Full Length Research Paper

Highly heterogeneous *Ty3/Gypsy-like* retrotransposon sequences in the genome of cassava (*Manihot esculenta* Crantz)

Michael A. Gbadegesin^{1*} and John R. Beeching²

¹Department of Biochemistry, University of Ibadan, Ibadan 200005, Nigeria. ²Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, U. K.

Accepted 2 March, 2011

The use of PCR has enabled the survey of transposable elements in many plants; thereby making the study of their diversity and applications possible in species where the full genome sequence data are not yet available. In the present study, we used PCR primers anchored on the conserved domain of reverse transcriptase and endonuclease to amplify the Ty3/Gypsy-like polyprotein fragment from the genome of cassava (*Manihot esculenta* Crantz). The PCR product was cloned and sequenced. Sequence analysis of individual clones clearly identified the conserved domain of the polyprotein enzymes and showed the cassava Ty3/Gypsy-like retrotransposon, Megyp (for *Manihot esculenta* gypsy-like), sequences to be highly heterogeneous. Some Megyps clustered with other plants' Ty3/Gypsy-like retrotransposons, while some clustered with Gypsy of *Drosophila melanogaster* and Ty3-2 of *Saccharomyces cerevisiae* in the comparative multiple sequence analysis. This suggests that the later belong to the retrovirus lineage of this group of elements. Southern analysis showed that, the Megyps and analogues were highly repeated within the genomes of cassava cultivars.

Key words: Cassava, transposable-elements, retrotransposons, retroviruses, Manihot esculenta, Ty3/Gypsy.

INTRODUCTION

There are two major super-families of transposable elements (TEs) based on their transposition intermediate and transposition mechanisms (Finnegan, 1992). DNA TEs (Class II elements) move by excision and reintegration via a DNA intermediate. They transpose by a 'cut and paste' mechanism mediated by a transposase that recognises their short terminal inverted repeated sequences (TIRs). On the other hand, retrotransposons or retro-elements (Class I elements) move and amplify through RNA intermediates, which are reverse transcribed before their integration into the nuclear genome. They have been divided into two principal groups, the long terminal repeat (LTR) retrotransposons and the non-LTR retrotransposons.

Non-LTR retrotransposons lack LTRs and are transcribed from an internal promoter. They are subdivided into long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). The LTR retrotransposons are further divided into two groups Ty1/copia and Ty3/gypsy. These were so named after the elements first described in *Saccharomyces cerevisiae* (*Ty1* and *Ty3*) and *Drosophila melanogaster* (*Copia* and *Gypsy*). Transcription of LTR retrotransposons starts at the 5' LTR and ends at the 3' LTR. The LTRs usually contain the regulatory sequences for promoting and terminating transcription of the element.

The use of PCR primers based on the highly conserved amino acid sequence of enzymes domains has proved highly successful in the survey of transposons in many plants (Flavell et al., 1992; Hirochika and Hirochika, 1993; Suoniemi et al.,1998; Vershhinin et al., 2002; Staginnus et al., 2001). It is making the study of transposable elements diversity, abundance and applications possible in species where full genome sequence data are not yet available.

Although *Ty1/copia*-like elements have been reported in many higher plants, fewer *Ty3/gypsy*-like retrotran-

^{*}Corresponding author. E-mail: magbadegesin@yahoo.com. Tel: +234 2 7504769. Fax: +44 8712 564876.



Figure 1. Structural organisation of Ty3/gypsy-like retrotransposons. They are bounded at their termini by long terminal repeats (LTRs). The primer binding site (PBS) and polypurine tract (PPT) is represented as grey rectangles. In between PBS and PPT (coloured boxes), are the two open reading frames (ORFs) with coding potential for the structural and enzymatic proteins needed for the retrotransposition cycle: the group antigenic glycoprotein (gag) domain coding for the protein that forms the nucleocapsid core; the protease (pr) domain encoding the proteins necessary for the maturation of the different proteins; the reverse transcriptase (RT) domain encoding the enzyme responsible for the creation of a DNA copy from the genomic RNA template; the ribonuclease H (RNaseH) domain encoding the enzyme for degradation of RNA hybridised to the first strand DNA: the endonuclease (endo) domain, encoding proteins necessary for the integration of the DNA copy into the host genome. Most of these proteins are encoded as polyproteins (pol) sometimes with overlapping ORFs and are processed into individual components by pr. In addition, a third open reading frame (ORF3) encoding an env-like activity is frequently found in Ty3/gypsy retrotransposons. The block arrow heads indicate the position of the forward and reverse primers for the PCR.

sposon sequences or elements have been identified in plant species (Su and Brown, 1997). *Ty3/gypsy*-like retrotransposons share common features with *Ty1/copia*-like elements but the order of the domains between the two long terminal repeats (LTRs) in *Ty3/gypsy*-like elements resembles those of the retroviruses (LTR-*gag-pr-rt-RNaseH-endo*-LTR) (Figure 1). Some members of *Ty3/gypsy* superfamily also sometimes contain an additional open reading frame (ORF3) encoding an *env*-like gene.

Cassava (*Manihot esculenta* Crantz) is the world's sixth most important crop in terms of production (Mann, 1997) and the staple food of over 500 million people in the tropical regions of the world. It however, has been grossly understudied. In this study we isolated, cloned, sequenced and analysed cassava polyprotein fragment unique to Ty3/gypsy-like retrotransposons using degenerate PCR primers. Cassava Ty3/gypsy-like retrotransposons have been named *Megyp* for *M. esculenta gypsy*-like. The diversity and organization of *Megyp* within the cassava genome and their relationship to those of other plants are also analyzed. The nucleotide sequences described here have been submitted to the Genbank database and given the accession numbers AY946154 -

AY946199.

MATERIALS AND METHODS

Plant material and DNA isolation

Using the method of Dellaporta et al. (1983), DNA was extracted from young leaf samples of cassava cultivars grown in the tropical glasshouse at the University of Bath. The growth conditions include temperature at 22 to 28°C, relative humidity of 40 to 80% and a minimum light period of 12 h per day under day light, supplemented with 400 W Phillips high-pressure sodium lights when necessary.

PCR Amplification of polyprotein fragment of Megyp sequences and cloning

The PCR method used was as described by Suoniemi et al. (1998) with some modifications as described by Gbadegesin et al. (2008). Amplified DNA bands were gel purified (Qiagen, 'Qiaquik'), ligated into $pGEM^{\textcircled{B}}$ -T Easy vector (Promega) and used to transform competent *Escherichia coli* DH5 α according to standard procedures (Sambrook et al., 1989).

DNA gel blot analysis

Restriction digestions of genomic DNA (5 µg each) were carried out using buffer and reaction conditions specified by the manufacturer (Promega). Blotting and hybridisation were performed using standard procedures (Sambrook et al., 1989).

Sequence and phylogenetic analyses

DNA molecules were sequenced on an ABI 337 automated dye primer sequencer using universal primers for the cloning vector. The first line of sequence identification was by using BLASTN and TBLASTX searches against the GenBank non-redundant database at the default parameters (Altschul et al., 1990). The sequence fragments were assembled using the Vector NTI program. Consensus sequence data were aligned using CLUSTAL W (version 1.82) (Higgins et al., 1994). The PHYLIP program package version 3.63 (Felsenstein, 2004), available from the author at Department of Genetics, University of Washington, Seattle, Washington, was used for phylogenetic analysis. Consensus NEIGHBOR-joining trees (Saitou and Nei, 1987) were derived from equally parsimonious trees using the extended majority rule in the CONSENSE. Unless otherwise stated, distance matrices for phylogenetic analyses based on nucleotide sequences data were computed using DNADIST according to the Kimura 2-parameter model (Kimura, 1980). Trees were drawn using TREEVIEW program version 1.6.6 available from the author, Roderic D.M. Page of the Taxonomy Unit, Department of Zoology, University of Glasgow.

RESULTS

PCR amplification of cassava *Ty3/Gypsy-like* retrotransposon polyprotein fragment, cloning and sequence analysis

PCR was carried out as described in the materials and methods section. The amplified products were analysed by electrophoresis on ethidium bromide stained 0.8%



Figure 2. PCR amplification of Ty3/gypsy-like polyprotein fragment from cassava genomic DNA. PCR product was run on a 0.8% agarose gel stained with ethidium bromide. The size marker (lane M) is bioline DNA 100 bp ladder, while the PCR product is shown in the right lane

agarose gel. Approximately 1.6 kb cassava DNA was amplified (Figure 2). Amplified DNA was purified and cloned as described earlier.

A clone was selected at random and sequenced from both ends using T7 and SP6 primers. The sequence was then submitted to BLASTN and TBLASTX searches against the GenBank non-redundant database using the default parameters. The searches confirmed that, a Ty3/gypsy-like polyprotein fragment had been amplified in the PCR experiments. The cassava element was 67% identical (within the region of the alignment) to the Ty3/gypsy-like retrotransposon polyprotein of Olea europaea at the amino acid sequence level (Figure 3). Thirty-six (36) clones (named Megyp1, Megyp2.... Megy36) in total were randomly selected.

