African Journal of Biotechnology Vol. 10(20), pp. 3951-3963, 16 May, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.1315
ISSN 1684–5315 © 2011 Academic Journals

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. Beeching[2]

[1]Department of Biochemistry, University of Ibadan, Ibadan 200005, Nigeria.
[2]Department 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.

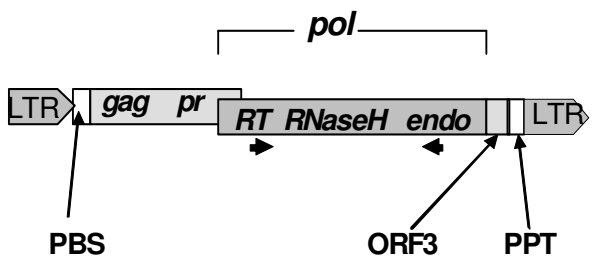

Figure 1. Structural organisation of *Ty3/gypsy*-like retrotransposons. They are bounded at their termini by long terminal repeats (LTRs). The primer binding site (PBS) and polypurine tract (PPT) is represented as grey rectangles. In between PBS and PPT (coloured boxes), are the two open reading frames (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®-T Easy vector (Promega) and used to transform competent *Escherichia coli* DH5α 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%

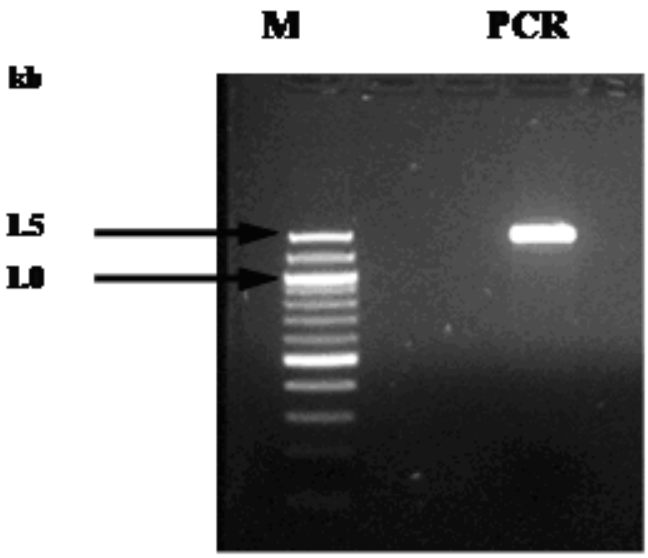

Figure 2. PCR amplification of *Ty3*/gypsy-*like* polyprotein fragment from cassava genomic DNA. PCR product was run on a 0.8% agarose gel stained with ethidium bromide. The size marker (lane M) is bioline DNA 100 bp ladder, while the PCR product is shown in the right lane

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 *Megyp*1, *Megyp*2…. *Megy*36) 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, *Megyp*5 and *Megyp*28, 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 *Megyp*s 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 quick check and provides strong evidence for it being part of the polyprotein sequence in all of the cassava *Ty3/gypsy*-like 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 *Megyp*s (left and right fragments for 15; full ~1.6 kb fragment sequences for *Megyp*s 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 consensus-unrooted tree was derived from 100 equally parsimonious trees. The consensus tree was drawn using TREEVIEW

```
Me: 55   CKIYQRVKLEHQKPAGMLNPLPIPEWKWENVVMDFVVGLPATSNRLNSIWVIVDRLTKSA 234
         C + Q+VK+EHQKPAG LNPL IPEWKWEN+ MDFVVG P ++   N+IWV+VDRLTKSA
Oe: 1535 CMVCQQVKVEHQKPAGWLNPLDIPEWKWENITMDFVVGFPKSAIGNNAIWVVVDRLTKSA 1714

Me: 235  HFIPVRSGYSVDKLAQVYVEEIIRLHGAPVSIVSDRRLQFTSRSWRSLQNAMGTRLDLST 414
         HF+PV+  +S+D+LAQ+Y+++++RL G PVSIVSDR L+FTS+ W+SLQ AMGT+L+ ST
Oe: 1715 HFLPVKMTFSLDQLAQLYIKDVVRLCGVPVSIVSDRDLRFTSKFWKSLQGAMGTKLNFST 1894

Me: 415  AFHPQTDGQSER 450
         A+HPQTDGQSER
Oe: 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 *Megyp*s, 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 *Megyp*s (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, *Megyp*5 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 *Megyp*s as homologous GMLNPLPI. Also present in the aligned *Megyp*s 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 *Megyp*s 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 *Megyp*s 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, *Megyp*18 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, *Megyp*18 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 *Megyp*s 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

A

```
   1 ctggggttggtcttgcagactttgagggaacatggcttgtatgccaagttctctaaatgt
     L  G  L  V  L  Q  T  L  R  E  H  G  L  Y  A  K  F  S  K  C
  61 gagttctggctgaggagcatttcgttcttggggcatgtagtgtcagagaatggtattgag
     E  F  W  L  R  S  I  S  F  L  G  H  V  V  S  E  N  G  I  E
