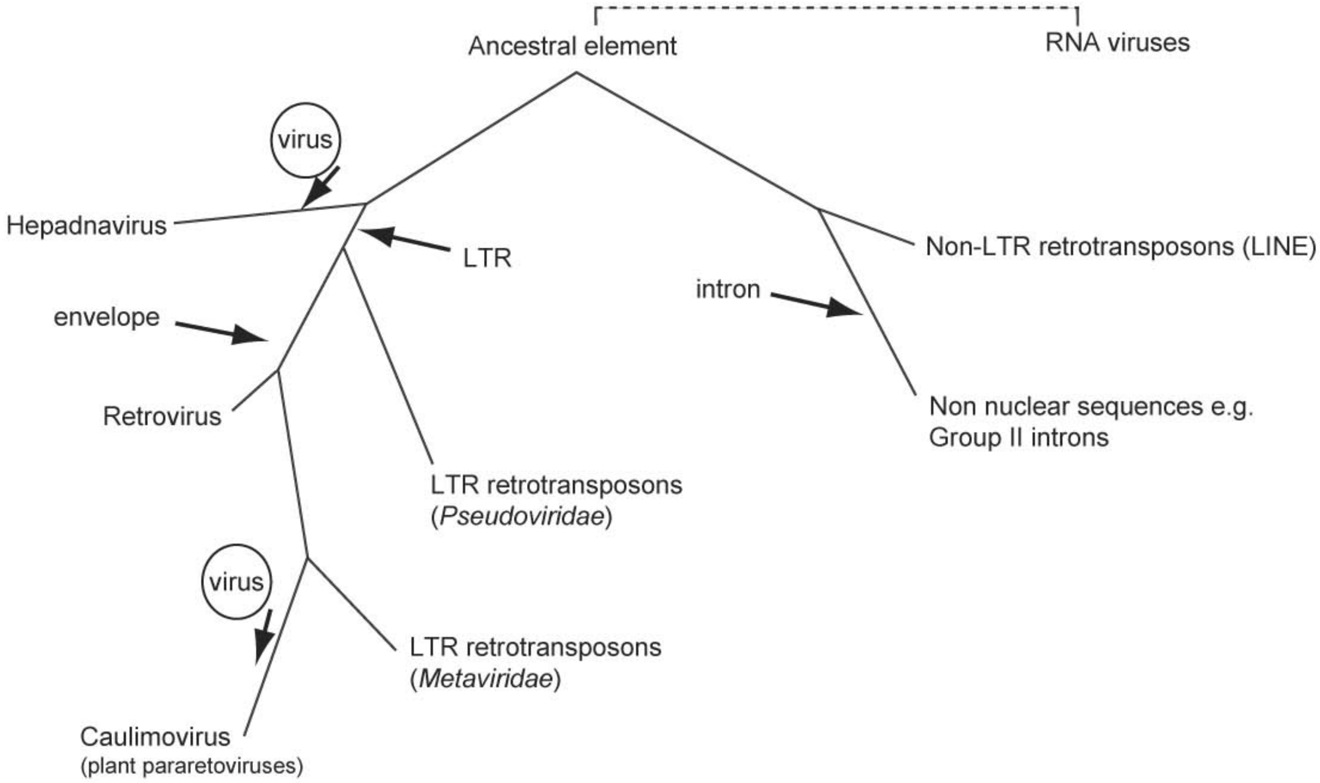
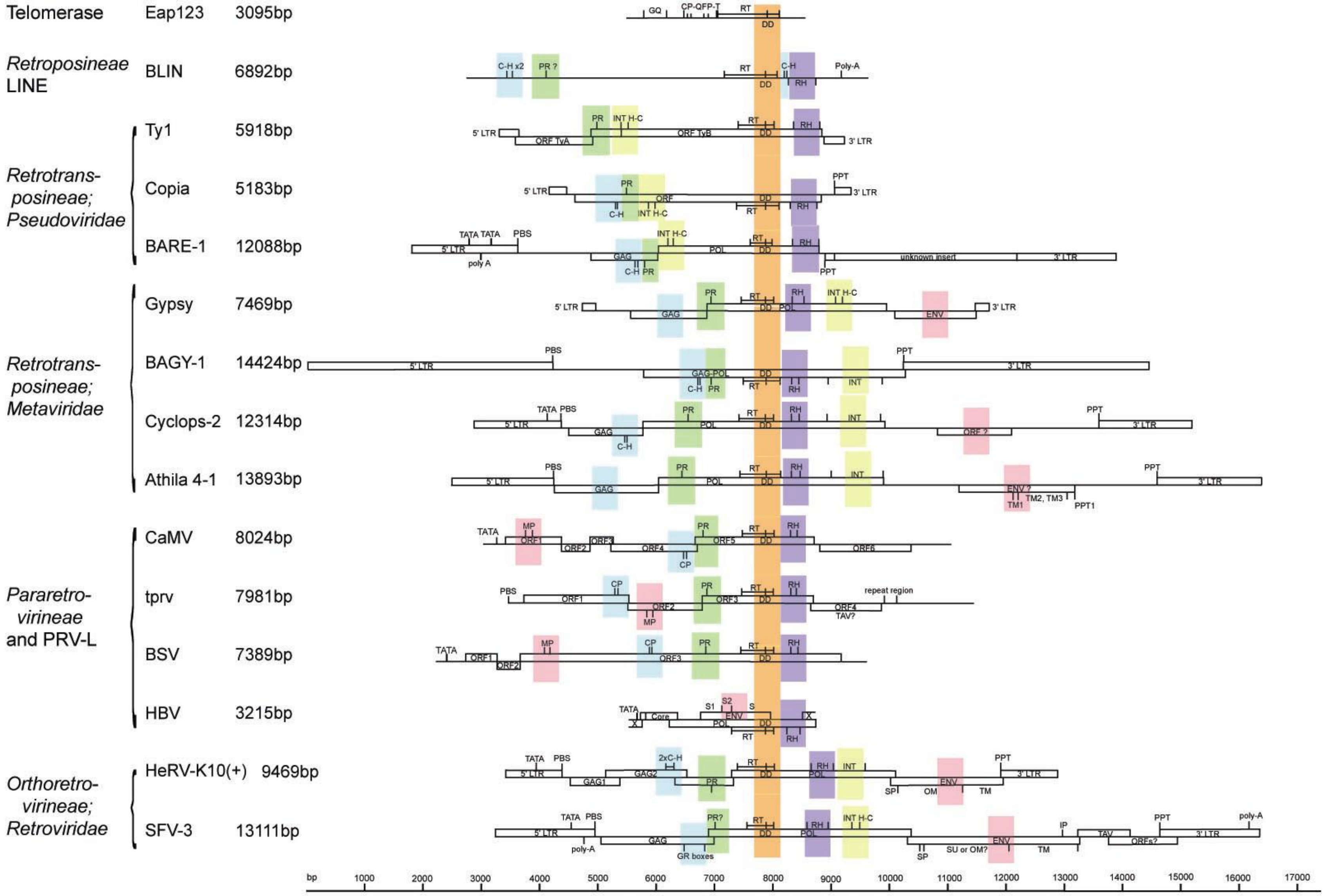

Ancestral element
RNA viruses
virus
Hepadnavirus
LTR
envelope
Retrovirus
intron
Non-LTR retrotransposons (LINE)
Non nuclear sequences e.g. Group II introns
LTR retrotransposons (*Pseudoviridae*)
virus
LTR retrotransposons (*Metaviridae*)
Caulimovirus
(plant pararetoviruses)

Figure 2. Hansen and Heslop-Harrison.



LOW RESOLUTION FROM VASCTOR GRAPHICS IN ADOBE ILLUSTRATOR

Figure 3. Hansen and Heslop-Harrison.

Figure 4.