Sequence and phylogenetic analysis

The selected clones were partially sequenced in both directions using the T7 and SP6 primers. The sequences were assembled using the Vector NTI contig assembly program. The NTI contig assembly allowed visualization and removal of vector sequences. The vector free sequence data were submitted to BLASTN and TBLASTX searches as before. Sequencing from the 5' end gave 26 *Megyp* clones with good sequences of which 20 (77%) showed clear homology to the polyprotein of *Ty3/gypsy*-like retrotransposons. However, sequencing from the 3' end gave 28 clones with good sequences of which 26 (93%) showed clear homology to the polyprotein of *Ty3/gypsy*-like retrotransposons (in most cases E-value)

were in the region of e^{-63}). These data show that, cassava *Ty3/gypsy*-like retrotransposons are more diverged at the 5' end of the amplified polyprotein fragment compared with the 3' end. Overall, the use of PCR primers anchored on RT and endonuclease domains proved useful and efficient for the isolation and characterisation of this group of cassava retroelements.

The deduced translations of the Megyps left and right nucleotide sequences were obtained using ORF finder (www.ncbi.nlm.nih.gov/gorf/) (data not shown). Twelve (70.6%) of the clones having good left and right sequences and clear homology to Ty3/gypsy-like retrotransposon, contain neither a frame shift nor a nonsense mutation, while five (29.4%) have these mutations within the sequences analysed. While it is possible to say that the latter group could be defective enzymes, full sequence data would be necessary to conclude that the former code for functional enzymes.

Two of the clones with uninterrupted open reading frames within the left and right sequences, Megyp5 and Megyp28, were fully sequenced. They contain no stop or frame shift mutations within the RT-RNaseH-endonuclease sequences analysed. The nucleotide sequences and deduced translation of these clones are shown in Figure 4a, b. The two shared 88% sequence identity at the nucleotide sequence level and 89% identity at the level of amino acid sequence. The 5' (RT) ends of the Meavps are more diverged than the 3' (ENDONUCLEASE) ends and the *Megyp*5 and *Megyp*28 nucleotide sequences did not align in the first 15 nucleotide base positions (data not shown). However, the presence of the highly conserved block YAKFSKCEF of the RT domain characteristic of Ty3/gypsy retrotransposons (highlighted grey in Figures 4a, b) is a guick check and provides strong evidence for it being part of the polyprotein sequence in all of the cassava Ty3/gypsylike retrotransposons.

To determine the relatedness of the cassava Ty3/gypsy-like retrotransposons to each other the nucleotide sequences (with the primer regions removed) for the 17 *Megyps* (left and right fragments for 15; full ~1.6 kb fragment sequences for *Megyps* 5 and 28) were aligned using CLUSTAL W) (Higgins et al., 1994). The aligned nucleotide sequences were used to compute a distance matrix using DNADIST of the PHYLIP package version 3.63 (Felsenstein, 2004), according to the Kimura 2-parameter model (Kimura, 1980). Trees were then produced using the neighbor-joining method.

This method is based on all pairwise comparisons in which positions for which there was no sequence data, for example, the central regions for all sequences other than *Megyp*5 and 28, were treated as missing data rather than as gaps (Felsenstein, 2004).

Using an extended majority rule in the CONSENSE program from the PHYLIP package, a consensusunrooted tree was derived from 100 equally parsimonious trees. The consensus tree was drawn using TREEVIEW

Me:	55	${\tt CKIYQRVKLEHQKPAGMLNPLPIPEWKWENVVMDFVVGLPATSNRLNSIWVIVDRLTKSA$	234
		C + Q+VK+EHQKPAG LNPL IPEWKWEN+ MDFVVG P ++ N+IWV+VDRLTKSA	
0e:	1535	${\tt CMVCQQVKVEHQKPAGWLNPLDIPEWKWENITMDFVVGFPKSAIGNNAIWVVVDRLTKSA$	1714
Me:	235	${\tt HFIPVRSGYSVDKLAQVYVEEIIRLHGAPVSIVSDRRLQFTSRSWRSLQNAMGTRLDLST}$	414
		HF+PV+ +S+D+LAQ+Y+++++RL G PVSIVSDR L+FTS+ W+SLQ AMGT+L+ ST	
0e:	1715	${\tt HFLPVKMTFSLDQLAQLYIKDVVRLCGVPVSIVSDRDLRFTSKFWKSLQGAMGTKLNFST}$	1894
Me:	415	AFHPQTDGQSER 450	
		A+HPQTDGQSER	
0e:	1895	AYHPQTDGQSER 1930	

Figure 3. Alignment of cassava (*Me*) amino acid sequence with *Olea europaea* (*Oe*) amino acid sequence of partial polyprotein *gypsy*-like retrotransposon (gi, 7283091). The two sequences show 67% identity and 86% homology.

(version 1.6.6) as shown in (Figure 5). Three families of *Megyps*, I, II and III emerged from the phylogenetic analysis (Figure 5). There are seven, six and four clones, respectively in these families.

The predicted amino acid sequences of the plant Ty3/gypsy-like polyprotein listed in Table 1 and that of *Gypsy* were aligned with those of *Megyps* (representative cassava Ty3/gypsy-like retrotransposons) using the CLUSTAL W programme (1.82) and colour shaded in GENDOC as shown in Figure 6. The alignment reveals blocks of residues previously identified as highly conserved (Barber et al., 1990; Kulkosky et al., 1992; Springer and Britten, 1993; Xiong and Eickbush, 1990). There is highly conserved block YAKFSKCEF (box a) that includes the invariant lysine (underlined) of reverse transcriptase (Barber et al., 1990).

The conserved TDAS motif that defined the RNase H region in most other Ty3/gypsy-like retrotransposons (Springer and Britten, 1993) is present in most cassava elements as CDAS (box b). In both cases, a key active-site aspartate (Campbell and Ray, 1993) is conserved. Also, conserved in the two fully sequenced cassava Ty3/gypsy POL fragments, Megyp5 and 28, is the motif N-3-DXL (box c) known to be essential in RNase H catalysis (Campbell and Ray, 1993).

The N-terminal DNA-binding domain of integrase (Kedar and Khan, 1990) is revealed as a conserved X-6-H-29-C-2-C motif (box d), from which all the cassava elements lack the first four upstream amino acids, a feature shared with many other published sequences of *Ty3/gypsy* POL (box d, Figure 6). The highly conserved N-terminal GLLQPLPI motif (box e) of integrase is present in all the *Megyps* as homologous GMLNPLPI. Also present in the aligned *Megyps* is a D-60-D-35-E motif of integrase domain, where E is part of the 3' primer sequence (not included in the alignment). The D,D-35-E motif is completely conserved in retroviral and retrotransposon integrases and is essential for enzymatic activity (Baker and Luo, 1994; Kulkosky et al., 1992).

Overall, the amino acid sequences of the predicted translation of cassava Ty3/gypsy-like POL compared well with other Ty3/gypsy elements. This therefore confirmed them again as authentic Ty3/gypsy-like polyprotein sequences.

Three families of *Megyps* and other plants *Ty3/gypsy*like retrotransposons emerged from the subsequent phylogenetic analysis (Figure 7). The cassava elements on the tree are indicated with arrowheads. These analyses revealed a high level of heterogeneity of Megyps among the reported plant Ty3/gypsy group retrotransposon using a PCR based assay. There are two monophyletic families (I and II) consisting of cassava Ty3/gypsy-like retrotransposons. The two clades were supported by bootstrap values of 49 and 45%, respectively, in the extended majority rule consensus tree (Figure 7). The third clade (III) supported by 100% bootstrap value consists of Gypsy of Drosophila melanogaster and the Ty3/gypsy-like retrotransposons of Arabidopsis thaliana rAt1, Ananas comosus, Oryza sativa, Hordeum vulgare rHv1, Lilium henryi del and one cassava element, Megyp18 and at 63% bootstrap, a second cassava element (Megyp 22) is included in this clade. The association within the sequences in this clade is very robust as shown by the high values of the bootstrap. Surprisingly, Megyp18 associated closely with Gypsy and Ty3-2 in clade III (Figure 7) Gypsy, like other retrovirus-like Ty3/gypsy retrotransposons, is known to encode env-like activity. Further studies would be required to classify the *Megyps* in this grouping as members of these endogenous retroviruses.