 121 gtagaccccaagaagacaaaaactgtggctaactggcctagacccacttcagtaacagag
     V  D  P  K  K  T  K  T  V  A  N  W  P  R  P  T  S  V  T  E
 181 attagaagtttcttgggtttggcaggttactacaggaggttcgttcaggacttctcaaag
     I  R  S  F  L  G  L  A  G  Y  Y  R  R  F  V  Q  D  F  S  K
 241 atagtagctcctctgaccagactgaccaggaagaatcagaagtttctgtggaccgacctg
     I  V  A  P  L  T  R  L  T  R  K  N  Q  K  F  L  W  T  D  L
 301 tgcgaggagagtttcgaagagcttaagaagaggttgacttcagcaccagtgttagctctg
     C  E  E  S  F  E  E  L  K  K  R  L  T  S  A  P  V  L  A  L
 361 ccatctagtgatgaggactttacagtcttttgtgatgcgtcccatatgggactgggttgt
     P  S  S  D  E  D  F  T  V  F  C  D  A  S  H  M  G  L  G  C
 421 gtactgatgcagaatgagagggtgatcgcttatgcttctaggcagctgaagaagcatgag
     V  L  M  Q  N  E  R  V  I  A  Y  A  S  R  Q  L  K  K  H  E
 481 ttgaattaccccacacatgaccttgagatggcagcagtaatctttgtactcaagatgtgg
     L  N  Y  P  T  H  D  L  E  M  A  A  V  I  F  V  L  K  M  W
 541 aggcattacctctatggggtgaaatgtgagatcttacagatcataagagcctgcagtac
     R  H  Y  L  Y  G  V  K  C  E  I  F  T  D  H  K  S  L  Q  Y
 601 atcttgagtcagagggatctgaatctgaggcagaggaggtgggtggagctgctgagtgac
     I  L  S  Q  R  D  L  N  L  R  Q  R  R  W  V  E  L  L  S  D
 661 tatgattgcaagattcagtatcatccgggtaaggcgaatgtcgtggcagacgccctaagc
     Y  D  C  K  I  Q  Y  H  P  G  K  A  N  V  V  A  D  A  L  S
 721 cggaagtcactaggcagtctatcccacatcgcggcagagaggagaccagtggtgaaggaa
     R  K  S  L  G  S  L  S  H  I  A  A  E  R  R  P  V  V  K  E
 781 ttctacaaagcttattgaggaaggtctacagttggagttgtctggtacaggtgccttagtg
     F  Y  K  L  I  E  E  G  L  Q  L  E  L  S  G  T  G  A  L  V
 841 gcccagatgagagtagcacccatgtttctggagcaggtggctcagaaacagcatgaggac
     A  Q  M  R  V  A  P  M  F  L  E  Q  V  A  Q  K  Q  H  E  D
 901 ccggagttagtgaaggttgccaggactgttcagtcaggcaaggatagcgagtacagattc
     P  E  L  V  K  V  A  R  T  V  Q  S  G  K  D  S  E  Y  R  F
 961 gacagtaaagggatcctccgctatgggagcagactatgtgtaccagatgacattgggcta
     D  S  K  G  I  L  R  Y  G  S  R  L  C  V  P  D  D  I  G  L
1021 aaaggagacattatgagagaggctcataatgcaagatacagcattcaccctggagccact
     K  G  D  I  M  R  E  A  H  N  A  R  Y  S  I  H  P  G  A  T
1081 aagatgtatcaagatttgaagaaagtttattggtggccagcgatgaagaaagaagtggca
     K  M  Y  Q  D  L  K  K  V  Y  W  W  P  A  M  K  K  E  V  A
1141 cagttcgtgtcagcctgcgaagtgtgtcagagggtgaagctggaacatcagaagccggct
     Q  F  V  S  A  C  E  V  C  Q  R  V  K  L  E  H  Q  K  P  A
1201 ggaatgcttaacccgctacctatcccagaatggaaatgggagaatatagctatggacttc
     G  M  L  N  P  L  P  I  P  E  W  K  W  E  N  I  A  M  D  F
1261 gtagtggggttaccggcggcgtccaacagagtggactccatatgggtgattgtggacaga
     V  V  G  L  P  A  A  S  N  R  V  D  S  I  W  V  I  V  D  R
1321 ctcaccaaatctgctcacttcattcctgtcaggagtggctactctgtagacaagttggcg
     L  T  K  S  A  H  F  I  P  V  R  S  G  Y  S  V  D  K  L  A
1381 caggtgtatgtagatgagatcgtcaggctgcatggggttcctgtttcgatagtgtcagat
     Q  V  Y  V  D  E  I  V  R  L  H  G  V  P  V  S  I  V  S  D
1441 agagggccccagttcacctccagattttggcggagtctgcagaatgccatgggtactagg
     R  G  P  Q  F  T  S  R  F  W  R  S  L  Q  N  A  M  G  T  R
1501 ttggatttcagtactgccttc 1521
     L  D  F  S  T  A  F
```

Figure 4. (A) Nucleotide sequence and deduced translation of *Megyp*5. 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), *Hin*d III (oval) and *Bgl* II (box) used in Southern analysis of cassava; (B) Nucleotide sequences and deduced translations of *Megyp*28. 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.

