THE STRUCTURES AND ABUNDANCE OF TRANSPOSABLE ELEMENTS CONTRIBUTING TO GENOME DIVERSITY IN THE DIPLOID AND POLYPLOID *BRASSICA* AND *MUSA* CROPS

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Faisal Nouroz
Department of Biology,
University of Leicester,
United Kingdom

March, 2012

Declaration

I hereby declare that no part of this thesis has been previously submitted to this or any other University as part of the requirements for a higher degree. The content of this thesis is result of my own experimentation and data analysis otherwise acknowledged in the text or by reference.

The work was conducted in the department of Biology, University of Leicester, during the period March, 2008 to March 2012.

Signed ……..…..

Faisal Nouroz.

Dedication

I wish to wholeheartedly dedicate this thesis to my parents Mr. Nouroz Khan and Mrs. Khursheed Nouroz and my lovely and caring wife Shumaila Noreen, who supported me in research work as well as at home at each step on my way.

Acknowledgements

It is a pleasure to thanks all those who made this work possible for me. First of all I am thankful to Almighty Allah, the most merciful who gave me the strength and courage to reach this stage and helped me at every step in my life.

My deep gratitude goes to my supervisor Prof. J.S (Pat) Heslop Harrison, whose encouragement, supervision and support from the preliminary to the concluding level enabled me to develop the understanding of my project. I am very thankful for his useful suggestions and caring at times when I needed them the most. I am also thankful to Dr. Trude Schwarzacher for her valuable assistance, moral support and help in learning the molecular techniques in the Lab.

My special gratitude goes to Dr. Richard Gornall for his valuable discussions and suggestions to improve and understand this work. My sincere thanks to members of biology dept Jean Liggins, Penny Butler, Lory Francescut and Jacqueline Frost for their assistance. I would also like to thanks Dr. John Bailey and my fellows in the lab, Niaz Ali for his help and support from the first day as we joined the lab together, Farah Badakshi and Mateus Mondin for their suggestions during my work and Nauf, Basma, Zuxin, Stuart, Chetan and other lab members for nice and friendly environment. My sincere gratitude goes to Gurpreet, Steven and Lidiya for their contribution in experimental work under my supervision.

I take the chance to thank all my family members. My parents whose prayers remained with me at every moment, as I was away from home for my studies. My wife Shumaila Noreen, whose love and support made my stay very easy in UK both at home and in University and her favour in my data analysis. I am also thankful to my sisters and brothers for their continuous support and understanding.

This thesis would not been possible without the funding provided by Hazara University, Mansehra and Higher Education Commission (HEC) of Pakistan for my stay in UK. I would like to thank the Staff of HRI, Warwick for providing the seeds for *Brassica* and Ashalatha (Asha) for providing banana DNA for analysis.

It is not possible to mention the names of all those people who contributed to this piece of work but I fully appreciate your valuable contributions. I therefore wish to sincerely thank all such people and may God bless them all.

Faisal Nouroz

The structures and abundance of transposable elements contributing to genome diversity in the diploid and polyploid *Brassica* and *Musa* crops

Faisal Nouroz

Abstract

Mobile DNA sequences - transposable elements (TEs) that amplify and move within genomes represent a high proportion of the DNA in most eukaryotes. The present study aimed to define TE nature, structure and abundance in two contrasting groups of diploid and polyploids crop genera, *Brassica* (dicotyledon) and *Musa* (monocotyledon). Rather than starting with known TE sequences, a sequence-data driven approach was used, comparing homologous and homoeologous BAC pairs. Over ~100 kb regions, any stretch of sequence was characterized that was inserted or deleted in the evolutionary time since divergence of the two BAC genomic sequences. Almost all the sequences were indeed TEs, representatives of existing and a few novel superfamilies. Polymorphisms due to activity were measured by PCR with flanking primers in 40 (*Brassica*) or 96 (*Musa*) accessions, and some families were localized on chromosomes by fluorescent *in situ* hybridization. Autonomous and non-autonomous TEs were found; class I retrotransposons like Copia and Gypsy (LTR) predominated in both genera, while SINEs and LINEs (Non-LTR) were abundant in *Brassica* genomes. Large retrotransposon derivatives (LARDs) were in both genera, with a very few terminal-repeat in miniature (TRIM) elements. Class II DNA transposons included CACTA, hAT, Harbinger, Mariner and Mutator like MITEs in *Brassica*, while CACTA and Mutator were uncommon in *Musa*. Among miniature inverted-repeat transposable elements (MITEs), Stowaway, Tourist, and Mutator-like MITEs were abundant with several novel families identified and characterized. In diploid and allopolyploid *Brassica* species, A- or C-genome specific elements were found while others were more active. PCR enabled accession identification and phylogenetic reconstruction in *Brassica* and *Musa*. As well as known element families, few novel types of TEs were identified, including several variable, short elements with characteristic structural features. The analysis provides insight into the nature and diversity of TEs as an important genomic component; results are useful for genome annotation and understanding evolution and variation within these crops and the associated pool of wild germplasm.

Abbreviations

aa	amino acids
AFLP	Amplified fragment length polymorphism
AP	Aspartic protease
ATP	A type transposase
BAC	Bacterial artificial chromosomes
Bo	*Brassica oleracea*
Bp	Base pair
Br	*Brassica rapa*
C	Current
C4	Calcutta 4
CDD	Conserved domain database
CHR	Chromatin organization modifier (chromodomain)
Cm	Centimeter
CMV	Cauliflower mosaic virus
CTAB	Cetyltrimethylammonim bromide
Cv	Cultivar
CYP	Cystein protease
DIRs	*Dictyostelium* intermediate repeat sequence
DNA	deoxyribonucleic acid
EBI	European bioinformatics institute
EMBL	European molecular biology laboratory
EN	Endonuclease
ENV	Envelop
ERV	Endogenous retrovirus sequences
EST	Expressed sequence tags
ETS	Extra transcription factor
FAO	Food and agriculture organization of the United nations
FITC	Fluorescein Isothiocyanate
g	gram
GAG	gag-nuclocapsid
GMGC	Global *Musa* genomics consortium
GSS	Genome survey sequence
GyDB	Gypsy database
HARB	Harbinger
Hel	Helicase
HP	Haemthiolate proteins
HVP	Tymovirus proteins
INT	Integrase

IPTG	isopropyl-β-thiogalactosid
IRAPs	Inter-retrotransposon amplification polymorphisms
Kb	Kilo base
l	Liter
LARDs	Large retrotransposons derivatives
LB	Luria-Bertani
LINEs	Long interspersed nuclear elements
LRR	Leucine rich repeat
LTR	long terminal repeat
M	Molar
Ma	*Musa acuminata*
Mb	*Musa balbisiana*
Mbp	Mega base pair
MFS	Major Facilitating Factor
Min	Minutes
MAP	MITE amplification polymorphism
MITE	Miniature inverted-repeat transposable elements
Ml	Milliliter
mM	milli molar
MT	Mannosyl transferase
Mu	Mutator
Mya	Millions years ago
NAM	No apical meristem-associated
NCBI	National centre for biotechnology information
ND5	NADH dehydrogenase subunit
NJ	Neighbour-joining
ºC	Degree celsius
ORFs	Open reading frames
PBS	Primer binding sites
PCR	Polymerase chain reaction
PIF	P instability factor
PKW	*Pisang Klutuk Wulung*
PLE	Penelope-like elements
PMV	Phage virion morphogenesis
POL B	B DNA polymerase
Pol	Polymerase
PPT	Polypurine tract
PRK	2´-Phosphodiesterase/3´-nucleotidase precursor protein
PRP	Pre-mRHA-splicing factor
PRV	Pararetrovirus
RAPD	Random Amplified Polymorphic DNA