Study on the genomic organization and diversity of *Ty3/Gypsy-like* retrotransposons in cassava cultivars

A representative cassava *Ty3/gypsy*-like polyprotein fragment, *Megyp*5, was used to probe Southern blots of restriction digests of genomic DNA from a range of

Α

1	;tggggttggtcttgcagactttgagggaacatggcttgtatgccaagttctctaa	atgt
	J G L V L Q T L R E H G L <mark>Y A K F S K</mark>	С
61	agttetggetgaggageatttegttettggggeatgtagtgteagagaatggtat	tgag E
121	jtagaccccaagaagacaaaactgtggctaactggcctagacccacttcagtaac	agag F
181	attagaagtteettggggttggeaggttactacaggaggttegteaggaetteet	aaag
241	itagtagctcctctgaccagaccagaagaatcagaagttcctgtggaccag	acctg
301	. V A P L I R L I R K N Q K F L W I D gegaggagagattegaagagettaagaagaggttgactcageaceagtgttage	L tctg
361	; E E S F E E L K K R L T S A P V L A ccatctagtgatgaggactttacagtcttttgtgatgcgtcccatatgggactggg	L Jttgt
421	' S S D E D F T V F C D A S H M G L G jtactgatgcagaatgagagggtgatcgcttatgcttctaggcagctgaagaagca	C atgag
481	′ L M Q N E R V I A Y A S R Q L K K H :tgaattaccccacacatgaccttgagatggcagcagtaatctttgtactcaagat	E :gtgg
541	」 N Y P T H D L E M A A V I F V L K M aggcattacctctatggggtgaaatgtg agatct ttacagatcataagagcctgca	W agtac
601	X H Y L Y G V K C E I F T D H K S L Q atcttgagtcagagggatctgaatctgaggcagagggggggg	Y ntqac
661	: L S Q R D L N L R Q R R W V E L L S	D
721	['] D C K I Q Y H P G K A N V V A D A L	S
781	K S L G S L S H I A A E R R P V V K	E
101	Y K L I E E G L O L E L S G T G A L	V
841	Jeceagatgagagtageaeceatgtttetggageaggtggeteagaaaeageatga A O M R V A P M F L E O V A O K O H E	iggac D
901	cggagttagtgaaggttgccaggactgttcagtcaggcaaggatagcgagtacag P E L V K V A R T V O S G K D S E Y R	jattc F
961	yacagtaaagggatcctccgctatgggagcagactatgtgtaccagatgacattgc	ggcta
1021) S K G T L R Y G S R L C V P D D T G	L
) S K G I L R Y G S R L C V P D D I G aaggagacattatgagagaggctcataatgcaagatacagcattcaccctggagc ; G D I M R E A H N A R Y S I H P G A	L cact
1081) S K G I L R Y G S R L C V P D D I G aaaggagacattatgagagaggeteataatgeaagataeageatteaeeetggage G D I M R E A H N A R Y S I H P G A agatgtateaagatttgaagaaagtttattggtggeeagegatgaagaaaga	L cact T :ggca
1081 1141	D S K G I L R Y G S R L C V P D D I G aaaggagacattatgagagagggctcataatgcaagatacagcattcacccctggagc K G D I M P G A aagatgtatcaagatttgaagaaagtttattggtggccagccgatgaagaaaga	L ccact T :ggca A :ggct
1081 1141 1201	D S K G I L R Y G S R L C V P D D I G aaaggagacattatgagagaggctcataatgcaagatacagcattcaccctggagc G D I M R E A H N A R Y S I H P G A aagatgtatcaagatttgaagaaagtttattggtggccagcgatgaagaaaga	L ccact T cggca A cggct A acttc
1081 1141 1201 1261	D S K G I L R Y G S R L C V P D D I G aaaggagacattatgagagaggctcataatgcaagatacagcattcaccctggagc G D I M R E A H N A R Y S I H P G A aagatgtatcaagatttgaagaaagttattggtggccagcgatgaagaagaagaagt M Y Q D L K K V Y W W P A M K K E V agttcgtgtcagcctgcgaagtgtgcagggggtgaagctggaacatcagaagcc P V S A C E V C Q R V K L E H Q K P ggaatgcttaacccgctacctatcccagaatggaaatgggagaatatagctatgga M L N P L P I P E W K W E N I A M D tagtggggttaccggcggcgccaacagagtggactcatagggtgatgtgga	L CCact T :ggca A :ggct A :cgct F icaga
1081 1141 1201 1261 1321	$ \begin{array}{ccccccc} & & & & & & & & & & & & & & & &$	L ccact T cggca A cggct A acttc F acaga R cggcg
1081 1141 1201 1261 1321 1381	D S K G I L R Y G S R L C V P D D I G aaaggagacattatgagagaggctcataatgcaagatacagcattcaccctggagc G D I M R E A H N A R Y S I H P G A aagatgtatcaagatttgaagaaagtttattggtggccagcgatgaagaagaagaagt M Y Q D L K K V Y W W P A M K K E V cagttcgtgtcagcctgcgaagtgtgtcagagggtgaagctggaacatcagaagc P V S A C E V C Q R V K L E H Q K P ggaatgcttaacccgctacctatcccagaatggaaatgggagaatatagctatgga M L N P L P I P E W K W E N I A M D gtagtggggttaccggcggcgccaacagagtggactccatatgggtgattgtgga V G L P A A S N R V D S I W V I V D ctccaccaaatctgctcacttcattcctgtcaggagtggctacctgtagacagtt J T K S A H F I P V R S G Y S V D K L aggtgtatgtagatgagatgacgtccaatggggttcctgttcgatagtgc	L ccact T cggca A cggct A acttc F acaga R cggcg A acagat
1081 1141 1201 1261 1321 1381 1441	D S K G I L R Y G S R L C V P D D I G aaaggagacattatgagagaggctcataatgcaagatacagcattcaccctggagc G D I M R E A H N A R Y S I H P G A aagatgtatcaagatttgaagaaagttattggtggccagcgatgaagaaaga	L ccact T cggca A ccttc F acaga R cggcg A cagat D ctagg

Figure 4. (A) Nucleotide sequence and deduced translation of *Megyp5*. The primer sequences are omitted. The highly conserved amino acid sequence block YAKFSKCEF of RT domain characteristic of *Ty3/gypsy* retrotransposons is highlighted grey. Recognition enzyme sequences are shown in bold face for *Eco* RI (underlined), *Hind* III (oval) and *BgI* II (box) used in Southern analysis of cassava; (B) Nucleotide sequences and deduced translations of *Megyp28*. Primer sequences are omitted. The highly conserved amino acid sequence block YAKFSKCEF of RT domain characteristic of *Ty3/gypsy* retrotransposons is highlighted grey. Restriction enzymes sequences are shown as detailed in Figure 4a.