B

```
   1 ctgaggataatattacagaccttgagggaacatggcttgtatgccaagttctccaagtgt
     L  R  I  I  L  Q  T  L  R  E  H  G  L  Y  A  K  F  S  K  C
  61 gagttctggttaaggagcatatcattcttggggcatatagtgtcagagaatggaatagag
     E  F  W  L  R  S  I  S  F  L  G  H  I  V  S  E  N  G  I  E
 121 gtagaccccaagaagatagaagctgtgactaactggccaagacccacctcagtgacagag
     V  D  P  K  K  I  E  A  V  T  N  W  P  R  P  T  S  V  T  E
 181 atcaaagctttcttgggtttggctggctactacaggaggttcgttcaggacttctctaag
     I  R  S  F  L  G  L  A  G  Y  Y  R  R  F  V  Q  D  F  S  K
 241 attgcagctcctttaaccagattaaccagaaagaatcagagattcgagtggaccgatcag
     I  A  A  P  L  T  R  L  T  R  K  N  Q  R  F  E  W  T  D  Q
 301 tgtgaagaaagtttcgaagagcttaagaagaggttgacttcagcaccagtgttagctctg
     C  E  E  S  F  E  E  L  K  K  R  L  T  S  A  P  V  L  A  L
 361 ccaaacagcaatgaggatttcacagtgttctgtgatgcatccagagtaggcctgggttgt
     P  N  S  N  E  D  F  T  V  F  C  D  A  S  R  V  G  L  G  C
 421 gtgttgatgcagaatggtaaggtgatcgcttatgcttctagacagccgaagaggcatgag
     V  L  M  Q  N  G  K  V  I  A  Y  A  S  R  Q  P  K  R  H  E
 481 ttgaattaccccacacacgacctggaaatggcagcagttatctttgccctcaagatgtgg
     L  N  Y  P  T  H  D  L  E  M  A  A  V  I  F  A  L  K  M  W
 541 aggcattacctctatggggtaaaatgtgagatctcacagatcataagagcctgcagcac
     R  H  Y  L  Y  G  V  K  C  E  I  F  T  D  H  K  S  L  Q  H
 601 atcttgaaccagagagagctgaacttgaggcagaggagatgggtagaactgttgagtgac
     I  L  N  Q  R  E  L  N  L  R  Q  R  R  W  V  E  L  L  S  D
 661 tacgattgcaagatccagtaccatccgggtaaggctaatgtagtagctgatgccttaagc
     Y  D  C  K  I  Q  Y  H  P  G  K  A  N  V  V  A  D  A  L  S
 721 cggaaatcacttggcagtctatcccacatcacggcagagaggagaccggtggtgaaggag
     R  K  S  L  G  S  L  S  H  I  T  A  E  R  R  P  V  V  K  E
 781 ttttataagctcattgaggagggtctacagatggagttgtctggtacaggtgctttgatt
     F  Y  K  L  I  E  E  G  L  Q  M  E  L  S  G  T  G  A  L  I
 841 gcacagatgaaagtaacccccgtgtttctggagcaagtggctcagaaacagcacgaggac
     A  Q  M  K  V  T  P  V  F  L  E  Q  V  A  Q  K  Q  H  E  D
 901 ccagagttagtgaagattgccaggactgttcagtcaggcaaagatagtgagttcagattt
     P  E  L  V  K  I  A  R  T  V  Q  S  G  K  D  S  E  F  R  F
 961 gatgataaggggatcctccgctatgggaacagactatgtgtaccagatgacatcgggcta
     D  D  K  G  I  L  R  Y  G  N  R  L  C  V  P  D  D  I  G  L
1021 aaaggagacattatgagagaggctcataatgcaaggtacagtgttcaccctggagccacc
     K  G  D  I  M  R  E  A  H  N  A  R  Y  S  V  H  P  G  A  T
1081 aagatgtaccaggatctgaagggagtgtattggtggccagctatgaagagggaagtggca
     K  M  Y  Q  D  L  K  G  V  Y  W  W  P  A  M  K  R  E  V  A
1141 cagttcgtgtcagcctgcgaaatatgtcagagggtgaagctggaacatcagaagccggct
     Q  F  V  S  A  C  E  I  C  Q  R  V  K  L  E  H  Q  K  P  A
1201 ggaatgcttaacccactgccgattccagagtggaaatgggagaacatagctatggatttt
     G  M  L  N  P  L  P  I  P  E  W  K  W  E  N  I  A  M  D  F
1261 gtagtggggttaccggcaacatccaacagactagactccatatgggtgattgtggacaga
     V  V  G  L  P  A  T  S  N  R  L  D  S  I  W  V  I  V  D  R
1321 ctcaccaaatctgctcacttcatccctgttaggagcaactactctgtggataagttagcg
     L  T  K  S  A  H  F  I  P  V  R  S  N  Y  S  V  D  K  L  A
1381 caggtttatgtggatgaagttgtcaggctgcatggggtcccagtttctatagtgtcagat
     Q  V  Y  V  D  E  V  V  R  L  H  G  V  P  V  S  I  V  S  D
1441 agagggccccagttcacctccaggttttggcggagtctgcagaatgctatgggtaccagg
     R  G  P  Q  F  T  S  R  F  W  R  S  L  Q  N  A  M  G  T  R
1501 ttggatttcagtactgccttc 1521
     L  D  F  S  T  A  F
```

Figure 4. Contd.