```
<    domain 3    >             <   domain 4   >             < domain 5 >        <     domain 6     >             < domain 7 >

Eap123
MDIEKCYDSVNREKLSTFLKTTK103FYKQTKGIPQGLCVSSILSSFYY·······NVNLLMRLTDDYLLITTQ··FIEKLINVSRENGFKFNMK KLQTSF······ NIVQDYCDWIGISIDMKTLAL
BLIN
LDLARAFDSVSWPFLFEVLRCHG27PAIWHRRGLHQGDPVSPQLLVLAV········IPAISLYADDVILLCHP··AVKEILQLFGRASGLHVNFQKSAAAL······  IVDFPLTYLGIPLKLRRPTAGQLQ
Ty1
LDISSAYLYADIKEELYIRPPPH·NDKLIRLKKSLYGLKQSGANWYETI········QVTICLFVDDMVLFSKN           LNSNKRIIEKLKMQYDTKIINLGESDEEIQYDILGLEIKYQRGKYMKLGMENS
Copia
MDVKTAFLNGTLKEEIYMRLPQG·· NVCKLNKAIYGLKQAARCWFEVF··········YVLLYVDDVVIATGD               MTRMNNFKRYLME   KFRMTDLNEIKHFIGIRIEMQEDKIYLSQSAYVKKILSKFNM
BARE-1
MDVKAAFLNGLLKEELYMMQPEG··· ACKLQGSIYGLVQASRSWNKRF·········AFLILYVDDILLIGNG               VEFLENIKDYLNK   SFSMKDLGEAAYILGIKIYRDRSRVIGLSQSTYLDKVLKRFK
Gypsy
LDLKSGYHQIY··EKTSFSV··        FCRLPFGLRNASSIFQRALDDVLREQIGKICYVYVDDVIIFSEN··HIDTVLKCLIDANMRVSQE KTRFFK              EYLGFIVSKDGTKS····EPDCVYKVRSFLG
BAGY-1
MDLRLGYHQIK··PKKAFVT··        YTVMSFGLTNAPATFSRLMNSIFMEYLDKFVVVYLDDILIYSMN··HLRLVLMKLREHRLYAKFS KCEFWY           HKVTYLGHVISGKGIAV····QPESVKQVRSFLG
Cyclops
LDGYSGYNQIA··*KTAFTC··        YRKMSFGLCNAPTTFQRCVQAIFADLNEKTMEVFMDDFSVFGVS··NLKTVLERCVKTNLVLNW* KCHFMV           TEGIVLGHKVSSRGLEV····PPVNVKGIRSFLG
Athila
LDGYSGFFQIP··EKTTFTC··        YKRMPFGLCNAPATFQRCMTSIFSDLIEEMVEVFMDDFSVYGPS··NLGRVLTRCEETNLVLNWE KCHFMV           KEGIVLDHKISEKGIEV····PPKTVKDIRSFLG
CaMV
FDCKSGFWQVL··PLTAFTC··        WNVVPFGLKQAPSIFQRHMDE AFRVFRKFCCVYVDDILVFSNN··HVAMILQKCNQHGIILSKK KAQLFK           KKINFLGLEIDEGTHKPQGHILEHINK
tprv
FDCKSGFYHLK··KLTAFTV··        WNVLPFGYKNAPGRYQHFMDN  YFNQLENCIIYIDDILLYSRT··LLEKFIHIVEISGISLSKK KAEVMK           NQIEFLGIQIDKNGIKMQTHVVQKI
BSV
FDLKSGFHQVA··PWTAFWA··        WLVMPFGLKNAPAIFQRKMDN CFRGTEDFIAVYIDDILVFSET··HLKKFMTICEKNGLVLSPT KMKIGT           RQIDFLGATIGNSKIKLQPHII
HBV
LDVSAAFYHIP58FGRKLHLYSHPIILGFRKIPMGVGLSPFL LAQFTS·· RRAFPHCVAFSYMDDVVLGAKS··LFTSITNFLLSLGIHLNPN KTKRWG           YSLNFMGYVIGSWGTLPQEHI
HERV-K
IDLKDCFFTIP··EKFAFTI·· EPATRFQWKVLPQGMLNSPTICQTFV··VREKFSDCYIIHYIDDILCAAET··CYTFLQAEVANAGLAIASD KIQTST            PFHYLGMQI
SFV
LDLSNGFWAHS··WLTAFTWLGQQYCW    TRLPQGFLNSPALFTADV  VDLLKEVPNVQVYVDIYISHDDP··LEKVFSLLLNAGYVVSLK  KSEIAQ           HEVEFLGFNITKEG
```

Figure 5.