В

1	ctgaggata	atatta	acaga	cctto	gaggga	acato	ggctt	gtatg	ccaa	gttct	ccaa	gtgt
61	L R I	L L	Q T accact	L t	R E	H (ј Ц Natati	Y A	K	F S	6 K	C
6 I	gagttetgg E F W	ttaag L R	gagcat S I	s S	F L	ggggg G I	tatata I I	agtgt V S	caga E	gaatg N G	gaata ; I	agag E
121	gtagacccc V D P	aagaa K K	gataga T E	aagct A	cgtgad V T	ctaact N V	zggcc≀ √ P	aagac R P	ссас т	ctcag S V	rtgaca 7 T	agag E
181	atcagaagc		gggtti	zggct	tggcta	actaca	aggag	gttcg	ttca	ggact	tctct	taag
241	attgcagct	ccttt	G L aaccag	A gatta	G I aaccag	ı ı Jaaaga	A R Aatca	gagat [.]	Q tcga	р ғ gtgga	iccgat	к tcag
301	I A A tqtqaaqaa	P L agttt	T R cqaaqa	L agctt	T R Laagaa	K 1 aqaqqt	N Q tqac	R F ttcag	E cacc	W I aqtqt	D D	Q tctq
361	C E E	SF	E E	L	K K	RI	T	S À	P	V L	Ă	L
201	P N S	aatga N E	D F	T T	V F	C I) A	S R	yagı. V	G I	G G	C
421	gtgttgatg V L M	cagaa Q N	tggtaa G K	aggto V	gatcgo I A	ttato Y A	gcttc [.] A S	tagac R Q	agcc P	gaaga K R	lggcat K H	tgag E
481	ttgaattac	cccaca	acacga	accto	ggaaat E M	ggcag	gcagt	tatct	ttgc	cctca	agat	gtgg W
541	aggcattac	ctcta [.]	п D tggggt	Laaaa	atgtg	agatct	tcac	agatc	ataa	gagec	tgca	w gcac
601	R H Y atcttgaac	L Y caqaqa	G V agagct	K Igaad	C E cttgad	I I Iqcaqa	F T aqqaqa	D H atqqq	K taga	S L actqt	y Q tqaqt	H tqac
CC1	I L N	QR	E L	N	LR	QH	R R	WV	Ē	LL	s S	D
001	Y D C	aagat K I	Q Y	H H	P G	K A	A N	V V	Lage A	D A	L L	s S
721	cggaaatca R K S	cttgg L G	cagtct S I	tatco S	ссасаt н т	cacgo T 7	gcaga A E	gagga R R	gacc P	ggtgg V V	rtgaa K	ggag E
781	ttttataag	ctcat	tgagga	agggt	tctaca	agatgo	gagtt	gtctg	gtac	aggtg	rcttt	gatt
841	f i k gcacagatg	aaagta	aaccco	ccgto	ц Q gtttct	 ggago	caagt	ggctc	ı agaa	acago	acga	ı ggac
		77 77	ΤР	V	F L	ΕÇ	V C	A Q	K	Q H	ΙE	D
901	A Q M ccagagtta	n v ataaa	gattg	ccado	ractot	tcaqt	cagg	caaaq	atag	tgagt	tcaga	alll
901	A Q M ccagagtta P E L	gtgaa V K	gattgo I A	ccago R	gactgt T V	tcagt Q S	cagg G G	caaag K D	atag S	tgagt E F	tcaga 'R	F
901 961	A Q M ccagagtta P E L gatgataag D D K	gtgaa V K gggat G I	gattgo I A cctcco L R	ccago R gctat Y	gactgt TV ggggaa GN	tcagt Q S acagao R I	cagg G G ctatg L C	caaag K D tgtac V P	atag [.] S caga [.] D	tgagt E F tgaca D I	tcaga 'R tcgga G	F gcta L
901 961 1021	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D	gtgaa V K ggggat G I attato I M	gattgo I A cctcco L R gagaga R E	ccago R gctat Y aggct A	gactgt TV ggggaa GN ccataa HN	Ctcagt Q Acagac R Atgcaa A	Cagg GG Ctatg GC Aggta RY	caaag K D tgtac V P cagtg S V	atag [.] S caga [.] D ttca H	tgagt E F tgaca D I ccctg P G	tcaga R tcgga G gagca A	F gcta L cacc
901 961 1021 1081	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac	gtgaa V K gggat G I attat I M cagga	gattgo I A cctcco L R gagaga R E tctgaa	ccago R gctat Y aggct A aggga	gactgt TV cgggaa GN ccataa HN agtgta	Caga Q R R Atgcaa A A A A A A A A A A A A A A A A A A	Caggi G G Ctatgi G C Aggtai R Y Cggcca	caaag K D tgtac V P cagtg S V agcta	atag S caga D ttca H tgaa	tgagt E F tgaca D I ccctg P G gaggg	tcaga R tcgga G gagca A gaagta	F gcta L cacc T ggca
901 961 1021 1081 1141	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg	gtgaa V K gggat G I attat I M cagga Q D tcagc	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga	ccago R gctat Y aggct A aggga G aaata	gactgt TV cgggaa GN ccataa HN agtgta VY atgtca	Caga Q Acagao R I Atgcaa A I Atgcaa W W W Agaggg	cagg G Ctatg C Aggta Aggta Y C C C C C C C C C C C C C C C C C C	caaag K D tgtac V P cagtg S V agcta A M gctgg	atag S caga D ttca H tgaa K aaca	tgagt E F tgaca D I ccctg P C gaggg R E tcaga	tcaga R tcgga gagca A gaagta V agcca	F gcta L cacc T ggca A ggct
901 961 1021 1081 1141 1201	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt	gtgaa V K gggat G I attat I M cagga Q D tcagc S A aaccc	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco	R R gctat Y aggct A aggga G aaata I cgatt	gactgt T V cgggaa G N ccataa H N agtgta V Y atgtca C Q	Contraction of the second seco	Ccagg G Ctatg C Aggta C Gggcc V P Gggaa K Aaatg	caaag K D tgtac V P cagtg S V agcta A M gctgg L E ggaga	atag S caga D ttca H tgaa K aaca H acat	tgagt E F tgaca D I ccctg P G gaggg R F tcaga Q K agcta	tcaga R G G G G G G G G C C P C C C C C C C C C	F gcta L cacc T ggca A ggct A tttt
901 961 1021 1081 1141 1201	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt G M L	gtgaa V K gggat G I attat I M cagga Q D tcagga S A aaccc N P	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco L P	R R gctat Y aggct A aggga G aaata I cgatt I	gactgt T V cgggaa G N ccataa H N agtgta V Y atgtca C Q cccaga P E	Contractions of the second sec	Cagg G Ctatg C Aggta Y Y P Jtgaa V P Jtgaa V K Aaatg K Aaatg K	caaaga K D tgtac V P cagtg S V agcta A M gctgg L E ggaga E N	atag S Caga D ttca H tgaa K aaca H acat	tgagt E F tgaca D I ccctg P G gaggg R F tcaga Q K agcta A M	tcaga R tcgga G gaagto A aagto C P tggat I D	F gcta L cacc T ggca A ggct A tttt
901 961 1021 1081 1141 1201 1261	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt G M L gtagtgggg V V G	gtgaa V K gggat G I attat I M cagga Q D tcagga S A aaccc N P ttacc L P	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco L P ggcaao A T	R R gctat A aggct A aggga G A aaata I cgatt I catco S	yactyt TV Ggggaa GN Cataa HN Agtyta VY Atytca CQ Cccaga PE Caacao NR	L Cagt Q 2 Acagac R I Atgcaa A F Atggt Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q	Cagg G Ctatg C Aggta C Gggcc V P Gtgaa V P Gtgaa V R Aaatg C W Gactco D S	caaaga K D tgtac V P cagtg S V agcta A M gctgg L E ggaga E N catat I W	atag S Caga D ttca H tgaa K aaca H acat J gggt V	tgagt E F tgaca D I ccctg P G gaggg R E tcaga Q K agcta A M gattg I V	tcaga R G G G G G G G G C C C C C C C C C C C	F gcta L cacc T ggca A ggct A tttt F caga R
901 961 1021 1081 1141 1201 1261 1321	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt G M L gtagtgggg V V G ctcaccaaa L T K	gtgaa gtgaa V K gggat G I attat I M cagga Q D tcagc S A aaccc N P ttacc L P tctgc S A	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco L P ggcaac A T tcactt H F	R R gctat Y aggct A aggga G aaata I cgatt I catco S ccatco I	yactyt TV Ggggaa GN cataa HN agtyta VY atytca CQ cccaga PE caacao NR ccctyt PV	Classical Control Cont	Cagg Cagg Caggta Caggta Cggcc V P Gtgaa V P Gtgaa V R Caaatg C W Gactco D S Cagcaa S N	caaaga K D tgtac V P cagtg S V agcta A M gctgg L E ggaga E N catat I W ctact Y S	atag S Caga D ttca H tgaa K aaca H acat gggt V ctgt	tgagt E F tgaca D I ccctg P G gaggg R F tcaga Q K agcta A M gattg I V ggata D K	tcaga R Cgagco A Gaagto CP CQ CP CQ CP CQ CP CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ	F gcta L cacc T ggca A ggct A tttt F caga R agcg A
901 961 1021 1081 1141 1201 1261 1321 1381	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt G M L gtagtgggg V V G ctcaccaaa L T K caggtttat	gtgaa gtgaa V K gggat G I attat I M cagga Q D tcagca S A aaccca N P ttacca L P tctgc S A gtgga	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco L P ggcaac A T tcactt H F	R R R R R R R R R R R R R R R R R R R	yactgt TV Ggggaa GN cataa HN agtgta VY atgtca CQ cccaga PE caacao NR ccctgt PV caggct	Classical Control Cont	Cagg G G Ctatg G C Aggta Y P Gtgaa V P Gtgaa V P Gtgaa V P Gtgaa G N Ggacto S N Ggggt	caaag K D tgtac V P cagtg S V agcta A M gctgg L E ggaga E N catat Y S cccag	atag S caga D ttca H tgaa K aaca H acat gggt V ctgt V tttc	tgagt E F tgaca D I ccctg P G gaggg R E tcaga Q K agcta A M gattg I V ggata D K	tcaga R Cgagco A Gaagto A C P C C P C C P C C P C C P C C C C C	F gcta L cacc T ggca A ggct A tttt F caga R agcg A agat
901 961 1021 1081 1141 1201 1261 1321 1381 1441	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt G M L gtagtgggg V V G ctcaccaaa L T K caggtttat Q V Y agagggccc	gtgaa V K gggat G I attat I M cagga Q D tcagga S A aaccc S A ttaccc L P ttaccc S A gtgga V D cagtt	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco L P ggcaac A T tcactt H F tgaagt E V caccto	R R R R R R R R R R R R R R R R R R R	yactyt TV Ggggaa GN Cataa HN agtyta VY atytca CQ Cccaga PE Caacao NR Cctyt PV Caggct RL gtttto	Classific de la consecutiva de	Cagg Gagga Caggta Caggta Caggta Caggta V P Jtgaa V P Jtgaa V P Jtgaa V P Jtgaa V S S S S S S S S S S S S S S S S S S S	caaaga K D tgtac V P cagtg S V agcta A M gctgg L E ggaga E N catat Y S cccag P V gcaga	atag S Caga D ttca H tgaa K aaca H acat S Ctgt V tttc S atgc	tgagt E F tgaca D I ccctg P G gaggg R F tcaga Q K agcta A M gattg I V ggata D K tatag I V tatagg	tcaga R G G G G G G G G C C C C C C C C C C C	F ggcta L cacc T ggca A ggct A tttt F caga R aggcg A agat D cagg
901 961 1021 1081 1141 1201 1321 1381 1441 1501	A Q M ccagagtta P E L gatgataag D D K aaaggagac K G D aagatgtac K M Y cagttcgtg Q F V ggaatgctt G M L gtagtgggg V V G ctcaccaaa L T K caggtttat Q V Y agagggccc R G P ttggattc	gtgaa V K gggat G I attat I M cagga Q D tcagga S A aaccc S A gtgga V D cagtt Q F aqtac	gattgo I A cctcco L R gagaga R E tctgaa L K ctgcga C E actgco L P ggcaac A T tcactt H F tgaagt E V caccto	R R R R R R R R R R R R R R R R R R R	yactgt TV Cgggaa GN Cataa HN agtgta VY atgtca CQ Cccaga PE Caacao NR Ccctgt PV Caggct RL gtttto FW	Classific de la consecutiva de	Caggues Gaggta Caggta Caggta Caggta Caggta V P Jtgaa V P Jtgaa V P Jtgaa V P Jtgaa V S S S S S S S S S S S S S S S S S S S	caaag K D tgtac V P cagtg S V agcta A M gctgg L E ggaga E N catat Y S cccag P V gcaga Q N	atag S caga D ttca H tgaa K aaca H acat gggt C tgt V tttc S atgc A	tgagt E F tgaca D I ccctg P G gaggg R E tcaga Q K agcta A M gattg I V ggata D K tatag I V tatgg M G	tcaga R G G G G G G G G G G G G G G G G G G	F gcta L cacc T ggca A ggct A tttt F caga R agat D cagg R

Figure 4. Contd.