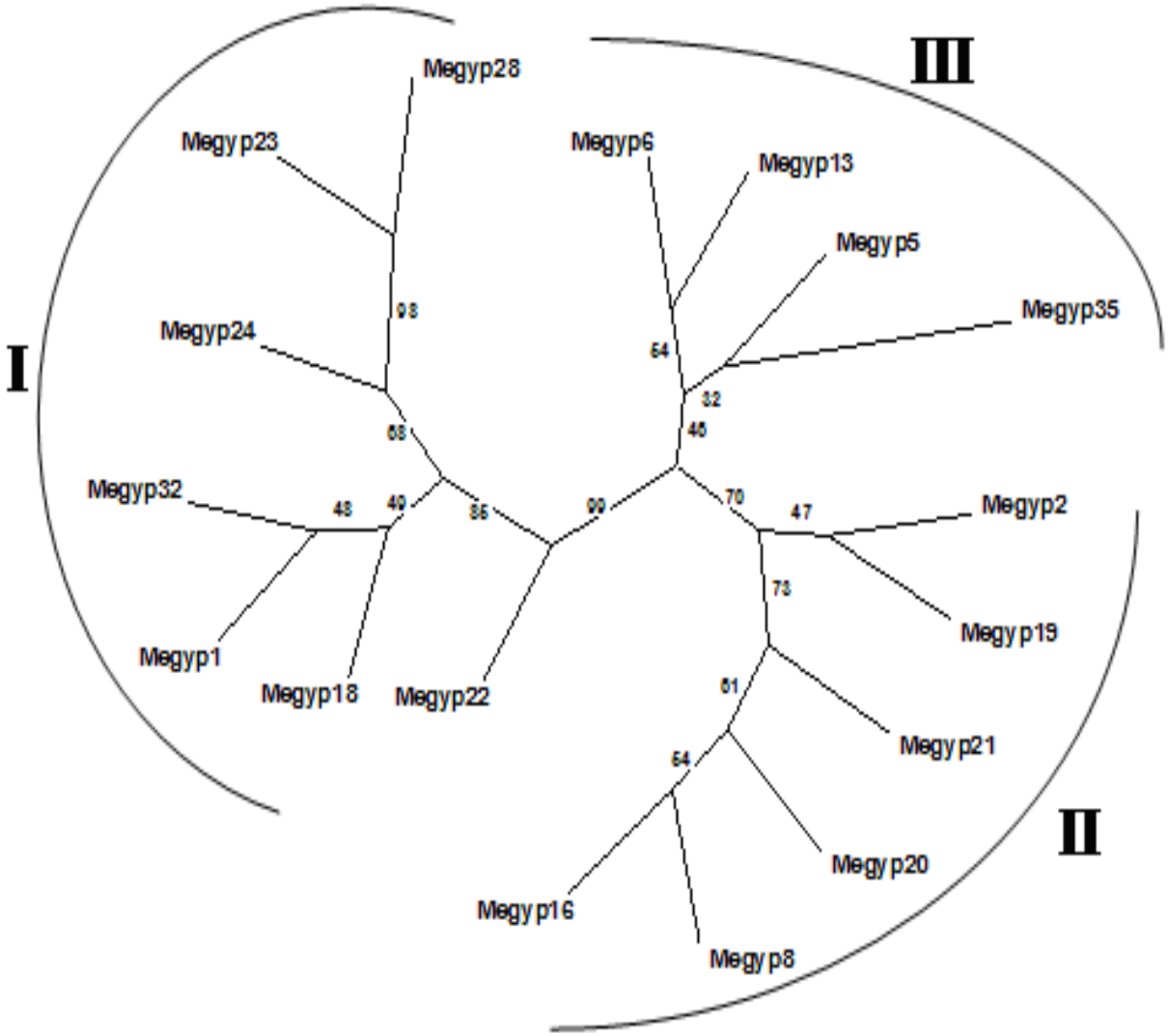

Figure 5. Phylogenetic analysis of 17 cassava *Ty3/gypsy*-like retrotransposons (*Megyp*s). The tree is based on 17 nucleotide sequences of *pol* gene fragments (*Megyp*s): 15 are partial sequences from the two ends of the ~1.6 kb gene fragments, while *Megyp*s 5 and 28 were full 1.6 kb length. This 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.