RBIP	Retrotransposon-based insertion polymorphisms
REMAP	Retrotransposons-microsatellite amplified polymorphism
RFLP	Restriction fragment length polymorphism
Rep	Replication initiator
RNA	Ribonucleic acid
RP	Retropepsin
RPA	Replication protein A
RRM	RNA recognition motif
RT	Reverse transcriptase
RTAP	Reverse transcriptase amplification polymorphisms
Sec	Seconds
SINEs	Small interspersed nuclear elements
SLG	S locus glycoprotein
SNP	Single nucleotiode polymorphism
Spp.	Species
SSAP	Sequence-specific amplification polymorphism
SSR	Single sequence repear
TAE	Tris-Acetyl-EDTA buffer
TAP	Transposase amplification polymorphism
TEs	Transposable elements
TIP	Transposon insertional polymorphism
TIR	Terminal inverted repeat
TNP	Transposase
TPRT	Target primed reverse transcription
TR	Transcriptional regulator
TRIMs	Terminal-repeat retrotransposons in miniature
tRNA	Transfer RNA
TRX	Thioredoxin
TSD	Target site duplication
UD	Undetermined
UKP	Unknown proteins
UN	Unknown
UTR	Untranslated regions
UV	Ultra violet
V	volt
v/v	Volume per volume
w/v	Weight per volume
WGS	Whole genome shotgun
ZF	Zinc finger
ZK	Zinc knuckle
µl	Microliter

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 *Brassica*

The genus *Brassica* (family *Brassicaceae*) includes many important crops such as oilseed rape (canola), brown mustard, Chinese cabbage, turnip, cabbage, cauliflower, broccoli, Brussels sprouts, collards, kale and kohlrabi, and is a close relative of the model plant *Arabidopsis*. *Brassicas* are a valuable and long-standing food source in both developing and industrialized countries (Monteiro and Lunn, 1999). *Brassica* oil production is rapidly increasing worldwide and accounts for 12% of global vegetable oils (after only soybean and cotton), with *Brassica rapa*, *Brassica oleracea*, *Brassica juncea* and *Brassica carinata* all used for oil (FAOStat database, 2012). The six most economically important *Brassica* species include three diploids and three allotetraploids. The evolutionarily-recent polyploidy events led to the formation of three tetraploid species from the diploids *Brassica rapa* (2n=2x=20), *Brassica nigra* (2n=2x=16) and *Brassica oleracea* (2n=2x=18), forming the "Triangle of U" (U.N, 1935), where the three allotetraploid species *Brassica juncea* (2n=4x=36), *Brassica napus* (2n=4x=38) and *Brassica carinata* (2n=4x=34) represent hybrids of each pair of the three diploid species (Figure 1.1).

The haploid genomes of *Brassica rapa*, *Brassica nigra* and *Brassica oleracea* have been named A, B and C, respectively and the resulting amphidiploids become AB, AC and BC for *Brassica juncea*, *Brassica napus* and *Brassica carinata* respectively. Other non-domesticated *Brassica* taxa have been described, with at least ten related to *Brassica oleracea* (Ostergaard and King, 2008). The genome of *Brassica* has shown close relation to the model plant *Arabidopsis*, as they are the close genera of the same family. It is estimated that *Brassica* diverged from the *Arabidopsis thaliana* lineage between 20-24 million years ago (Mya) (Koch *et al.*, 2000), or based on nucleotide substitution rate as recently as 14.5-20 Mya (Yang *et al.*, 1999). While the common ancestor itself involves polyploidy, a 2.2 Mbp region from *Brassica oleracea* with a high homology to *Arabidopsis thaliana* has been used to show that gene loss occurred after the separation of *Brassica oleracea* and *Arabidopsis* so some polyploidy events occurred more recently (Town *et al.*, 2006). With the emerging demands of *Brassica* crops as vegetables and oil, the scientific community is thinking to develop latest methodologies for its improvement.

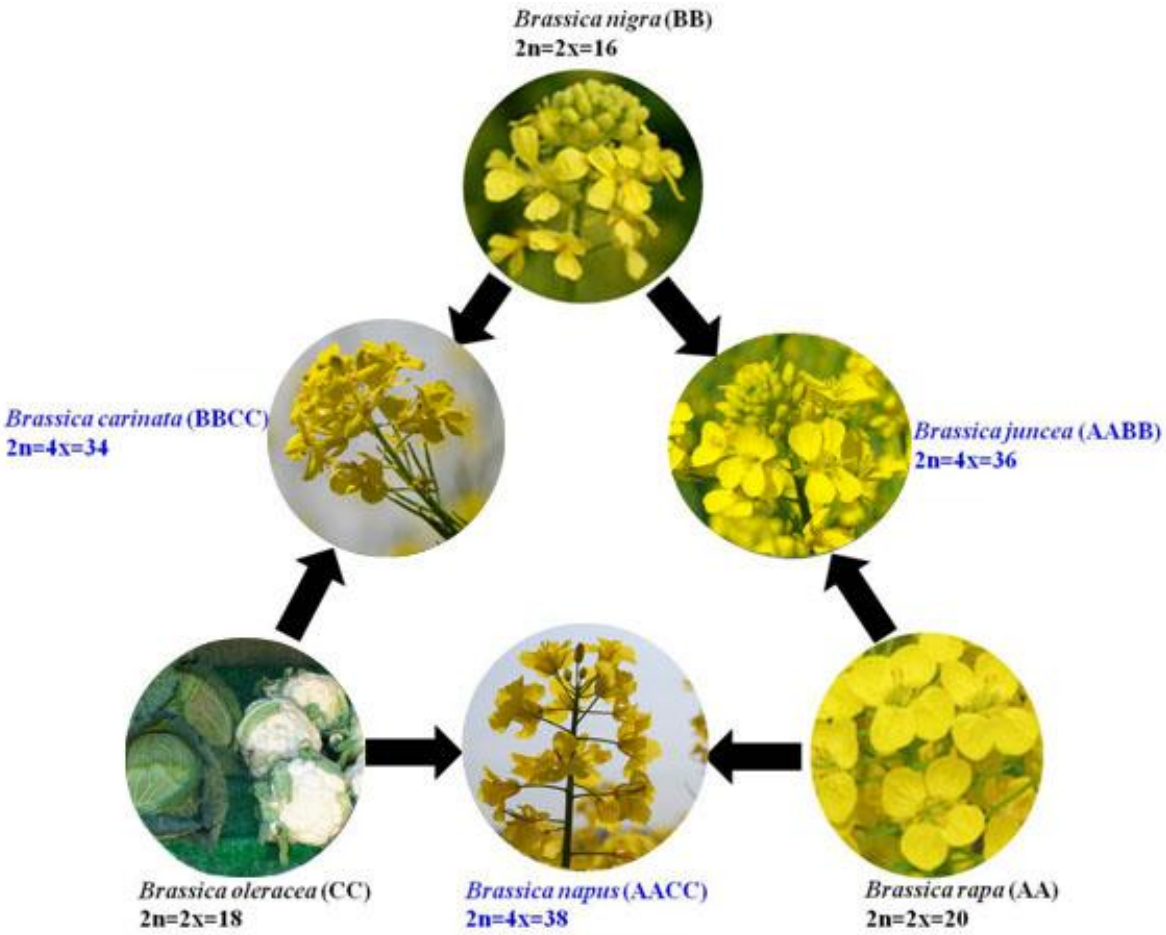


Figure 1.1: The genetic relationship and origin of allotetraploids from diploid *Brassica* species of the "Triangle of U". Diploid species are represented by black and allotetraploids (amphidiploids) species with blue colour. The genome and chromosome numbers are also indicated. (Molecular Cytogenetics Lab, University of Leicester photographs).

The Multinational *Brassica* Genome Project (MBGP) (*Brassica*.info; http://www.*Brassica*.info/index.php) was established in 2002 to gather the international research community working on *Brassica*. A number of genetic maps have been produced, along with SSR markers, EST and BAC end libraries, and the *Brassica rapa* sequencing project is in its final stages. Well established clone libraries, genetic maps, genetic markers, proteomics, sequencing data and many other databases are available providing significant information about *Brassica* genomic structures, biodiversity and polyploidy. A standard Gene Nomenclature was proposed by Ostergaard and King (2008) which will distinguish the alleles associated with the various genomes, and various paralogous loci. Several other projects are in progress and many are in pipeline to fully explore the genomic structure of several *Brassica* crops. The Korean *Brassica* genome project aims to sequence chromosome 1 of *Brassica rapa*; the sequenced chromosome 1 shows several conserved regions with *Arabidopsis* chromosomes, and as high as 82% homology was found by comparing five *Brassica rapa* sequenced BACs with *Arabidopsis* chromosome sequences (Yang *et al.*, 2005).