Figure 5. Phylogenetic analysis of 17 cassava *Ty3/gypsy*-like retrotransposons (*Megyps*). The tree is based on 17 nucleotide sequences of *pol* gene fragments (*Megyps*): 15 are partial sequences from the two ends of the ~1.6 kb gene fragments, while *Megyps* 5 and 28 were full 1.6 kb length. This is a consensus neighbor-joining unrooted tree constructed with the PHYLIP package from the distance matrix following the Kimura 2-parameter model (Kimura, 1980). Bootstrap values (100 replicates) are shown.

Locus or sequence name	Source species	Gi number
A. comosus	Ananas comosus	2995405
O. sativa	Oryza sativa	37532428
Del	Lilium henryi	19442
rHv1	Hordeum vulgare	3413486
rAt1 right	Arabidopsis thaliana	3413430
rAt1 left	Arabidopsis thaliana	3413431
Ту3-2	Saccharomyces cerevisiae	1084606
Gypsy	Drosophila melanogaster	130583

Table 1. Sources of polyprotein amino acids sequences used in comparative phylogenetic analysis with the 16 cassava *Megyps* amino acid sequences.

The table shows the name of Ty3/gypsy retrotransposons and the GI (Geneinfo identifier) number of corresponding polyprotein as well as the name of the source organisms. The *romani* elements are *rHv1* and *rAt1*.