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

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
Ty3-2	*Saccharomyces cerevisiae*	1084606
Gypsy	*Drosophila melanogaster*	130583

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*.

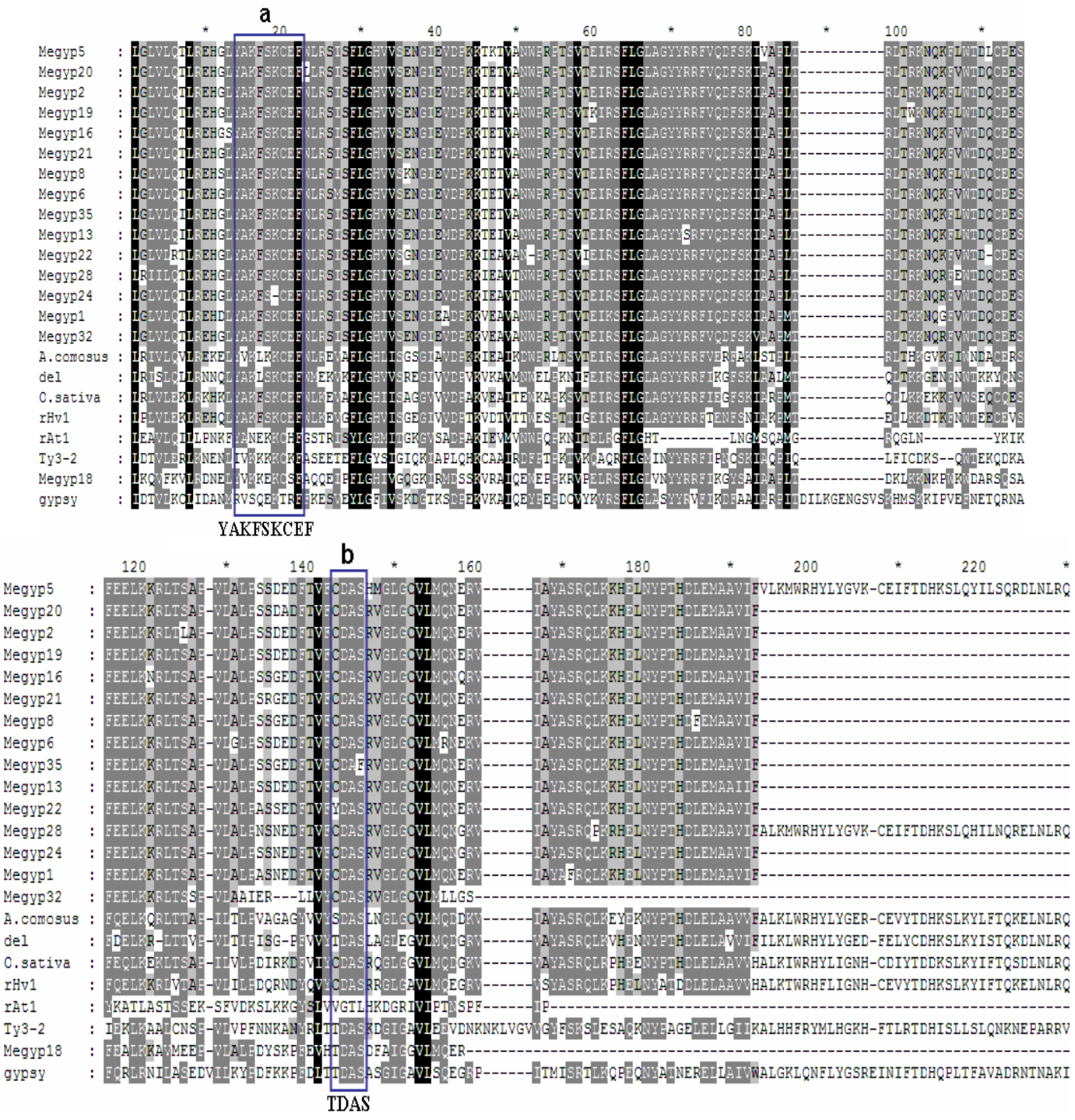

Figure 6. Alignment of the predicted amino acid sequences for 16 