```
               ↓                              ↓                     ↓                  ↓        ↓
BLIN
TGLALRMRWQW        LSRVDVSRAWSGLDLHFAPEERALFFASTTMAIGS           GQRALFWEDRWINGLAIREIAPLLFDLIPKQRRKS  RTVADGLHEN QWAADIHGIIGIPEIGEYLRLWHAMAKT     VLTDAPD
Ty1
TRDKQLIWHKN· EPDNKLVAISDASYGNQPY YKSQIGNIYLLNGKVIGGKSTKASLTC TSTT     EAEIHAISESVPLLNNLSYLIQELNKKPII KGLLTDSRSTISIIKSTN····AMRLRDEVSGNNL YVYYIETKKNIADVMT
Copia
TIDMKLIFKKNLAFENKII GYVDSDWAGSEIDRKSTTGYLFKMFDFNLICWNTKRQNSVAASST     EAEYMALFEAVREALWLKFLLTSINIKLENPIKIYEDNQGC   ISIAN·····HFAREQVQNNVI CLEYIPTENQLADIFT
BARE-1
TTEMFLVYGGDKELAVK   GYVDASFDTDPDDSKSQTGYVFILNGGVVSWCSSKQSVVA DSTC     EAEYLAASEATKEGVWMKQLMTDLGVVSSALNPITLFCDNMGV IALAK·····NLIRDYVEEEDV· KVHMDLNVAPAD
Gypsy
FQRLRNILASE··DFKKPFDLTTDASASGIGAVLSQEGRPITMISRTLKQPEQNYATNE            RELLAIVWALGKLQNFLYGSRE        INIFTDHQPLTFAVADRNTNAKIKRWKSYIDQHNAKVFYKPGKENFVADALS
BAGY-1
TSALVLLPPDF    SKDFVIYCDTSRQGLGCILMQDRHVIA            YASRQLHPHEDNYP AHDLELAAVVHALKT** HYLLGNR       CEIFTDHQSLKYIFTQPDLNLRQRRWVELISDYDLGITYTPGKPMLWVMH*V
Cyclops
TLKEKLVIAPI·PNWNLNFELMCDASNYAIGAVLGQRKEKKFHAIH        YASKVLNEAHN    TEKELLAIVYAL EKFR SYLIGSKVVVYTNHSAIKYLLTKPDSKQRLIR       WILLLQEFDVEIKDKKGSENLVADHLS
Athila
TIKDALVSAPV·PNWDYPFEIMCDASDYAVGAVLGQKIDKKLHVIY        YASRTLDDAQGRYATTEKELLAVVFAF EKFR SYLVGSK       VTVYTDHEL*ALRHLYAK·KPRLLRWILLLQEFDMEIVDKKGIENGAADHLS
CaMV
YMQKVKKNLQG··· EEKLIIETDASDDYWGGMLKA IKINEGTNTELICR   YASGSFKAAEKNYHSNDKETLAVINTI KKFS IYLTPVH       FLIRTDNTHFKSFVNLNY··GRNIRWQAWLSHYSFDVEHIKGTDNHFADFLS
tprv
QKIKNMCKKLP··· QFTYIVETDSSDHSYGGVLKY   KYDNEKIEHHCR   YYSGSYTEPQLKWEINRKELFGLYKCL LAFE PYIVYNK       FIVRTDNTQVKWWITRKV··KEIRRLVLNIQNFTFTIEVIRTDKNVIADYLS
BSV
IVKEVKEVVAN··· KAIMIIETDGCMEGWGGVCKWKTDSLQPRWSEKICA   YASGKFTPIKSTID   AEIQAVINSLD KFKIYYLDKKE       LIIRTDSQAIVSFYKKSS···LAFTDYITGTGLEIKFEHIDGKDNVLADTLS
HBV
TYKAFLCQQYL··· SGLCQVFADATPTGWGLAIGHRRMRGTFVAPLPIH                    TAELLAACFARSRSGA               KLIGTDNSVVLSRKYTSF   PWLLGCAANWILRGTSFVYVPSALNPADDPS
HERV-K
TRREPLENAL         TVFTDGSSNGKAAYTGPKERVIKTPYQSAQR                     DELVAVITVLQDFDQP              INIISDSAYVVQATRDVE··········· YIRAHTNLPG· KANEQADLLV
SFV
TWMSYLEDPR27HPSEFSMVFYTDGSAIKHPNVNKSHNAGMGIAQVQFKPEFTVINTWSIPLGDHTA  QLAEVAAVEFACKKALKIDGP           VLIVTDSFYVAESVNKEL·············EKGHQPTAS·TEGNNLADKLA
```

Figure 6.