		*	a 20	*	40	*	60	,	* 8	0 *	100	*
Megyp5	: ICLVLQI	REHGL	AK <mark>e</mark> skce <mark>e</mark> vlr	SISFIC	VVSENGIEV:	DE <mark>KK</mark> TKT <mark>V</mark> AN	WE <mark>R</mark> ETS <mark>V</mark> TE	IRS <mark>FLG</mark> LAGI	YYRRFVÇDESI	NIVAEIII	RL <mark>T</mark> RK <mark>NQ</mark> I	KFLWTDLCEES
Megyp20	: IGLVLOI	IREHGL /	AK F SKCE <mark>F</mark> LLR	SISELC	IV <mark>V</mark> S <mark>EN</mark> GI <mark>E</mark> VI	DE <mark>KRTETV</mark> AN	NE <mark>R</mark> ets <mark>n</mark> te	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIRAFIT	RL T RK NO	KFVNTDCCEES
Megyp2	: IGLVLQI	IREHGL /	AK <mark>E</mark> SKCE <mark>E</mark> NLR	SISELCE	IV V S <mark>EN</mark> GI <mark>E</mark> VI	DE KRTETVA N	NE <mark>R</mark> ets <mark>n</mark> te	IRS <mark>FLG</mark> LAGI		KIAAFIT	RL T RK <mark>NQ</mark> I	KFLWTDCCEES
Megyp19	: <mark>I</mark> GLVLQ <mark>I</mark>	IREHGL /	AKESKCE <mark>E</mark> NLR	SISFIC	IV <mark>V</mark> S <mark>ENGI</mark> EVI	DE <mark>KKTETVA</mark> N	NERETSVIK	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIAAF <mark>I</mark> T	RLTWKNQ	KFLWTDQCEES
Megyp16	: <mark>I</mark> GLVLQ <mark>I</mark>	I REHGS /	AM <mark>F</mark> SKCE <mark>F</mark> NLF	SISFIC	IV <mark>V</mark> S <mark>EN</mark> GI <mark>E</mark> VI	DE <mark>RRTETVA</mark> N	NE <mark>retsvt</mark> e	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIIAAFIT	RL <mark>T</mark> RK <mark>NQ</mark> I	KFVWT <mark>DQ</mark> CEES
Megyp21	: <mark>I</mark> GLVLQ <mark>I</mark>	IREHGL ?	AM <mark>F</mark> SKCE <mark>F</mark> NLF	SISFIC	IV <mark>V</mark> S <mark>EN</mark> GI <mark>E</mark> VI	DE <mark>RRTETVA</mark> N	NE <mark>retsvt</mark> e	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIIAAFIT	RL <mark>T</mark> RK <mark>NQ</mark> I	KFVWT <mark>DQ</mark> CEES
Megyp8	: <mark>I</mark> GLVLQ <mark>I</mark>	l r ehs l <i>i</i>	AK <mark>f</mark> skce <mark>f</mark> nlf	SISFICE	IV <mark>V</mark> S <mark>KN</mark> GI <mark>E</mark> VI	DE rrtta n	WE <mark>retsvt</mark> e	IRS <mark>FLG</mark> LAGI	YYRRFVÇDESI	KIRAFIT	RL <mark>T</mark> RK <mark>NQ</mark> I	KFLWTDQCEES
Медурб	: <mark>I</mark> GLVLQ <mark>I</mark>	i r ehg l <i>i</i>	AK <mark>F</mark> SKCE <mark>F</mark> NLR	SESELCE	IV <mark>V</mark> SENGIEVI	DE <mark>KKTETV</mark> AN	WE <mark>retsvt</mark> e	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIRAFIT	RL <mark>T</mark> RK <mark>NQ</mark> I	KFVWTDQCEES
Megyp35	: <mark>I</mark> GLVLQI	l r e hgl 1	AK F SKCE <mark>F</mark> NLR	SIS <mark>FIC</mark> E	IV <mark>V</mark> S <mark>ENGI</mark> EVI	DP <mark>KKTETV</mark> AN	NE <mark>retsvt</mark> e	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIAAFIT	RL <mark>T</mark> RK <mark>NQ</mark> I	KFLWTDQCEES
Megyp13	: <mark>I</mark> GLVLQI	IREHGL /	AM <mark>F</mark> SKCE <mark>F</mark> NLF	SIS <mark>FIC</mark> H	IV <mark>V</mark> S <mark>EN</mark> GI <mark>E</mark> MI	DE <mark>RRTEIVA</mark> N	NE retsvt e	IRS <mark>FLG</mark> LAGI	YY <mark>S</mark> RFVÇDFSI	КІААР <mark>І</mark> Т	RL <mark>T</mark> RK <mark>NQ</mark> I	KFLWTDQCEES
Megyp22	: IGLVLRT	IREHGL /	AK <mark>f</mark> skce f nlf	SISELCE	IV <mark>V</mark> S <mark>GNGIE</mark> VI	DE RRIEAN AN	-FRFTSVIE	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIAAP <mark>I</mark> T	RL <mark>T</mark> RK <mark>NQ</mark> I	KFVWTD-CEES
Megyp28	: <mark>I</mark> RIILQI	IREHGL /	AKESKCEENLR	SISTICE	IIVSENGIEVI	DP KKIEAVT N	WE RETSVT E	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIAAFLT	RL <mark>T</mark> RK <mark>NQ</mark> I	RFEWTDQCEES
Megyp24	: IGLVLQI	IREHGL?	AMES-CEENLE	SISFIC	IV V S EN GI <mark>E</mark> VI	DF REIEAVT N	NE RETSVT E	IRS <mark>FLG</mark> LAGI	YYRRFVÇDFSI	KIAAFIT	RL T RK <mark>NQ</mark> I	RFVWIDQCEES
Megyp1	: IGLVLQI	IREHDLY	AKESKCEENLR	SISFICE	IIVSENGIEA	de ksveav an	WE RETINT E	IKS <mark>FLG</mark> LAGI	YYRRFIÇDFSI	KIRAF <mark>M</mark> T	RL T KK NQ	GEVWIDQCEES
Megyp32	: IGLVLQI	IREHGLI	AMESKCEFNLR	SISTLC	W VSENGIE VI	DE KRVEAVAN	NE RETINT E	IRS <mark>FLG</mark> LAG	YYRRFVÇDFSI	XVAAEMT	RL T KK NQ I	REVNIDÇCEES
A.comosu	15 : URIVLOV	IR <mark>E</mark> KEL I	VELKECEPILE	ENAFLC	ILUSGSCIAVI	I KRIEAIK	NERLISVIE	IRS <mark>FLG</mark> LAG	YYRRFVE R FAI	STERN	RUTHRGV	KFINNDACERS
del	: IRISLQI	I RNNQL I	ARISKCER (ME	KKELG	WWSREGIWU	H VRVKAMMN	RELEKNIFE	IRS <mark>FLG</mark> LAG	YYRRFI <mark>KG</mark> FSI	angalmn	QUTERGE	NENWIKKYONS
U.Sativa	I RLVLEK	IRKHKLI I DRUGI I	AMESKCEPNLK Smrchoffnlk	E ALLG		I ASVEALTE	KAPKSVIL	IKSELGLAG	TIKKFILGFS:	319KE01	QLLKSEK	KEVWELQUQLE
rHV1	I FLVLEK	LRENGL I	AND SKULL VLF	EVELLE:			NESETTERE	IKSELGLAG. I D <mark>ori out</mark>	IIKKI <mark>I</mark> ENISI		EII Laigu II	VUTU
THUS_2	IDTWIER	TENEL	UNERSCHAUSE UNERSCHAUSE	RIDILG.	STCTONEVER	LOHNCA AT DIS	NEVENNELL FETERANDO	AODELONIA MODELONIA	VVDDFTFNOSI		IFICDES	ININ
IyJ-2 Megup18	IKOWERV	TEDNET	VERTRE SPACE	RUPPLO		SCHOCHAIN	URDERDWER 1 FLENDWER	TREFICIENT	VVDDRT <mark>KO</mark> VG		DKI NEMK	
negypio	TDTVI KO	TTDANMO	VSOFATRAKE	STRVICE	TYSKDETKS	DEFAURATOR	VEREDOWNK	VRSFICIAS	VYRMETKEEN VYRMETKEEN		NGSVSOHMSOKT	PURENETOPNA
97597			- SADA INC								NOD Y DEMINISTRY	I VIENLI XNNA
		YA	KFSKCEF									
	120	*	140	b	*	160	*	180	*	200	*	220 *
Megyp5 :	FEELM <mark>K</mark> RLTS	AF-VLA	LESSDEDET <mark>V</mark>	FCDASH	GLG <mark>OVI</mark> MÇI	NERV	IAYASRÇI	.K <mark>KH</mark> E L NYPI	T <mark>H</mark> DLEMAAVI	FVLKMWRHYLY	GVK-CEIFTDHE	SLQYILSQRDLNLRQ
Megyp20 :	FEELK <mark>K</mark> RLTS	AF-VLA	LESSDADET <mark>V</mark>	FCDASRV	/GLG <mark>CVL</mark> MC1	NERV	- IAYASRCI	.K <mark>KH</mark> ELNYP:	T <mark>H</mark> DLEMAAVI	F		
Megyp2 :	FEELKKRLT	AF-VLA	LESSDEDETV	FCDASRV	GLG <mark>OVL</mark> MO	NERV	- IAYASRCI		T <mark>H</mark> DLEMAAVI	F		
Meavp19 :	FEELKKRLTS	AF-VLA	LESSDEDETW	FCDASRV	GLG <mark>OVL</mark> MO	NERV	- IAYASRCI		T <mark>H</mark> DLEMAAVI	F		
Megyp16 :	FEELMNRLTS	AF-VLA	SSGEDETW	FCDASRV	GLGCVINCI	NORV	IAYASRCI	.K <mark>KH</mark> B L NYP:	THCLEMAAVI	F		
Megyp21 :	FEELMKRLTS	AF-VLA	SRGEDET	FCDASRV		VERV	IAYASRCI	K KHEL NYP	THCLEMAAVI	F		
Megyp8 :	FEELBRRLTS	AF-VIA	SSGEDET	FCDASRV		VERV	IAYASRCI	K KH ELNYP'	THEFEMAAVI	F		
Megyn6 :	FEELWRRLTS	AF-VIG	SSDED	FCDASRV		VERV	TAYASRCT	KKH RINYP'	THELEMAAVI	F		
Megyp35 :	FFELMKRLTS		LESSGEDETV	FCDAFRY		VERV	TAYASROT	KKH RINYP	THELEMAAN	F		
Megun13	FFFLWNRLTS	A TV- 2	SSDEDUT	FORASPI		15 RV	TAVASBOT	RKHRIMYP	Т <mark>Н</mark> СТ.ЕМААТІ	F		
Megup22 :	FEFT WEDT TO		LEASSEDET	FWFASD			TATADAQL	NTRINCIAL STREET	THEIRMAN	F		
Magun20	VERIVEDITO	A	INCNED	TODACAN			TAVAGEO	VDURTMVD'	IIIDIENAAVI IV66M3I9		OUR CETETRU	STORT NODELNI DO
Megyp20 : Megyp20 :	FEEDWARDIN	A A	CONFRICT	TODADAN TODADAN			TAVASDOT	NNNDLNIF. NNNNNNNN	IBLLENAAVI TBLLENAAVI	PALKEWRITLI	SVR-CEIFIDH	PERMITRANCTRENT AND
Megyp24 :	FEELWERLIG	AA	LED SNEDET	TODAD N			TATABAQI	NKRELNIF.	THELENAAYI THELENAAYI	F		
Megypi :	FELLARALIS	A - A	TED TIM	LODACKY		OLRV	THINGRY	N <mark>KN</mark> ELNIF.	I <mark>U</mark> LTENAAA1	·		
Megyp32 :	TODINODITI	а н.	AIERLLV	10DASKY			T1 12 CD OT					
A.comosus :	FÖFTRÖKTI I	A 1		ISDASLI I TODO OLI	GLEUVINU	DENV	TATASKČI	KETERNIF.	THELELAAV)	FALKLWRHILI	SER-CEVIIDH	VETRATEIÖKETNTKÖ
aer :	FLEERR-LTT	VE-VLI	TETRO-FEAM	TIDASLA	GLEGVENÇI	GEV	VATASRÇI	NVHENNYP.	THCLELA <mark>V</mark> VI	LITKTMKHATA	SED-FELYCDH	VETRATELÖKDTNUKÖ
U.Sativa :	FEÇLMEKLTS	AF-ILV	TEDIKKLEVI	10DASR,	GLEGVINÇI		VATASRÇI	REMEDINTE:	INCLELAAV\	HALKIWRHYLI	JNH-CDIYTDDE	VPTKATLIČZDTNUKČ
rHv1 :	FÇELMKRIVI	AS-WII	TE DÖKND JÖV	TODASRE	GLICAVIIMQI	LGRV	V <mark>S</mark> YASRÇI	NFHELNIA'	I <mark>D</mark> ULELAAVV	HALKTWRHFLI	SNH-CEVYTDHE	KETKATŁIČKETNŢĽĞ
rAtl :	TKATLASTS	EK-SFV	UKSLKKGISI	VUGTLH	DGRIVIPT	SPF						
туз-2 :	IERLEAAICI	ISE-VIV	FFNNKAN I R <mark>I</mark>	TTEASKI	IG I GAVI E E V	VDNKNKLVG	NGIFSKSI	ESAÇKNYE	AGELELLGI	KALHHFRYMLH(GKH-FTLRTDHI	ISLLSLQNKNEPARRV
Megyp18 :	FEALRKAVME	EE-VLA	LEDYSKEEEV	HICASOR	AIG <mark>QVI</mark> MQI	ER						
dAbaA :	FÇRLENILAS	EDVIIK	AFDEKKEED I .	TTCASAS	IGI G AVII SQI	EGRP	ITMISRTI	NGFEQNYA:	INERELLAIV	WALGKLONFLY	GSREINIFTDHÇ	QPLTFAVADRNTNAKI
				TDAS								

Figure 6. Alignment of the predicted amino acid sequences for 16 polyprotein fragments of cassava *Ty3/gypsy*-like retrotransposons with those of eight other plants. *Ty3-2* and *Gypsy* are included for reference (detail of the polyproteins in Table 4). Each of the clones was represented by either translation of full ~1.6 kb or partial sequences from the two ends of the POL fragments. Colour blocking indicates sequence conservation. Black = 100% identity, deep grey = > 80% identity, light grey = > 60% identity and non-shaded = < 60% identity. Letters in bold orange colour below the alignment and boxed regions labelled with small letter a-e indicate the key residues conserved in all related enzymes as explained in the text.

cassava cultivars. Following high stringency washes (0.2 X SSC, 0.1% SDS, 65° C), strong signals were observed

in all the digests (Figure 8) and the autoradiograph required a short exposure time. This showed that, the



Figure 6. Contd.