polyprotein fragments of cassava *Ty3/gypsy*-like retrotransposons with those of eight other plants. *Ty3-2* and *Gypsy* are included for reference (detail of the polyproteins in Table 4). Each of the clones was represented by either translation 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

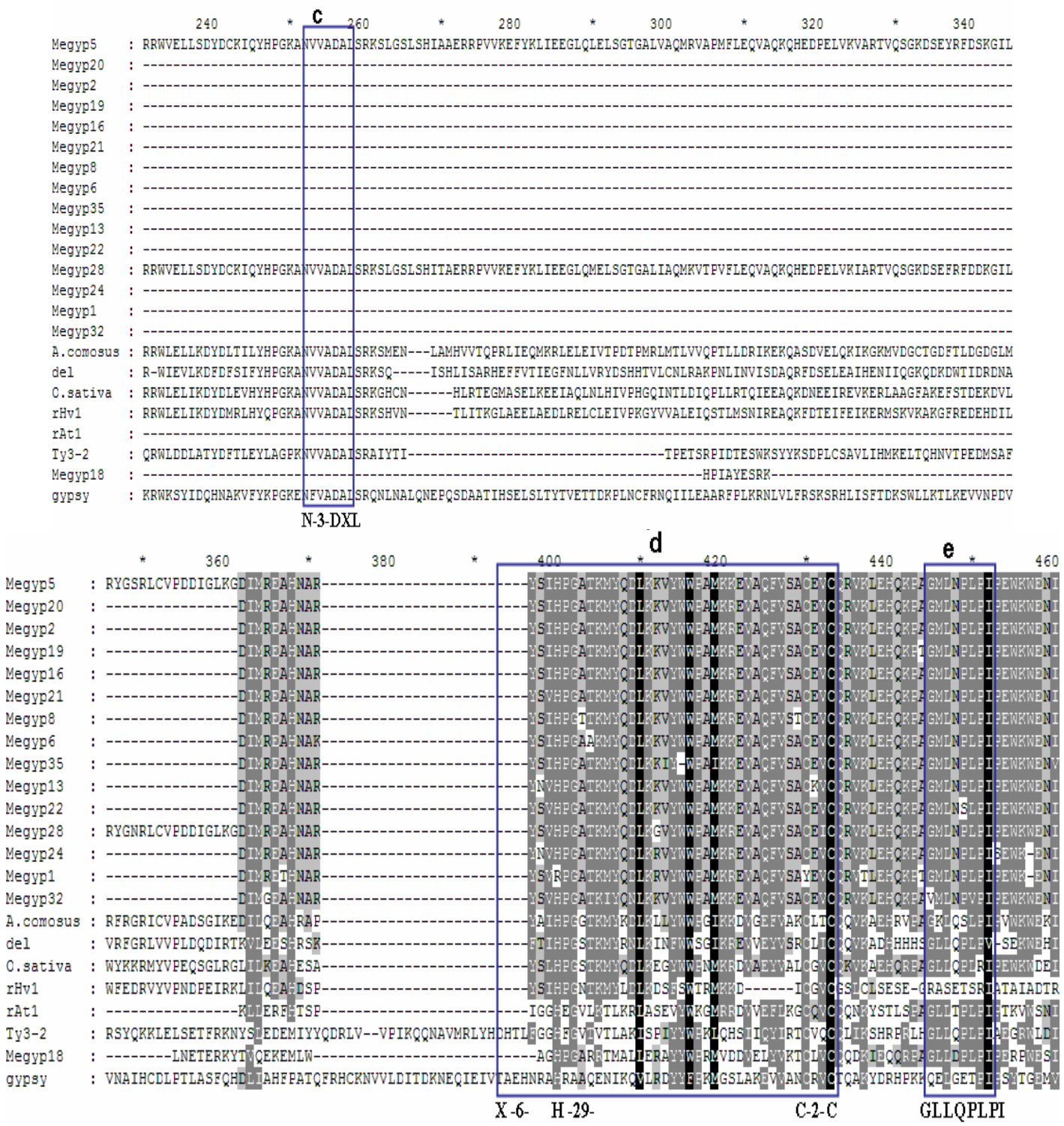

Figure 6. Contd.