In 2011, the analysis and annotation of the draft genome sequence of *Brassica rapa* Chiifu-401-402 (Chinese cabbage) was published. Around 41174 protein coding genes were detected in *Brassica rapa* genome, which underwent genome triplication after splitting from *Arabidopsis thaliana* (itself used an out group to investigate the genome triplication). This *Brassica rapa* genome sequence is providing insight to the evolutionary mechanism of polyploids and genetic improvement of *Brassica* crops and its oil yield (Wang *et al.*, 2011). Single nucleotide polymorphism (SNP) linkage maps of the tetraploid *Brassica napus*, were constructed comprising 23037 markers, which were used to align the *Brassica napus* genome with its related species *Arabidopsis thaliana* and to genome sequence of its progenitor species, *Brassica rapa* and *Brassica oleracea*. Transcriptome SNP assays can be successfully used across the mapping populations to develop SNP-based linkage maps and transcriptome sequencing will increase the efficiency of predictive breeding even without fully sequenced genome (Bancroft *et al.*, 2011).

1.1.1 Polyploidy and phenotypic evolution in *Brassica*

Polyploidy or whole genome duplication (WGD) has played a major role in the evolution of higher plants, with about 70% of the angiosperms experiencing one or several events of

polyploidization during their evolutionary phases (Heslop-Harrison, 2010). This polyploidy is the result of doubling the chromosomes within a species (autopolyploidy) or by the combination of chromosomes sets from two different but related species (allopolyploidy). Both polyploid types are frequent and may be well adapted and genetically stable, and fertile where diploid chromosome behaviour is regained. However, synthetic polyploids or neo-allopolyploids may show genetic instability, low fertility and low embryonic viability (Osborn *et al.*, 2003; Comai, 2005; Chen, 2007; Ge *et al.*, 2009; Heslop-Harrison, 2012). The events of polyploidy played a major role in the tribe *Triticeae* and polyploidy events are well studied in these crops (Ma and Gustafson, 2008). The impact of ploidy events on gene expression and phenotypic characters have been carefully studied in *Brassica* allotetraploids including *Brassica napus* (Gaeta *et al.*, 2007). Like most angiosperms, there have been several rounds of polyploidy or whole genome duplication during evolution of *Brassica* species (Van de Peer *et al.*, 2009). Evolution of hexaploid *Brassica* progenitors advanced by several events, which occurred sequentially at different times (Ziolkowski *et al.*, 2006). Van de Peer *et al* (2009) have reviewed the evidence for three whole genome duplications occurring during the origin of the *Brassica* group (including *Brassica* and *Arabidopsis*): the gamma duplication early in evolution sometime after separation of angiosperms from gymnosperms, while the beta and alpha duplication are more recent.

Polyploidy contributes to phenotypic variation and the ability for genome mutation through several mechanisms: new phenotypes with new gene control, transpositions, altered regulatory interactions or changes that affect the expression of gene and cause phenotypic variation. Continued research on *Brassica*, cereal crops and related species should provide insight into the relative importance of these evolutionary mechanisms for generating novel variation in polyploids (Osborn *et al.*, 2003; Heslop-Harrison and Schwarzacher, 2011; Heslop-Harrison, 2012). The allopolyploid *Brassica* species can be resynthesized by crossing diploid species and doubling the chromosomes of the hybrids. Resynthesized *Brassica* polyploids are ideal for studying polyploid evolution in action, because the exact diploid progenitors are known and completely homozygous polyploids can be created by chromosome doubling of amphihaploids. Studies of resynthesized *Brassica* polyploids provided the first molecular evidence of novel phenotypic variation in newly formed polyploids (Schranz and Osborn, 2000; Barker *et al.*, 2009).

1.2 *Musa* (Banana)

Banana and plantains are the fourth most important tropical crop of the world; there is no botanical distinction, but often 'banana' describes the dessert fruit and 'plantains' the form cooked as a vegetable. They are upto 3 meter tall herbaceous monocotyledonous plants from genus *Musa*, family *Musaceae* and order Zingiberales (Tomlinson, 1969). There are more than 1000 banana cultivars with a high genomic diversity and variability; most of the cultivated species are triploids, while diploids and tetraploids are also common (Heslop-Harrison and Schwarzacher, 2007). The genus *Musa* is classified into four sections on the basis of morphology and chromosomes numbers as Eumusa (n=11), Rhodochlamys (n=11), Callimusa (n=9/10) and Australimusa (n=10).

Most edible bananas belong to section Eumusa and are diploid or triploid hybrids from *Musa acuminata* (A-genome) alone, or in combination with B-genome diploid banana *Musa balbisiana* (Perrier *et al.*, 2011); cultivars have multiple origins from cultivated and wild cultivars by hybridisation (Hippolyte *et al.*, 2012). Cultivated bananas are mostly sterile and parthenocarpic (fruit formation without seeds and fertilization) and are mostly triploid (2n=3x=33) having the genome constitution of AAA, AAB, ABB. Most cooking types are interspecific hybrids (AAB or ABB) while sweet dessert bananas are mostly triploid *Musa acuminata* i.e. AAA (Pollefeys *et al.*, 2004; Heslop-Harrison, 2011). This triploidization occurred independently in various areas between the diploids and from various parental combinations, and new triploids and tetraploids can be synthesized in breeding and selection programmes. Three well established subgroups from triploids are African AAA 'Mutika Lu-jugira', AAB 'African Plantains' and AAB 'Pacific Plantains'. Other seedless cultivated bananas include diploids with a genome set of AA and AB and tetraploids with a genome constitution of AAAA, AAAB, AABB and ABBB (Heslop-Harrison and Schwarzacher, 2007; Perrier *et al.*, 2011).

The Global *Musa* Genomics Consortium (GMGC) is an international network of scientists, working for the improvement in breeding and management of banana by genomic techniques. Comparative genomics is a valuable tool to investigate genomic evolution and *Musa* is considered as an interesting model crop for understanding genomic evolution. The consortium is working to improve and explore the genomic structure of *Musa* including *Musa* Genome Resource Centre (MGRC), Genotyping Centre and

Bioinformatics databases and a germplasm resource centre. The research is in progress in areas including banana bioinformatics, genetic diversity, mapping, sequencing, expression, validation, proteomics and pathogen genomics (GMGC website; http://www.*Musa*genomics.org/home_page.html). In recent years bacterial artificial chromosomes (BAC), expressed sequence tags (EST) and several other DNA marker sequences have been generated within the GMGC framework: in total 51 BACs from various *Musa* accessions were recently sequenced and annotated for genes and TEs with automated approaches (Guignon *et al.*, 2012). The BAC end sequences were searched against several databases and high homology was observed against various repetitive sequences and transposons (30%). Approximately 600 BAC end-sequences contained protein sequences that were not found in the existing available *Musa* ESTs, repeat or transposon databases. These results suggested that these BAC end sequences from *Musa acuminata* had shown significant homology to *Oryza sativa* and *Arabidopsis* genome sequences (Cheung and Town, 2007).

After a great development in *Musa* BAC sequencing, there is huge progress in whole genome sequencing of *Musa* and investigating the repetitive part of the genome. The repetitive part of banana was investigated by sequencing the *Musa acuminata* 'Calcutta4' genome. The major components from transposable elements were the LTR retrotransposons, from which more than 16% of the genome is composed of Copia, while 7% is represented by Gypsy elements. The phylogenetic study of the elements indicate that the majority of the Copia elements belong to the *SIRE/maximus* lineage (13%), while the remainder belongs to *Angela*, *Tnt1* and *Hopscotch* lineages. Most of the Gypsy elements (87%) belonged to the lineage of chromoviruses. The non-LTR retrotransposon component of banana was very rare with only 1% genome composition covered by LINE elements. No SINEs were found actively proliferating in the genome, while the DNA transposons were found to be very rare, covering less than 1% of total 'Calcutta 4' genome. Very few hAT-like elements were identified due to their low copy number. In addition to retrotransposons and DNA transposons, two satellite repeats were also identified, which were considered as best cytogenetic markers. A specific *Musa* repeat database was created to assist the banana sequence annotation (Hribova *et al.*, 2010).