Megyp5 sequence and its homologues were highly repeated within these genomes. The probe contained one each of the *Bgl* II, *Eco* RI and *Hin*d III recognition sites (Figure 4a), which could explain the presence of

two bands in the DNA digestions by each of these enzymes. However, multiples of two hybridising bands were observed for each of the three enzymes (Figure 8), indicating that multiple copies of *Megyp5* and relatives

			*	480	*	500	*	520	÷	540	*
Megyp5	÷	<mark>A</mark> MDFVVGLE	AASNRV	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVP <mark>V</mark> SIVS <mark>D</mark> R <mark>G</mark>	PQFTSRFWRS	l <mark>qna</mark> mgt <mark>r</mark> l	DFST <mark>A</mark> F
Megyp20	÷	<mark>A</mark> MDFVVGLE	AASNRV	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSREWRS	l <mark>qna</mark> mgt <mark>r</mark> l	dfst <mark>a</mark> f
Megyp2	÷	<mark>A</mark> MDFVVGLE	ATSNRL	D <mark>S</mark> IWVVVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFT <mark>FR</mark> FWRS	l <mark>qda</mark> mgt <mark>r</mark> l	dfst <mark>a</mark> f
Megyp19	÷	<mark>A</mark> MDFVVGLE	ATSNRL	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSREWRS	l <mark>qda</mark> mgt <mark>r</mark> l	dfst <mark>a</mark> f
Megyp16	÷	<mark>A</mark> MDFVVGLE	ATSNRL	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSREWRS	l <mark>qna</mark> mgt <mark>r</mark> l	dfst <mark>a</mark> f
Megyp21	÷	<mark>A</mark> MDFVVGLE	ATSNRL	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSREWRS	l <mark>qda</mark> mgt <mark>r</mark> l	dfst <mark>a</mark> f
Megyp8	÷	<mark>A</mark> MDFVVGLE	ATSNRL	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSREWRS	l <mark>qna</mark> mgt <mark>r</mark> l	dfst <mark>a</mark> f
Медурб	÷	<mark>A</mark> MDFVVGLE	AASNRL	D <mark>S</mark> IWVIVDRLT	K <mark>S</mark> AHFI <mark>A</mark> VRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSRSWRS	l <mark>çsa</mark> mgt r l	dfst <mark>a</mark> f
Megyp35	;	<mark>A</mark> MDFVVGLE	AASNRL	D <mark>S</mark> INVIVDRLT	K <mark>S</mark> AHFI <mark>H</mark> VRS	GYSVEKLVQ	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS D R <mark>G</mark>	PQFTSRFWRS	l <mark>qna</mark> mgt <mark>r</mark> l	DFST <mark>A</mark> F
Megyp13	;	AMDFVVGLE	AASNRL	D <mark>S</mark> IWVIVDRLT.	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>I</mark> SIVS <mark>D</mark> R <mark>G</mark>	PQFTS <mark>R</mark> FWRS	l <mark>qna</mark> mgt <mark>r</mark> l	DFST <mark>T</mark> F
Megyp22	;	AMDFIVGLE	AASNRL	D <mark>S</mark> IWVIVDRLT.	K <mark>S</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>K</mark>	W <mark>HVD</mark> EIVRLHO	GVE <mark>I</mark> SIVS <mark>D</mark> R <mark>G</mark>	PQFTS <mark>R</mark> FW <mark>W</mark> S	l <mark>qna</mark> mgt <mark>r</mark> l	DFST <mark>A</mark> F
Megyp28	;	AMDFVVGLE	AT SNRL	D <mark>S</mark> IWVIVDRLT.	K <mark>S</mark> AHFIPVRS	NYS <mark>VE</mark> KLA <mark>Q</mark>	VYV D EVVRLH(GVE <mark>V</mark> SIVS <mark>D</mark> R <mark>G</mark>	PQFTSREWRS	l <mark>qna</mark> mgt <mark>r</mark> l	DFST <mark>A</mark> F
Megyp24	;	AMDFVVGL <mark>I</mark>	AASNRL	D <mark>S</mark> IWVIVDRLT.	K <mark>S</mark> AHFIPVRS	NYS <mark>VE</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIVRLHO	GVE <mark>V</mark> SIVS <mark>D</mark> R <mark>G</mark>	PQFTS <mark>R</mark> FWRS	l <mark>qsa</mark> mgt r l	DFST <mark>A</mark> F
Megyp1	;	AMDFVVGLE	AASNRL	D <mark>ST</mark> WVIVDRLT.	K <mark>S</mark> AHFIPVRS	GYSMEKLAQ	VYV <mark>D</mark> EVVRLHO	GIP <mark>V</mark> SIVS D R <mark>G</mark>	PQLTSRFWRS	l <mark>qna</mark> mgt <mark>r</mark> l	dfs <mark>ia</mark> f
Megyp32	;	AMDFVVGL	AASNRL	D <mark>S</mark> IWVVVDRLT	K <mark>F</mark> AHFIPVRS	GYS <mark>VD</mark> KLA <mark>Q</mark>	VYV <mark>D</mark> EIV <mark>G</mark> LHO	GVE <mark>I</mark> SIVS <mark>D</mark> R <mark>G</mark>	PQFTSRFWRS	l <mark>qna</mark> mgt <mark>r</mark> l	DFST <mark>v</mark> f
A.comosus	;	TMDFV <mark>T</mark> GLE	RSQAGH	D <mark>A</mark> IWVIVDRLT.	K <mark>S</mark> AHFIPI <mark>H</mark> I	TNT <mark>GE</mark> RLA <mark>Q</mark>	VYL <mark>D</mark> EIVRLH(GVE <mark>T</mark> SIVS D R <mark>D</mark>	TREVSHEWRS	l <mark>qda</mark> lgt <mark>r</mark> l	DFSTAF
del	;	LMDFIIG <mark>F</mark> e	LSKRCH	D <mark>S</mark> IWVIVDR <mark>F</mark> T.	K <mark>S</mark> AHFIPI <mark>H</mark> I	TIS <mark>GKD</mark> LA-	LYI <mark>K</mark> EIIRLH(GIF <mark>T</mark> TIVT D R <mark>D</mark>	TKFTS <mark>R</mark> FW <mark>G</mark> S	l <mark>K-S</mark> lgT <mark>E</mark> l	FSTAF
0.sativa	;	<mark>G</mark> MDFIVGLE	KTATGY	D <mark>S</mark> IWVIMDRLT	K <mark>TA</mark> RFIPVKI	NYS <mark>SA</mark> KLA <mark>E</mark>	LYM <mark>TR</mark> IV <mark>C</mark> LHO	GVE <mark>kr</mark> iis d r <mark>g</mark>	T <mark>ç</mark> ftshfw ek	V HEA LGS <mark>Y</mark> L	AFSTAY
rHv1	;	LEVGQA	RHGFHH	G]	TESAFRILD	GCSRP	FDKG	SSE		
rAt1	÷	SLDFI <mark>D</mark> GLE	KSE-GY	D <mark>V</mark> V <mark>L</mark> VVVDRLS	N <mark>YG</mark> HFVPLM <mark>P</mark>	IFYT <mark>TKS</mark> VA <mark>D</mark>	IFL <mark>H</mark> EIVRLHO	FFESMVSDRD	KIFISNFW <mark>S</mark> S	l fksh gt <mark>s</mark> l	DKST <mark>S-</mark>
Ту3-2	÷	SMDFV <mark>T</mark> GLE	PT <mark>SN</mark> NL	N <mark>MIL</mark> VVVDR F S.	K <mark>r</mark> ahfi <mark>at</mark> r	TLDATQIID	LERYIESYHO	FFRTITSDRD	VRMTADKYQE	I tkr lg iks	TMSSAN
Megyp18	;	S <mark>MDFIIGLE</mark>	-RVNGY	RNI <mark>M</mark> VVVDR <mark>F</mark> S.	K <mark>y</mark> avfv alpe	KFDAKDTAR	LE <mark>FRD</mark> VVK <mark>YW</mark> (GIF <mark>R</mark> SIIS D R <mark>D</mark>	TR F <mark>VG</mark> k fw <mark>s</mark> e	l <mark>fki</mark> lgt <mark>d</mark> l	NFST <mark>S</mark> F
gypsy	;	HIDIFS	TDR	KLELTCICKES	NV-IVQE-VV	SRIIVDITA	PLIQIINLFPI	IKTVYCDNEP	AFNSETVISM	LKNSFGIDI	VNAPPL
		D -	60 -					D -			

Figure 6. Contd.

were integrated in the genome. Many of the bands are very strong and distinct but there are few weak ones suggesting that *Megyp*5 is cross hybridising with sequences highly homologous to the probe, represented by the strong major bands, as well as related diverged fragments, seen as weak signals. The cultivars showed no clear polymorphism of hybridisation fragments with *Megyp*5 probe used (Figure 8).

DISCUSSION

The detection of Ty3/gypsy-like retrotransposons using heterologous primers based on conserved domain of RT in PCRs has not been efficient due to the relatively high sequence heterogeneity among these elements (Su and Brown, 1997). The main problem with the use of these primers alone has been in the frequent amplification of other sequences (other retroelements, transposons and non-transposons), in addition to the desired Ty3/gypsy sequences. For instance, of forty four sequenced clones following genomic DNA PCR amplification using primers based on the conserved RT domain in *Brassica* sp, only twenty were similar to any of the known transposon types and just fifteen were Ty3/gypsy-like of all known lineages (Alix and Heslop-Harrison, 2004). In contrast, the use of degenerate PCR primers anchored on both the integrase and reverse transcriptase (Suoniemi et al., 1998) has proved a more robust and reliable tool, as the design of these primers exploited the differences in the domainorder between the Ty1/copia and Ty3/gypsy groups of retrotransposons. The usefulness of these primer sets has been confirmed here by the isolation and characterisation of the polyprotein fragment diagnostic of Ty3/gypsy-like retrotransposon in cassava. This has enabled a study of diversity of this group of retrotran-



Figure 7. Comparative phylogenetic analysis of 16 cassava Ty3/gypsy-like retrotransposons (*Megyps*) with other eight from other organisms. The tree is based on predicted amino acid sequences of *pol* gene fragments. This is a consensus neighbor-joining unrooted tree constructed with PHYLIP package. Distance matrix used the Jones-Taylor-Thornton model (Jones et al., 1992). Three groups of Ty3/gypsy-like retrotransposons were revealed. The cassava elements are indicated with arrowheads. Fourteen of them clustered into two monophyletic groups but *Megyp*18 and *Megyp*22 associated with *Gypsy* and *Ty3-2* group. The identities of other sequences used in comparative analyses with cassava's are as in Table 2. Bootstrap values (100 replicates) = > 45% is shown.



