Full Length Research Paper

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

Michael A. Gbadegesin¹* 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 (Finnegean, 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; Vershhiinin 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-
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, “Qiapure”), 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% agarose gels.
Ty3/gypsy showed clear homology to the polyprotein of end gave 28 clones with good sequences of which 26 like retrotransposons. However, sequencing from the 3’ Megy 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 retrotransposon. 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-50). 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 3a, 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 Megyps are more diverged than the 3’ (ENDONUCLEASE) ends and the Megyp5 and Megyp28 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 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 Megyp5 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
and Britten, 1993; Xiong and Eickbush, 1990). There is a block of residues previously identified as highly conserved in gypsy-like retrotransposons as shown in Figure 6. The alignment reveals four families using the CLUSTAL W programme (1.82) and colour shaded in 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 Megypsy (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; Kulikosky et al., 1992; Springer and Britten, 1993; Xiong and Eickbush, 1990). There is a 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 Megyp as homologous GMLNPLPI. Also present in the aligned Megyp 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; Kulikosky 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 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 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 Megyp 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, Megyp5, was used to probe Southern blots of restriction digests of genomic DNA from a range of

---

**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.
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 *Bgl* 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.
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.

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

<table>
<thead>
<tr>
<th>Locus or sequence name</th>
<th>Source species</th>
<th>Gi number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. comosus</td>
<td>Ananas comosus</td>
<td>2995405</td>
</tr>
<tr>
<td>O. sativa</td>
<td>Oryza sativa</td>
<td>37532428</td>
</tr>
<tr>
<td>Del</td>
<td>Lilium henryi</td>
<td>19442</td>
</tr>
<tr>
<td>rHv1</td>
<td>Hordeum vulgare</td>
<td>3413486</td>
</tr>
<tr>
<td>rAt1 right</td>
<td>Arabidopsis thaliana</td>
<td>3413430</td>
</tr>
<tr>
<td>rAt1 left</td>
<td>Arabidopsis thaliana</td>
<td>3413431</td>
</tr>
<tr>
<td>Ty3-2</td>
<td>Saccharomyces cerevisiae</td>
<td>1084606</td>
</tr>
<tr>
<td>Gypsy</td>
<td>Drosophila melanogaster</td>
<td>130583</td>
</tr>
</tbody>
</table>

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 *Hind* 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
were integrated in the genome. Many of the bands are very strong and distinct but there are few weak ones suggesting that \textit{Megyp5} 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 \textit{Megyp5} probe used (Figure 8).

**DISCUSSION**

The detection of \textit{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 \textit{Ty3/gypsy} sequences. For instance, of forty four sequenced clones following genomic DNA PCR amplification using primers based on the conserved RT domain in \textit{Brassica} sp, only twenty were similar to any of the known transposon types and just fifteen were \textit{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 primer sets has been confirmed here by the isolation and characterisation of the polyprotein fragment diagnostic of \textit{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 Megyp18 and Megyp22 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.
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) 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, 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. Megyp18 clustered closely with Ty3-2 retroelements (Figure 7) suggesting that, Megyp18 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 are
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