1.3 Transposable elements (TEs)

Transposable elements (TEs) also termed as mobile genetic elements, jumping genes, or transposons are a major component of all eukaryotic genomes, representing 40% of the entire genome in humans (Mills *et al.*, 2006) and 50-90% in important agricultural crops like maize, wheat, barley, rye and sugar beet (Pearce *et al.*, 1997; Kubis *et al.*, 1998; Wicker and Keller, 2007; Wicker *et al.*, 2007; Kapitonov and Jurka, 2008). The larger genomes are made up of abundant tandemly repetitive sequences and transposable elements, which compose a major proportion of DNA, sometimes representing more than half of the DNA (Heslop-Harrison and Schwarzacher, 2011). The first transposable element to be described, the Ac element of maize, was discovered by Barbara McClintock in the early 1950s (McClintock, 1950). With the advancement in computer-assisted analyses and genome-sequencing projects, it is now known that TEs are important components of all eukaryotic genomes and play a major role in their evolution (Flavell *et al.*, 1997; Wicker *et al.*, 2007).

1.4 Outline of Transposable Elements (TEs) classification system

The first classification system of transposable elements (TEs) was proposed by David Finnegan in 1989. His classification system distinguished the TEs in two major classes on the basis of their transposition intermediate. Class I or retrotransposons transpose via an RNA intermediate and Class II or DNA transposons transpose directly from DNA with a transposase. By analogy with computer word-processing programs, the Class I transposition mechanism is called “copy and paste” and that of Class II, “cut and paste” (Finnegan, 1989). Another detailed classification and nomenclature based on the Finnegan (1989) classification was proposed by Capy *et al*. (2005). In general, all the eukaryotic TEs were classified into two major types by many authors; retrotransposons and DNA transposons based on their copy and paste and cut and paste transposition mechanism respectively (Jurka *et al*., 2007; Kaptinov and Jurka, 2008), although this system must be adjusted for miniature inverted-repeat elements (MITEs), which copy and paste without an RNA intermediate or a transposase in the genome. The grouping of TEs in two classes is still the most accepted system of classification used by many authors (Hansen and Heslop-Harrison, 2004; Wicker *et al.*, 2007).

The TE classification system proposed by Wicker *et al*. (2007) includes (in hierarchical order) the levels of class, subclass, order, superfamily, family and subfamily. This classification is adopted and followed in the present work. The highest level of the hierarchy is the class, which divides TEs into two classes by the presence or absence of an RNA transposition intermediate. Subclass is used to distinguish elements that copy themselves for insertion from those that leave the donor site to reintegrate elsewhere. The Order in this system is used for Class I elements or retrotransposons, which are divided into 5 orders as LTR, DIRS, PLE, LINEs, SINEs. The order LTR is composed of five superfamilies as Copia, Gypsy, Bel-Pao, Retrovirus and ERV. The order LINE consists of R2, RTE, Jockey, L1 and I type superfamilies, while SINEs constitute 3 superfamilies. The Class II or DNA transposons are divided into two sub-classes. Sub-class 1 is composed of common superfamilies as Tc1-Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF-Harbinger, CACTA and Crypton, while sub-class 2 includes the superfamilies of Helitron and Meverick elements (Figure 1.2) (Wicker *et al.*, 2007).

A second TE classification system, differing at some levels, was proposed by Kapitonov and Jurka (2008), where TEs are classified into two major types (retrotransposons and DNA transposons) and five major classes as LTR retrotransposons, non-LTR retrotransposons, cut-and-paste DNA transposons, rolling-circle DNA transposons (Helitrons) and self-synthesizing DNA transposons (Polintons). These classes are further divided into superfamilies, which in turn are composed of numerous families of TEs. Thus the class LTR retrotransposons is composed of the Gypsy, Copia, BEL and DIRS superfamilies including the ERV1, ERV2 and ERV3 superfamilies. The non-LTR retrotransposons class includes several superfamilies as CR1, CRE, I, Jockey, L1, NeSL, Penelope, R2, R4, RandI, Rex1, RTE, Tx1 (LINEs) and SINE1, SINE2, SINE3 (SINEs) superfamilies. The DNA transposons consist of 15 superfamilies as Chapaev, CACTA, hAT, Harbinger, ISL2EU, Kolobok, Mariner, Merlin, MuDR, Mutator, Novosib, P, PiggyBac, Mirage, Rehavkus, Tourist and Stowaway. The fourth and fifth class is composed of Helitron and Polintons superfamilies respectively, which are not common and well proliferated like other TEs (Kapitonov and Jurka, 2008).

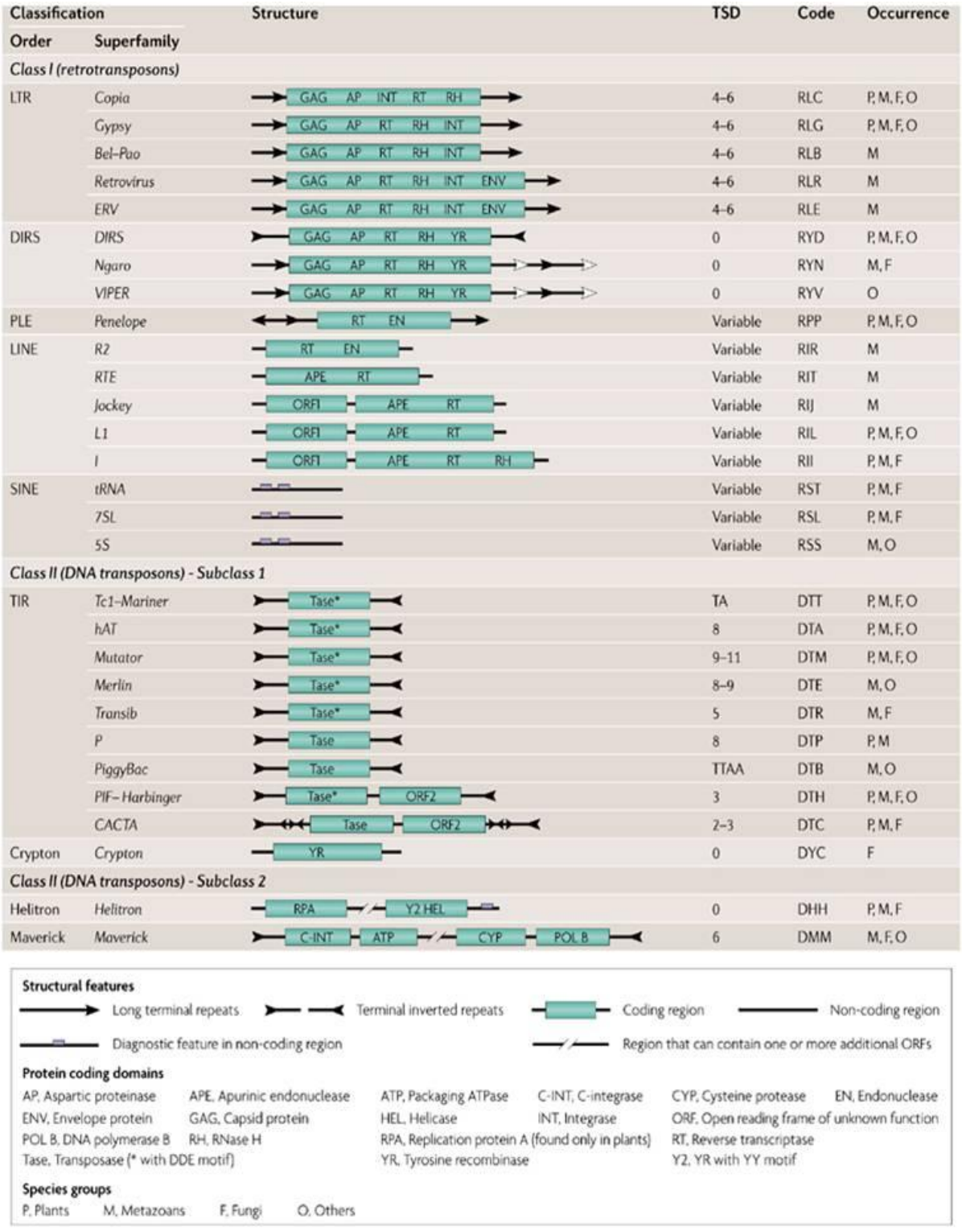


Figure 1.2: Classification system for transposable elements from Wicker *et al*., (2007). This scheme was adopted in the present work. The system includes (in hierarchical order) the levels of class, subclass, order, superfamily, family and subfamily. The transposable elements are divided into two major classes (retrotransposons and DNA transposons) by the presence or absence of an RNA transposition intermediate. The Copia and Gypsy are the only LTR retrotransposons present in plants, while a few LINEs and many SINEs are present in plant genomes. Among DNA transposons, Tc-1 Mariner, hAT, Mutator, PIF-Harbinger and CACTA are frequent in plant genomes. Origional image taken from Wicker *et al*., (2007).