Figure 8. Southern blot analysis of *Ty3/gypsy*-like polyprotein of 12 cassava cultivars. 10 μg of genomic DNA from each of the cassava cultivars; lanes 1 (MGA1); 2 (MNGA2); 3 (MDOM5); 4 (MNGA19); 5 (MCOL22); 6 (CMC40); 7 (MVEN77); 8 (CG402); 9 (SM627); 10 (SM985); 11 (SM1088); 12 (CM2177) were digested with *Bgl* II, *Eco* RI or *Hind* III. The digested DNAs were separated on 0.8% agarose gels, transferred to nylon membrane and hybridised with the *Megyp5* probe. *Hind* III-digested lambda DNA was used as a DNA size marker (M).

sposons in this important food crop. This approach has also provided better information from all the conserved domains for better resolution of the Ty3/gypsy group than could be afforded by the use of individual enzymatic domains (Springer and Britten, 1993; Wright and Voytas, 1998; Xiong and Eickbush, 1990).

Alignments of the nucleotide sequences of cassava Ty3/gypsy clones (*Megyps*) and subsequently, the inferred phylogenetic tree led to the identification of diverse members of this group of elements in cassava. In addition, the predicted translation of *Megyps*, aligned with other plant Ty3/gypsy sequences, revealed the presence of conserved residues established to be critical for enzymatic activity of integrase, reverse transcriptase and RNase H (Campbell and Ray, 1993; Kedar and Khan, 1990; Baker and Luo, 1994; Kulkosky et al., 1992) and proving that they represent authentic Ty3/gypsy-like retrotransposon sequences and suggesting that they were probably derived from recently active elements.

Phylogenetic analysis of cassava and other plant Ty3/gypsy polyproteins revealed a level of heterogeneity in cassava elements that has not been reported in many of plant Ty3/gypsy group retrotransposons. Most cassava Ty3/gypsy-like retrotransposons are clustered into monophyletic sub-groups (Figure 7). The groupings were supported by bootstrap values of 49 and 45%. The low bootstrap values are most probably due to the heterogeneity of the cassava sequences. *Megyp*18 clustered closely with Ty3-2 retroelements (Figure 7) suggesting that, *Megyp*18 represents the retrovirus lineage of Ty3/gypsy retrotransposons in the genome of cassava.

The findings in this study also support the suggestion

that, the use of primers based on conserved domains of RT, which have proved inefficient for the isolation of plants Tv3/qvpsv-like retrotransposons, could be the limiting factor in the study of the diversity and heterogeneity among plants Ty3/gypsy-like retrotransposons. In fact, the availability of the whole genome sequence has revealed the presence of seven families of Ty3/gypsy-like retrotransposons in A. thaliana (Wright and Voytas, 2002). Availability of more Ty3/gypsy-like sequences from other plants may give a better picture of the diversity of this group of retrotransposons among plants. In their study, Wright and Voytas (2002) further used primers specific for the conserved domains of RT of endogenous retroviruses lineage of Ty3/gypsy-like retrotransposons to make a survey of this family of retrotransposons among plants. The PCR assay revealed that, they are almost universally present in genome of dicots and old-world monocots (Wright and Voytas, 2002). Their ubiquitous nature and potential for horizontal transfer by infection implicates these retrotransposons as important vehicles for plant genome evolution (Wright and Voytas, 2002).

Also of interest is the fact that, Ty3/gypsy-like retrotransposons from a single or related plant species were clustered in a subfamily indicating that, sequence divergence during vertical transmission has a major influence on the evolution of this group of retrotransposons in plants. The presence of more than one family of Ty3/gypsy-like retrotransposons in one plant species indicates that, the retrotransposons of a family could evolve independently within a species without affecting the evolution of the members of other families. Southern hybridisation supports the diversity identified by sequencing and highlights that, multiple copies of *Megyps* are integrated in the genome of all cassava cultivars tested. However, these cultivars have the same pattern of hybridisation with the three different restriction enzymes (Bgl II, Eco RI or Hind III) digestion of the genomic DNA. This suggests that, there are no recent retrotransposition activities among these cassava elements. However, a unique distribution of Ty3/gypsy-like sequences was found for each of the four basic genomes of Hordeum genus except for the subspecies H. vulgare and H. spontaneum that were reported to show no polymorphism of hybridisation fragments with all the Ty3/gypsy clones used as probe (Vershinin et al., 2002). Also, study on the genomic organization of the Ty3/gypsy-like retrotransposons of different oil palm species and accessions by southern hybridisation revealed minor differences (Kubis, et al., 2003).

ACKNOWLEDGEMENTS

The first author would like to acknowledge funding from the Commonwealth Scholarship Commission, UK. This publication is part of an output from a research project funded by the United Kingdom, Department for International Development (DFID) for the benefit of developing countries: R8156 Crop Post-Harvest Programme. The views expressed are not necessarily those of DFID. This work has been carried out in compliance with the current laws governing genetic experimentation in the U.K.

REFERENCES

- Alix K, Heslop-Harrison JS (2004). The diversity of retroelements in diploid and allotetraploid brassica species. Plant Mol Biol. 54: 895-909.
- Altschul SF Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215: 403-410.
- Baker T, Luo L (1994). Identification of residues in the mu transposase essential for catalysis. PNAS, 91: 6654-6658.
- Barber AM, Hizi A, Maizel JV Jr, Hughes SH (1990). Hiv-1 reverse transcriptase: Structure predictions for the polymerase domain. AIDS Res. Hum. Retroviruses, 6: 1061-1072.
- Campbell AG, Ray DS (1993). Functional complementation of an escherichia coli ribonuclease h mutation by a cloned genomic fragment from the trypanosomatid crithidia fasciculata. Proc. Natl. Acad. Sci. USA, 90: 9350-9354.
- Dellaporta SL, Wood J, Hicks JB (1983). A plant DNA minipreparation. Plant Mol. Biol. Rep. 1: 19-21.
- Felsenstein J (2004). Phylip: Phylogeny inference package, version 3.63. University of Washington, Seattle, WA.
- Finnegan DJ (1992). Transposable elements. Curr. Opin. Genet. Dev. 2: 861-867.
- Flavell AJ, Dunbar E, Anderson R, Pearce SR, Hartley R Kumar A (1992). Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res. 20: 3639-3644.

- Gbadegesin MA, Wills MA, Beeching JR (2008). Diversity of Itrretrotransposons and enhancer/suppressor mutator-like transposons in cassava (manihot esculenta crantz). Mol. Genet. Genomics, 280: 305-317.
- Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994). Clustal w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.
- Hirochika H, Hirochika R (1993). Ty1-copia group retrotransposons as ubiquitous components of plant genomes. Jpn. J. Genet. 68: 35-46.
- Kedar P, Khan AS (1990). Nucleotide sequence of the integrase (in) gene of an endogenous murine leukemia retroviral DNA. Nucleic Acids Res. 18: p. 4022.
- Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111-120.
- Kubis SE, Castilho AM, Vershinin AV, Heslop-Harrison JS (2003). Retroelements, transposons and methylation status in the genome of oil palm (elaeis guineensis) and the relationship to somaclonal variation. Plant Mol. Biol. 52: 69-79.
- Kulkosky J, Jones KS, Katz RA, Mack JP, Skalka AM (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. Mol Cell Biol. 12: 2331-2338.
- Mann C (1997). Reseeding the green revolution. Science 277: 1038-1043.
- Saitou N, Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.
- Sambrook J, Fritsch E, Maniatis T (1989). Molecular cloning: A laboratory manual, Cold Spring Harbour Laboratory Press, Cold Spring Harbour.
- Springer M, Britten R (1993). Phylogenetic relationships of reverse transcriptase and rnase h sequences and aspects of genome structure in the gypsy group of retrotransposons. Mol. Biol Evol. 10: 1370-1379.
- Staginnus C, Huettel B, Desel C, Schmidt T, Kahl G (2001). A pcrbased assay to detect en/spm-like transposon sequences in plants. Chromosome Res. 9: 591-605.
- Su PY, Brown TA (1997). Ty3/gypsy-like retrotransposon sequences in tomato. Plasmid, 38: 148-57.
- Suoniemi A, Tanskanen J, Schulman AH (1998). Gypsy-like retrotransposons are widespread in the plant kingdom. Plant J. 13: 699-705.
- Vershinin AV, Druk A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS (2002). Lines and gypsy-like retrotransposons in *hordeum* species. Plant Mol Biol. 49: 1-14.
- Wright DA, Voytas DF (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 DA, Voytas DF (2002). Athila4 of arabidopsis and calypso of soybean define a lineage of endogenous plant retroviruses. Genome Res. 12: 122-131.
- Xiong Y, Eickbush T (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 9: 3353-3362.