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

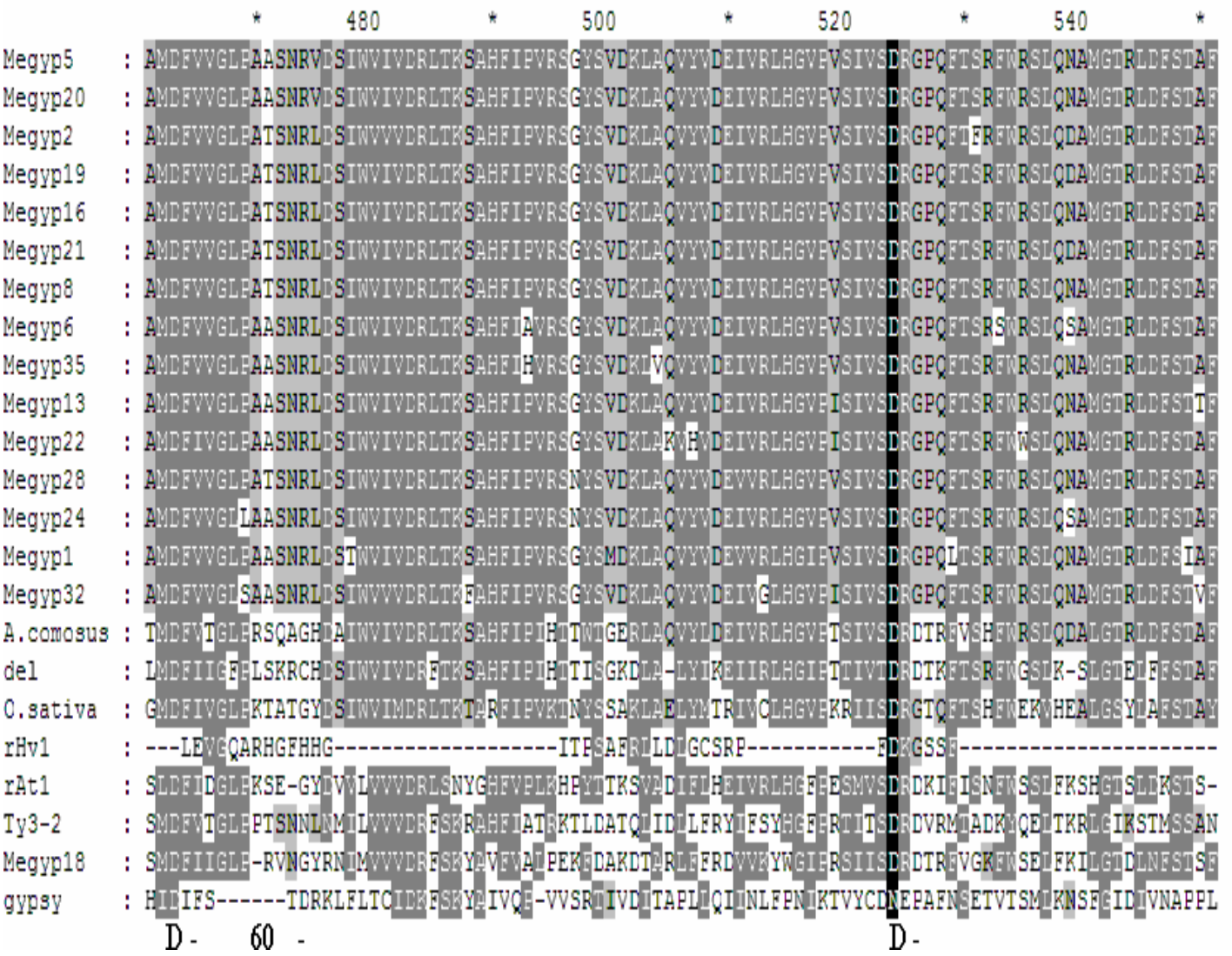

Figure 6. Contd.

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 domain-order 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-

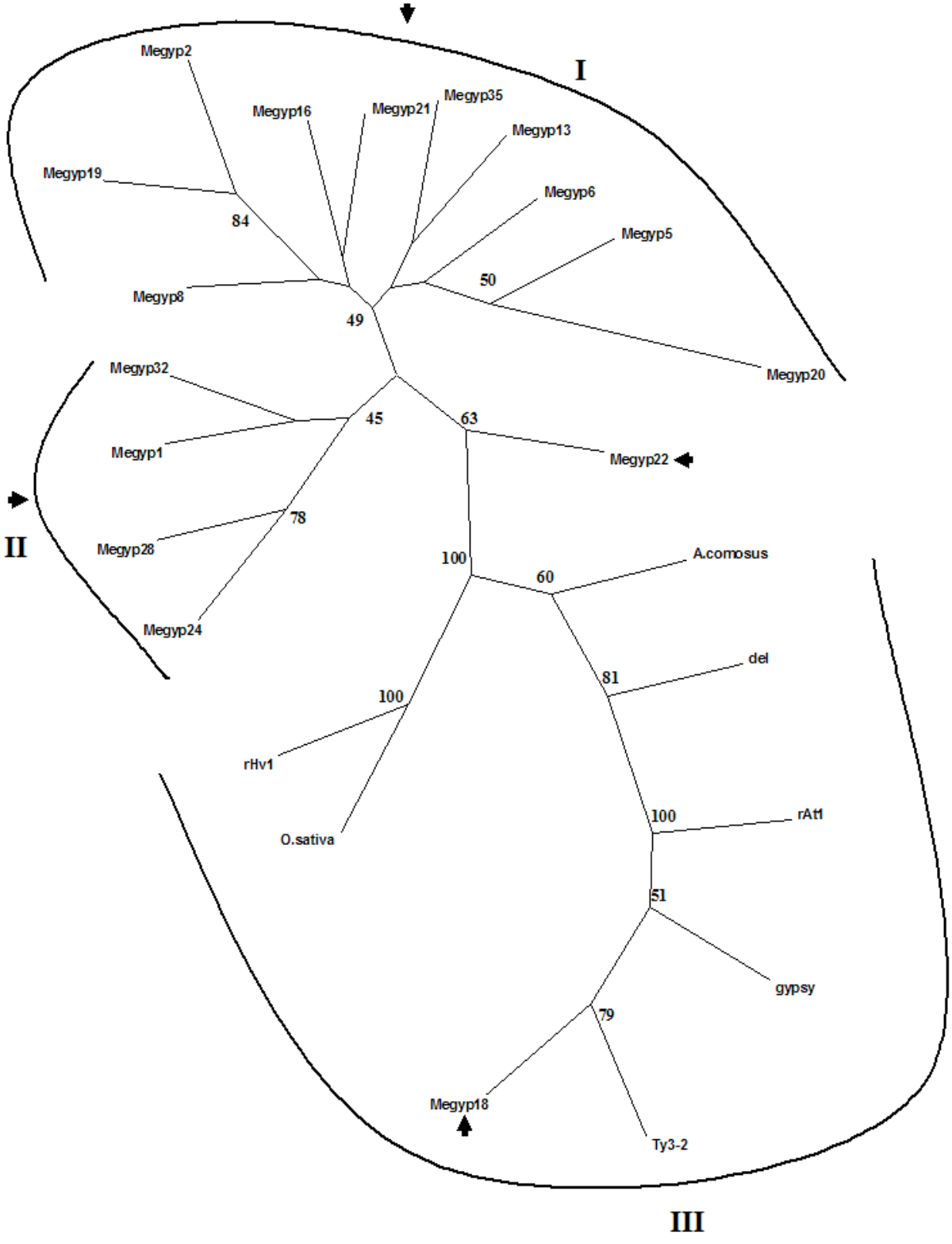

Figure 7. Comparative phylogenetic analysis of 16 cassava *Ty3/gypsy*-like retrotransposons (*Megyp*s) with other eight from other organisms. The tree is based on predicted amino acid sequences of *pol* gene fragments. This is a consensus neighbor-joining unrooted tree constructed with PHYLIP package. 