1.5 Autonomous and non-autonomous transposable elements

Retrotransposons and DNA transposons are further classified as autonomous and non-autonomous based on presence or absence of the genes encoding the enzymatic machinery required for their transposition. DNA transposons possess a transposase, which assists in transposition and integration of the element. In autonomous retrotransposons, *gag-pol* genes are organized to encode the proteins necessary for retroelement transposition. The *gag* gene catalyses the production of virus like particles, while *pol* gene encodes polyproteins (RT, RH, INT), where RT and INT catalyse the transposition and integration of newly synthesized copy into a new site. In contrast the non-autonomous elements lack the necessary encoded proteins in their internal regions. The non-autonomous elements transpose by borrowing the enzymatic machinery of autonomous relative belonging to the same superfamily. Almost all the superfamilies from retrotransposons and DNA transposons have autonomous and non-autonomous elements which may differ by as little as a single frameshift or base pair change, or may have major deletions (Feschotte *et al.*, 2002; Jurka *et al.*, 2007; Vukich *et al.*, 2009).

1.6 Class I transposable elements

In maize 70% of the nuclear DNA is contributed by retrotransposons (SanMiguel and Bennetzen, 1998), a typical result for many species. The genomic and extra-chromosomal copies proliferate by an RNA intermediate copied into DNA by reverse transcriptase (Feschotte *et al.*, 2002; Kapitonov and Jurka, 2008; Kapitonov *et al.*, 2009). Retroelements have been categorized on the basis of phylogeny of their reverse transcriptase (RT), *gag-pol* domain organization, proliferating devices and structural features into long terminal repeat retrotransposons (LTRs), long interspersed nuclear elements (LINEs), short interspersed elements (SINEs), DIRs-like elements and Penelope-like elements. The LTR retrotransposons are further classified into superfamilies as Ty1/copia, Ty3/gypsy, Bel-Pao, Retrovirales and ERV-like elements (Figure 1.2). Copia and Gypsy elements are the most abundant and diverse group of retrotransposons studied in several organisms (Wicker *et al*; 2007).

1.6.1 Structural characteristics of long terminal repeat (LTR) retrotransposons

LTR retrotransposons are bounded by long terminal repeats (LTRs) ranging in sizes from a few hundred base pairs to 5 kb. They are predominant in plants and their size ranges from a few kb to 25 kb. Upon insertion, they generate target site duplications (TSDs) of 4-6 bp and their flanking LTRs usually have the conserved termini as 5′-TG-3′ and 5′-CA-3′. They contain open reading frames (ORFs) for *gag-pol* genes and sometime have one or more additional ORFs of unknown function. The *gag* gene is a structural protein for virus like elements, which encodes proteins aiding in maturation and packaging of retroelement RNA and proteins required for genome integration. The *pol* gene is a long ORF having polyprotein domains like aspartic protease (AP), reverse transcriptase (RT), RNase H (RH) and integrase (INT). RT is the most conserved domain present in all the retrotransposons and helps in the transposition mechanism of the element with the association of RH. The LTR retrotransposons exhibit the primer binding site (PBS) downstream to 5′ LTR and a polypurine tract (PPT) towards the upstream of 3′ LTR. Some other specific signals are also present in LTR retrotransposons for packaging, dimerization, reverse transcription and integration of the elements into new sites. LTR retrotransposons are further divided into orders and superfamilies. Five well known families are characterized from LTR retrotransposons (Kumar and Bennetzen, 1999; Wicker *et al.*, 2007; Vukich *et al.*, 2009). The LTR retrotransposons are described below.

1.6.1.1 Ty1/Copia

Ty1/copia elements are abundant elements present in most organisms including plants. They differ from Ty3/gypsy elements in the order of protein domains encoded by *pol* gene. In Copia the INT is towards the upstream of RT, whereas in Gypsy INT is downstream to the RT and RH domains (Figure 1.2). They range in size from few kb to many kb, flanked by LTRs and displayed the PBS and PPT motifs. The canonical Ty1/copia element is flanked by LTRs, displays PBS and PPT towards downstream and upstream of 5′ LTR and 3′ LTR respectively, and has internal *gag-pol* genes, which encode the proteins as 5′-GAG-INT-RT-RH-3′. Some elements also encode some additional protein domains of known or unknown nature in their *pol* gene and their copy numbers vary in various host species (Flavell, 1992; Flavell *et al.*, 1997; Wicker *et al.*, 2007).

The Copia elements are a diverse and heterogeneous group of LTR retrotransposons present in almost all eukaryotic genomes including the model fly *Drosophila* (Flavell, 1984), amphibians, reptiles (Flavell and Smith, 1992; Flavell *et al.*, 1995) and in higher plant genomes (Flavell *et al.*, 1992a). The chromosomal localization of Copia elements have been studied by fluorescent *in situ* hybridization (FISH) in *Beta vulgaris* showing their massive distribution on *Beta* chromosomes (Schmidt *et al.*, 1995) and *Arabidopsis* (Heslop-Harrison *et al.*, 1997). More recently the Copia retrotransposons have been studied in several plants including wheat, barley, rice and *Arabidopsis* (Wicker and Keller, 2007; Tsukahara *et al.*, 2009; Tomita *et al.*, 2010), sugarcane (Muthukumar and Bennetzen, 2004), cotton (Hawkins *et al.*, 2008), jute (Ahmed *et al.*, 2011), grapevine (Moisy *et al.*, 2008), melon (Ramallo *et al.*, 2008), tomato (Tam *et al.*, 2007; Cheng *et al.*, 2009), cassava (Gbadegesin *et al.*, 2008), tomato (Karlov *et al.*, 2010), sunflower (Vukich *et al.*, 2009; Kawakami *et al.*, 2010), pea (Macas *et al.*, 2007), sweet potato (Okpul *et al.*, 2011), *Medicago truncatula* (Wang and Liu, 2008) and *Arabidopsis* (Tsukahara *et al.*, 2009). The characterization of Copia from higher plants revealed their abundance and distribution among plant genomes (Flavell *et al.*, 1992b).

1.6.1.2 Ty3/Gypsy

Ty3/gypsy elements constitute a superfamily of LTR retrotransposons, which are widely distributed among fungi, animals and plants. They are characterized by generating 4-6 bp TSDs, LTRs ranging from few hundred bp to few kb, internal regions displaying *gag-pol* genes encoding protein domains and a PBS and PPT towards downstream and upstream of 5´ LTR and 3´ LTR respectively. On the basis of structural features, they resemble Copia except in the order of INT in the *pol* gene. In Gypsy elements INT is present downstream to RT and RH, while in Copia; it is upstream to RT and RH domains (Figure 1.2). This significant difference separates the two major groups of LTR retrotransposons. On the basis of presence or absence of another protein domain chromatin modifier organizer (Chromodomain; CHR), they are further divided into chromodomain-bearing Gypsy and non-chromodomain Gypsy elements (Kumar and Bennetzen, 1999; Malik *et al.*, 1999).