```
Ty1
HRMLA  HANAQTIRYSLKNNTITYFNESDVDWSSAIDYQ CPDCLIGK······EPFQYLHT DIFGPVHNLPNS···········TTILAFIKNQFQASVLVIQMDRGSEYTNRTLHKFLEKNGITPCYTTTADSRAHGVA ERLNR
Copia
HERFG  HISDGKLLEIKRKNMFSDQSLLNNLELS       CEICEPCL······KRPLFVVHSDVCGPITPVTLD··········FQDFVAKSEAHFNLKVVYLYIDNGREYLSNEMRQFCVKKGISYHLTVPHTPQLNGVS ERMIR
BARE-1
HCRLG  HIGVKRMKKLHTDGLLESLDT             CEPCLMGK······IIHTDVC  DPMSVEARSGYH·········FKQFQSEVENHYNKKIKFLRS DRGGEYLSFEFGAHLRQCGIVSQLTPPGTPQCNGVS ERRNR
Gypsy
HNRA   HRAAQENIKQVLRDYYFPKMGSLAKEVVAN    CRVCTQAK······TGEMVHI  DIFSTDRKLFLT······IVDITAPLLQIINLFPNIKTVYC  DNEPAFNSETVTSMLKNSFGIDIVNAPPLHSSSNGQVERFHS
BAGY-1
HDSTLTIHPRSTKMYQDLRQRFWWTRMKREIAEFVAN    CDVCRRVK······KWDKVSM  DFITGFPKTKKG·········QLAELYVSRIVFLHGVPLGINSDRGSIFTSRFWESFQNAMGTHLSFSTAFHAQSSGQV ERVNQ
Cyclops
HNSYGG HYNGVRTATKILQSGFYWPTIF   KDAHTHAQSCDSCQRSG······FDCWGI   DFVGPFPPLMVTSICLSQLRRLPHLGRMRKRLPEK71PRVLISDGGSHFCNAPLESILKHYGVSHRVATPYHPQANGQA EVSNR
Athila
HGSAYGGHFATFKTVSKILQAGFWWPTMF   KDAQEFVSKCDSCQRKG······FDVWGI   DFMGPFPSSYGN········KVVLKLFKTIIFPRFGVPRVVISDGGKHFINKVFENLLKKHGVKQVEISNREIKTIL   EKTVG
HERV-K
HALT   HVNAAGLKNKFDVTWKQA       KDIVQH   CTQCQVLH······WQM      DVTHVPSFGRLS········HVKKHLLSCFAVMGVPEKIKT  DNGPGYCSKAFQKFLSQWKISHTTGIPYNSQGQAIV ERTNR
SFV
  IILQAHNIAHTGRDSTFLKVSSKYWWPNLRKDVVKVIRQCKQCLVTN······FDKFFI   DYIGPLPPSNGY········SATVKALNMLTSIAVPKVIHS  DQGAAFTSATFADWAKNKGIQLEFSTPYHPQSSGKV ERKNS
```

Figure 7.

```
BLIN-1      CFR  CLEGG   HRVCA  C
BLIN-2      CCR  CLISG   HESNC  C
BLIN-3      CLRQGCLERDS  HPSAPRAC
Copia       CHH  CGREG   HIKKD  C
BARE-1      CYY  CKGMG   HWKRN  C
BAGY-1      CYM  CGEPG   HYS*E  C
Cyclops     CEL  CKGD    HDTGF  C
CaMV      CRCWI  CNIEG   HYANE  C
tprv      CTCYN  CGKLG   HLAKD  C
BSV-1     CRCYA  CGEEG   HFASE  C
BSV-2       CKA  CGSEAAPKHRIDCLKCEMTVCLMC
HERV-K-1    CYN  CGQIG   HLKKN  C
HERV-K-2    CPR  CKKGK   HWASQ  C
```

Figure 8.