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.

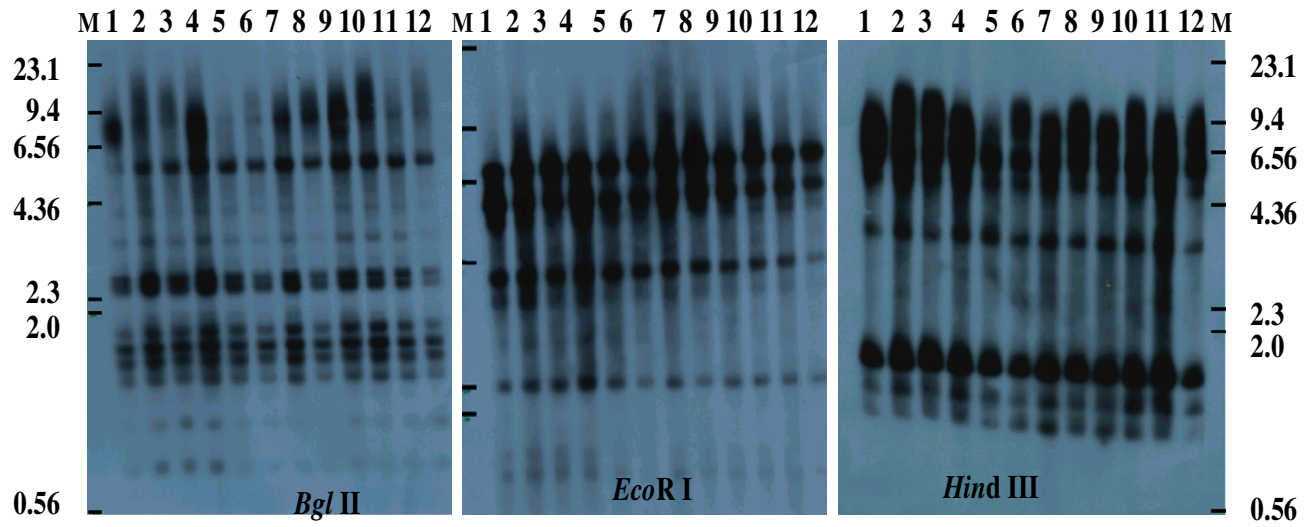

Figure 8. Southern blot analysis of *Ty3/gypsy*-like polyprotein of 12 cassava cultivars. 10 µg of genomic DNA from each of the cassava cultivars; lanes 1 (MGA1); 2 (MNGA2); 3 (MDOM5); 4 (MNGA19); 5 (MCOL22); 6 (CMC40); 7 (MVEN77); 8 (CG402); 9 (SM627); 10 (SM985); 11 (SM1088); 12 (CM2177) were digested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