The Gypsy elements are actively proliferating in plant genomes and have shown diversity and abundance in several plants like wheat (Tomita *et al.*, 2010; Salina *et al.*, 2011), sorghum (Muthukumar and Bennetzen, 2004), pinus (Rocheta *et al.*, 2007), jute (Ahmed

et al., 2011), citrus (Bernet and Asins, 2003), soybean (Du *et al.*, 2010), pepper and tomato (Park *et al.*, 2011), tomato (Peters *et al.*, 2009), sweet potato (Okpul *et al.*, 2011), chickpea (Rajput and Upadhyaya, 2009), sunflower (Staton *et al.*, 2009; Ungerer *et al.*, 2009) and *Arabidopsis* (Tsukahara *et al.*, 2009). This suggests the diversity, abundance and distribution of Gypsy elements and their impact on plant genome duplication and diversification.

1.6.1.3 Retroviruses and related elements

Retroviruses are similar to LTR retrotransposons from the evolutionary point of view. They are considered to be evolved from the Ty3/gypsy elements that adopted a viral lifestyle by gaining an envelope protein (ENV) and some other regulatory proteins. As a parasitic mode of reproduction, they are mostly present in vertebrates. They are classified as the superfamily of LTR retrotransposons; otherwise they were classified as viruses for a long time. Retroviruses can be transferred into LTR retrotransposons by losing or deleting their extra domains (Capy, 2005). Retroviruses are more similar to Ty3/gypsy elements as both exhibits the similar *pol* domain organization, while *gag* gene differs in retroviruses by encoding matrix functions and an extra capsid, which is not present in Gypsy-*gag*. The main characteristics of viruses is the retaining of envelop (ENV) domian, that has both surface and transmembrane units (Wicker *et al*., 2007).

1.6.1.4 Large retrotransposon derivatives (LARDs)

LARDs are non-autonomous LTR retrotransposons, which do not encode the *gag* or *pol* gene proteins necessary for transposition. They were discovered in maize, where large non-autonomous elements (5.5-8.5 kb) were found flanked by large LTRs, 4-6 bp TSDs but display non-coding internal regions, due to which the elements were named ‘large retrotransposons derivatives’ (LARDs). The non-autonomous Dasheng and Zeon-1 elements from maize genome are each represented by around 1000 copies (Hu *et al.*, 1995; Jiang *et al.*, 2002). They are mobilized in *trans* by using the proteins from the autonomous elements residing nearby. The internal region of these elements contains a long conserved non-coding DNA segment that may provide the important secondary structure to the mRNA, although it is unclear how these non-coding sequence works in the life cycle of the elements (Havecker *et al.*, 2004). The well known Sukkula elements are also LARDs

elements having a large internal non-coding region, flanked by LTRs. All these elements are considered as non-autonomous due to the lack of *gag-pol* protein domains. They exhibit the PBS and PPT, which serves as minus and plus strand of the primary sites for reverse transcription of the elements as observed in Copia and Gypsy elements (Kalendar *et al.,* 2004). The LARDs-like elements were identified in many plants and have more or less similar features. The LARDs identified in barley and members of the *Triticale* have well characterized LTRs of 4.5 kb and an internal region of 3.5 kb (Kalendar *et al*., 2004).

1.6.1.5 Terminal-repeat retrotransposons in miniature (TRIM)

TRIM are considered as non-autonomous LTR retrotransposons due to structural similarities with them. TRIM are small in size, generate 5 bp TSDs, flanked by 100-250 bp LTRs and display a PBS and PPT downstream and upstream of 5´ LTR and 3´ LTR respectively (Witte *et al.*, 2001; Antonius-Klemola *et al.*, 2006). TRIM have been studied in many monocot and dicot plant families including *Poaceae, Brassicaceae, Solanaceae*, and *Fabaceae*. The highest TRIM from monocots were studied in *Oryza sativa*, while from dicots, they were investigated in *Arabidopsis thaliana*. A total of 43 TRIM-like elements were identified from *Arabidopsis thaliana* (Witte *et al*., 2001). By the pairwise comparison of homoeologous BAC sequences, TRIM elements were detected in *Brassica rapa*. The elements were named *Br1, Br2, Br3* and *Br4*, which are 364, 385, 536-654 and 1311 bp in sizes and flanked by LTRs of 119, 125, 210 and 255 bp respectively. The *Br3* element is largest in size with 807 bp non-coding region inner to flanking LTRs. The copy numbers of TRIM in *Brassica rapa* and *Brassica oleracea* were also estimated as 530 and 660 copies respectively (Yang *et al.*, 2007). The structure of *Malus* (apple) TRIM was also similar to TRIM described in other plants. The apple TRIM was flanked by about 306 bp LTRs with a short internal non-coding domain. The terminal ends of both LTRs have a 10 bp terminal inverted repeat. Downstream to 5´ LTR of these TRIM is the PBS complementary to $tRNA_{Met,}$ while a PPT is present upstream to the 3´ LTR (Antonius-Klemola *et al*., 2006).

1.7 Non-LTR retrotransposons

Non-LTR retrotransposons or retroposons are characterized by having short LTR ranging from 1-50 bp. They are further divided into two groups on the basis of their sizes and presence or absence of their internal region encoding the domains. The larger elements encoding the RT, RH or an additional domain are called long interspersed nuclear elements (LINEs), while the small elements without coding regions are called small interspersed nuclear elements (SINEs). Both elements are present in plants (Kubis *et al.*, 1998) but LINEs are more abundant in mammals including the humans (Jurka *et al.*, 2007).

1.7.1 Long interspersed nuclear elements (LINEs)

LINEs lack LTRs and range in length from a kb to several kilobases. Upon insertion to a new site, they generate TSDs, one or two open reading frames (ORFs) and an internal RNA polymerase II promoter in its 5´ terminal region, which facilitate the retrotransposons in its transcription. The ORFs of the elements sometimes overlap by the frameshifts and the untranslated regions (UTRs) can be flanked at both ends of the coding regions (Xiong and Eickbush, 1990; Jurka *et al.,* 2007). Autonomous LINEs encode at least an RT and a nuclease in their *pol* ORF for their transposition (Figure 1.2). The most conserved domain present in all the autonomous LINEs is RT, which shows the amino acid conservation in seven domains, characteristic of retroviral RNA-directed DNA polymerases. However, lack of the LTRs and the presence of a poly(A) fragment distinguish the LINEs from the retroviruses. In may LINEs, a zinc finger (ZF) domain is present in *gag* or *pol* ORFs. The mechanism of mobilization and integration of the LINEs was studied in detail in a coupled process called target primed reverse transcription (TPRT) (Xiong and Eickbush, 1990; Malik *et al.*, 1999; Jurka *et al.*, 2007).

Few LINE elements have been investigated and characterized in plants till now, although the number of reported LINEs is growing with the advancement in genome sequences (Kubis *et al.*, 1998; Noma *et al.*, 1999). The first well characterized plant LINE was Cin4 detected in the A1 gene of *Zea mays* (maize), as an insertion in the 3´ untranslated region. Later on, many more copies identical to Cin4 were detected with variable sizes and variability in their 5´ end. The chromosomal localization of LINEs in *Beta* species and

five gymnosperms revealed their abundance and distribution (Kubis *et al.*, 1998). The well characterized LINEs from plants are *Karma* from *Oryza sativa* (Komatsu *et al.*, 2003), *Llb* described in *Ipomoea batatas* genome (Yamashita and Tahara, 2006), *BLIN* from *Hordeum vulgare* (Vershinin *et al.*, 2002), *del2* from *Lilium speciosum* (Leeton and Smyth, 1993) and *ATLN* from *Arabidopsis thaliana* (Noma *et al.*, 2001). A LINE family named *BNR* was described from the genome of *Beta vulgaris* having 3 well characterized elements (*BNR1-BNR3*). The elements range in size from 6.4-9.3 kb, flanked by 7-22 bp TSDs and exhibiting two non-overlapping ORFs. *BNR1* in 6.7 kb, *BNR2* is 6.4, and *BNR3* is 9.3 kb in size. BNR elements harbour an extra domain in ORF1 named as RNA recognition motif (RRM) (Heitkam and Schmidt, 2009).