```
BLIN
SVQLELRGILPQAWHLSTAEHIFGTGCWVERLHP  DTRSRADLAVFRLTVRVRDLASIRREAILELVEHVPADRPDLPPAFRTLEYPISIRLV          QSAALPRVVDDATNGNG   TGDGEADGSMPDPAGHG
Ty1
ISTTFILGQKLTESTVNHTNHSDDELPGH LLL   DSGASRTLIRSAHHIHSASSNPDINVVDAQKRNIPI NAIGDLQFHFQDNTKTSIKVLHTPNIAYDLLSLNELAAVDITACFTKNVLERSDGTVLAPIVKYGDFY
Copia
KQVQTATSHGIAFMVKEVNNTSVMDNCGF VL    DSGASDHLINDESLYTDSVEVVPPLKIAV        AKQGEFIYATKRGIVRLRND           HEITLEDVLFCKEAAGN    LMSVKRLQEAGMSIEF
BARE-1
KYLADKKAAKEKSGIFDIHVIDVYLTSSR SSAWVFDTGSVAHICNSKQELRNKRRLAKDEVT          MRVGNGSKVDAIAVGTISLQLPSGLVMNLNNCYLVSALSMNIIWILFIARRLLVFKSENNGCSVSMSN
Gypsy
VEFFRGRSRLPFI      ERRLAGRTLK MLI   DTDAAKNYIRPVKELKNVMPVASPFSVS         SIHGSTEIKHKCLMKVFKHISPFFLLDSLN                  AFDAIIGLDLLTQAGVKLNL
BAGY-1
YVSAEEAAENPDV   ILGTLLVNHHPTR VLF   DTGSSHSFISESYALLHNMSFCDMPIP LIV      QTPGSKWETSRITYDNEILVYRLVFLASLLALKSL             DINIILGMDWMSAHYAKIDT
Cyclops
QRTLPKKEVDPGR  VTLPVKIGDIYVGK GLI   DLGSSINLIPFSIVKRLGNIEIKSIRMTLQLADKSTLTKTSWATP*GWVLDKFFFPVDFIVIDMEEDD                 DAPLILGRPFMKTARMMIDV
Athila
KKIIPKKLSDPGS  FTLPCSLGPLAFNR CLC   DLGASVSLMPLSVAKRLGFTQYKSCNISLILADRSVRIPHGLLENLPIRIGAVEIPTDFVVLEMDEEP                 KDPLILGRHFLATAGAMIDV
CaMV
QTEQVMNVTNPNS IYIKGRLYFKGYKKI ELHCFVDTGASLCIASKFVIPEEHWVNAERPIMVK        IADGSSITISKVCKDIDLIIAGEIFRIPTVYQES              GIDFIIGNNFCQLYEPFIQF
tprv
         MPKI  YILSKIIVEGYYN RYYTPMVDTGAEANMCRHNCLPESKWEKKTPIVVTGF       NNEGSMITYKARNIKIQIWDKILTIEEIYSYEFT              NKDILLGMPFLDKLYHIITK
BSV
LEEVSINALRPRNNHLNIKCEIEVKNKKV VLNAILDTGATVCVADERMIPSGMKEQAKNKIIIR        GVNGVTEVNEVTSAGKLWVGKQWFYLPQTFIMPSLAD           GVHMIIGMNFIRTVGLRIEN
HERV-K
YWASQVSENRPVC      KAIIQGKQFE GLV   DTGADVSIIALNQWPKNWPKQKAVT            GLVGIGTASEVYQSMEILHCLGPDNQESTVQPMITS            IPLNLWGRDLLQQWGAEITM
SFV
PPRLVQVKMDPLQ    LLQPLEAEIKGT KLKAHWDSGATITCVPQAFLEEEVPIKNIWIK           TIHGEKEQPVYYLTFKIQGRKVEAEVISSPYDYILVSPSDIPWLMKKPLQLTTLVPLQEYEERLLKQT
```