1.7.2 Small interspersed nuclear elements (SINEs)

The small interspersed nuclear elements (SINEs) are small Non-LTR retrotransposons ranging in size from 100-500 bp having internal promoters for RNA polymerase III (Okada *et al.*, 1997; Kapitonov and Jurka, 2003). They have a complex structure, with a 5´ region similar to tRNA genes, or 7SL RNA genes, an internal polymerase III promoter, non-tRNA region of variable sizes, or 3´ region with similarity to the terminal regions of LINEs, a short segment of A or T at their 3´ terminal end and presence of flanking direct repeats. The SINEs are non-autonomous elements, as they lack their own reverse transcriptase protein necessary for transposition. Despite their non-autonomous nature they are mobile elements and utilize the enzymatic machinery of LINEs for their transposition. Like LINEs, they also generate TSDs upon integration to a new site (Deragon and Zhang, 2006; Kramerov and Vassetzky, 2011). The SINEs elements have been described in several plants (Cheng and Ling, 2006), PCR amplified in several species of *Gramineae*, *Fabaceae* and *Solanaceae* (Fawcett *et al.*, 2006) and well characterized in other plants as *S1* described in *Brassica* (Deragon *et al.*, 1996). The members of the *S1* family are ~170 bp in size and widely distributed among *Brassicaceae* family. Another well characterized family named *BoS*, has shown around 4290 copies in the *Brassica oleracea* genome covering 0.16% of the total genome (Deragon and Zhang, 2006). Recently a novel SINE Au element (GmAu1) was characterized from *Glycine max* (Shu *et al.*, 2011).

1.8 Class II transposable elements

Class II transposable elements or controlling elements constitute the major component of eukaryotes genomes. They are DNA fragments with an ability to insert in chromosomal sites, and generate duplicate copies during transposition. Class II transposable elements directly transpose via DNA and were first studied in plants. In size, they have shown an immense variability and start from a few hundred bases to about 10 kb. They have short terminal inverted repeats (TIRs), which are variable in different superfamilies. The classification proposed by Wicker *et al*., (2007) divides Class II elements into two sub-classes. Sub-class 1 is composed of common superfamilies as Tc1-Mariner, hAT, Mutator, Merlin, Transib, P, PiggyBac, PIF-Harbinger, CACTA and Crypton, while sub-class 2 includes the superfamilies of Helitron and Meverick elements (Figure 1.2). Out of these only Tc1-Mariner, hAT, CACTA, PIF-Harbinger and Mutator superfamilies are common in plants (Wicker *et al*., 2007) and are described here.

1.8.1 Ac/Ds-hAT

The first mobile DNA element to be discovered was the maize transposon *Activator* (*Ac*). Ac is an autonomous element while its non-autonomous partner Ds was identified soon after the discovery of the Ac element (McClintock, 1950). After that, many other transposons were studied and investigated in many species, sharing similarity to *Ac*-like elements suggesting their diverse nature and distribution in different organisms. The elements were named hAT after the discovery of *hobo* elements from *Drosophila*, *Ac* from maize, and *Tam3* from snapdragon (Rubin *et al*., 2001). The terminal inverted repeats (TIRs) of hAT elements are short and ill defined ranging from 5-27 bp long, generate 8 bp target TSDs upon transposition and exhibit transposase protein that catalyze the DNA breakage and rejoining reactions required for transposition. Several elements lack transposase and are non-autonomous transposons (Kempken and Windhofer, 2001). The transposase displays significant amino acid sequence similarity, with the highest primary structure conservation at their C-termini. The transposase is highly specific in hAT elements, which is composed of conserved blocks of amino acids and a DDE motif. A total of 147 hAT-related sequences in plants, animals, and fungi were studied and phylogenetic analysis and clustering of hAT sequences suggest that the hAT superfamily is very ancient, probably predating the plant-fungi-animal separation (Rubin *et al*, 2001).

Plant genomes contain multiple different members of the hAT superfamily. In maize, three distinct types of hAT element were investigated in addition to *Ac*, while rice exhibit four and *Arabidopsis thaliana* harbour five different types of hAT families. Some of these form loose phylogenetic clades, suggesting an ancient diversification of the superfamily before the monocot-dicot separation (Xu and Dooner, 2005). The hATs were studied in several plants like maize (Shimatani *et al.*, 2009; Du *et al.*, 2011; Fujino and Sekiguchi, 2011), sugarcane (de Jesus *et al.*, 2012), sugar beet (Menzel *et al.*, 2012), *Petunia hybrid*, *Phaseolus*, *Brassica napus* (De Keukeleire *et al.*, 2004) and *Arabidopsis* (Bundock and Hooykaas, 2005).

1.8.2 Tc1-Mariner

Tc1-Mariner elements move via DNA cut and paste mechanism. They are flanked by a 2 bp TSDs (TA), TIRs ranging from few to 33 bp and internal region encoding a transposase having characteristic amino acids designated as DDE/D motif. This motif consists of two aspartic acids and a glutamic acid residue (or a third D) with specific spacing of the nucleotides between the residues. A highly conserved domain of 150 aa surrounding the DDE/D motif is present in almost all the Tc1-Mariner elements. This conserved motif has established the evolutionary relationship of the Tc1-Mariner elements. The phylogenetic analysis of conserved regions of transposase and the distance between DDE/D motifs of Tc1-Mariner elements distinguished them into three monophyletic groups: Tc1-like, mariner-like and pogo-like (Doak *et al.*, 1994; Plasterk and van Luenen, 1997; Feschotte and Wessler, 2002). The Tc1-Mariner elements are abundant in animal genomes but their presence in plant genomes was investigated recently. The plant Tc1-Mariner elements have a long ORF, which is similar to the ORF of Tc1-Mariner elements from animals. The complete Tc1-Mariner element *Osmar1* (5259 bp) was studied in *Oryza sativa* (Tarchini *et al.*, 2000), *Soymar1* (3491 bp) in *Glycine max* (Jarvik and Lark, 1998) and *Vulmar1* (3909 bp) in *Beta vulgaris* (Jacobs *et al.*, 2004).

1.8.3 Mutator

Mutator (Mu) transposons are known to be the most mutagenic plant transposons and are widespread among angiosperms. They can capture the genetic sequences of the host and can mobilize the captured fragments to new sites causing evolution (Xian-Min, 2006). The

Mutator transposons are characterized by 9 bp target site duplications, 170-220 bp terminal inverted repeats and sometime have additional direct or indirect repeat sequences in their genomes. TIRs and TSDs are also conserved in these elements and remained constant with continued transposition activity, while internal regions are highly variable with no similarity to each other (Jiang *et al*., 2004; Xian-Min, 2006). Pack-MULEs are the non-autonomous Mutator elements capturing the host genes or gene fragments. Analysing the public EST data demonstrates that MULEs are not only important component of rice genome but they are also active in the genomes of other plants like wheat, sugar cane, rice and barley (Jiang *et al.*, 2004b). The maize Mu transposable elements are regulated by an autonomous element, MuDR encoding two genes, MuDRA and MuDRB. These two genes transcribe two segments from the opposite strands and produce 2.8 and 1.0 kb transcripts respectively (Lisch, 2002; Jiang *et al.*, 2004b).

Mutator family is a diverse family divided further into subfamilies including the Mu1/Mu2, Mu3, Mu4, Mu6/7, Mu8 and Mu9/Mu5. About 120-220 bp conserved terminal inverted repeats (TIRs) are shared within the subfamilies, while the internal sequences are variable and distinct (Lisch, 2002). Many Mutator-like elements are investigated and recently Pack-MULEs, which contain fragments of genes, were discovered. These MULEs are distributed among several species of *Gramineae* including wheat, barley, rice, sorghum and bamboo (Lisch *et al.*, 2001). The genome of rice harbour ~8000 copies of MULEs in their genomes, Out of which, 24% showed similarity to the coding regions of the other genes unrelated to transposons, indicating the capturing of genes (Juretic *et al.*, 2005). A complete (12089 bp) MuDR-like element designated as *CUMULE* was detected from *Cucumus melo* which were also investigated in *Arabidopsis* genome (van Leeuwen *et al.*, 2007).

1.8.4 En/Spm-CACTA

The En/Spm (Enhancer/Suppressor) elements are called CACTA due to their highly conserved motif in the termini of TIRs. The En/Spm are the autonomous CACTA elements while I/dSpm (Inhibitor/dSpm) are their non-autonomous counterparts. Both En and Spm are the autonomous elements due to possession of an active transposase. In contrast, their matching partners, the non-autonomous inhibitor and the defective Spm (dSpm) are the deletion derivatives of the autonomous elements (Gierl *et al.*, 1985; Pereira

et al., 1986). CACTA elements have 3 bp TSDs, flanked by short TIRs of 10-28 bp, widespread sub-terminal repeats and internal region encoding transposase. They are mostly recognised by their specific transposase, and 5′-CACTA.....TAGTG-3′ terminal motifs in their TIRs, due to which they are called as CACTA elements. Many subfamilies of CACTA superfamily have been described from the grass family as *Baldwin, Casper, Enac, Isaac, Jorge, Mandrake* and *TAT-1*. The internal sequences of the elements are highly divergent but 20-30 bp TIRs including CACTA motif are similar. They are not easily identified by computer aided database searches (Wang *et al.*, 2003; Wicker *et al.*, 2003; Tian, 2006). Generally, the autonomous CACTA elements contain a transposase protein but another additional protein in frequently present. One protein is named as TNPD, which is the transposase of the CACTA, responsible for its transposition and integration while the other is called as TNPA, a factor performing multiple functions (Gierl and Saedler, 1989; Trentmann *et al.*, 1993). The CACTA elements are investigated in several plants including maize and sorghum (Lee *et al.*, 2005), rice (Kwon *et al.*, 2006), temperate grasses and cereals (Langdon *et al.*, 2003), *Triticaceae* members (Wicker *et al.*, 2003), *Arabidopsis* (Miura *et al.*, 2004; Marsch-Martinez and Pereira, 2011), *Brassica* (Alix *et al.*, 2008), *Glycine max* (Zabala and Vodkin, 2008) and *Manihot esculenta* (Gbadegesin and Beeching, 2010).

1.8.5 PIF-Harbinger

PIF-Harbinger is a superfamily of DNA transposons characterized by generating 3 bp TSDs, flanked by 14-25 (50 bp in few elements) bp TIRs and a DDD/E transposase, which is the enzymatic machinery required for their transposition. The first Harbinger was identified in *Arabidopsis thaliana* (Kapitonov and Jurka, 1999), which showed similarity to maize P instability factor (PIF) elements (Zhang *et al.*, 2001; Zhang *et al.*, 2004). The diverse PIF-Harbinger elements are easily distinguishable into two subgroups, named PIF and Pong (Zhang *et al.*, 2004). Harbinger superfamily is highly diverse and its members are present in protists, insects, worms, vertebrates and plants. This is the only superfamily of DNA transposons where the autonomous elements encode two protein domains; the first is the superfamily specific transposase and the second is a DNA-binding protein domain. The DNA binding domain is characterized by having different conserved motifs as SANT/myb/trihelix (~70 aa), while the other region of DNA binding domain showed no significant similarities studied in different species (Kapitonov and Jurka, 2004). The

most conserved domain in all the Harbingers is the transposase, which is composed of 5 conserved motifs. The Harbingers are flanked by TAA target site duplications, but few families generate other TSDs. The most unusual feature of the Harbingers is the presence of second protein, which is not observed in any other superfamily of DNA transposons (Jurka and Kapitonov, 2001; Kapitonov and Jurka, 2004). Harbingers are identified from few plants like *Medicago truncatula* (Grzebelus *et al.*, 2007), carrot (Grzebelus *et al.*, 2006; Grzebelus and Simon, 2009) and *Arabidopsis* (Kapitonov and Jurka, 2004).

1.9 Miniature inverted-repeat transposable elements (MITEs)

Miniature inverted-repeat transposable elements are small elements present almost in many eukaryotic genomes. They are <600 bp in size, generate TSDs, flanked by TIRs of variable lengths and lack any ORF encoding protein domains. For their mobility, they rely on the activity in *trans* of transposases encoded by the nearest autonomous transposons. It is believed that the full length DNA transposons are the evolutionary progenitors of the MITEs, based on the similarity of TSDs and TIRs of the MITEs with other DNA transposons (Jiang *et al.*, 2004b). A 128 bp insertion in a mutant maize waxy gene was the first element, which led to the identification of a diverse group of elements named Tourist from various grass species (Bureau and Wessler, 1992; Bureau and Wessler, 1994a). Another 257 bp element from sorghum was the originator of second family of elements called Stowaway, which was laterally studied in many other plants. Majority of the Stowaway elements described are small in size (70-350), have conserved termini of 11 bp and have a TA dinucleotide preference for insertion (Bureau and Wessler, 1994b).

Tourist and Stowaway MITEs share some structural features like their sizes and short TSDs but their sequences are distinct. The members of both elements range in size from ~100-500 bp, flanked by conserved terminal inverted repeats (TIRs), terminated by target site repeats and no coding domain in their internal regions. The Stowaway elements generate 2 bp TSDs ‘TA’, whereas the Tourist elements generate a 3 bp ‘TAA’ TSDs upon integration to a new site. Due to the high copy numbers and uniformity of Tourist, Stowaway and other similar elements in many species, they were collectively brought under a same group called MITEs (Wessler *et al.*, 1995; Bureau *et al.*, 1996). MITEs are highly diverse and abundant group of TEs identified from several plants including *Spring* MITE from crops from *Gramineae* (Park *et al.*, 2003), barley (Takahashi *et al.*, 2006;

Lyons *et al.,* 2008), rice (Oki *et al.*, 2008; Lu *et al.*, 2012), potato (Momose *et al.*, 2010), pea (Macas *et al.,* 2005), apple (Han and Korban, 2007), peanut (Shirasawa *et al.,* 2012), grapevine (Benjak *et al.*, 2009), *Medicago truncatula* (Grzebelus *et al.*, 2009), *Beta vulgaris* (Menzel *et al.,* 2006), *Pennisetum glaucum*, (Remigereau *et al.*, 2006), *Arabidopsis* (Santiago *et al.*, 2002) and *Brassica* (Sarilar *et al.*, 2011).

1.10 Impact of Transposable elements (TEs)

1.10.1 TEs and evolution

Almost all the TEs have a wide range of activities within the genomes, being related to chromosomal breakage, chromosomal rearrangements, altered gene regulations, insertional mutations, sequence amplification and duplication (Bennetzen, 2000). Studies in various plant genomes have shown that TEs have played a major role in plant genome evolution (Flavell *et al.,* 1994). The activity of retrotransposons is one of the sources of evolution in yield-related trait variations in introgressed line populations of *Brassica napus* (Zou *et al.*, 2011).

1.10.2 TEs as mutagens

TEs are powerful mutagens that can alter any potential gene by their insertion in the coding region of the gene and produce the altered expression patterns by insertions into the regulatory regions. The positions of the TE insertions in the mutant genes can be identified by using the TEs as probes. The insertions can be PCR amplified by designing one oligonucleotide primer from the gene and other from the transposon insertion (Reviewed by Kunze *et al.*, 1997).

1.10.3 Transposon Tagging

Transposon tagging is a tool to investigate and study the localization of the genes and is very useful technique to locate the position of the genes or the insertional patterns of TEs in various genes. Transposon insertions into various gene loci cause several types of mutations and even the non-functioning of the normal genes. This integration of transposon insertion in the gene can cause the appearance of phenotypically new

phenotype. Excision of same insertion partially or completely restores the normal genes and wild type. However, reversion events often are followed by small mutations such as leaving small finger prints like TSDs or small segments of the element, or few bases of sequences produced either by illegitimate conversion or other forms of repair of the excision site. Thus the mutations in the genes are caused by both insertions and excisions of the transposons, although insertions contributed more than the excision (reviewed by Kunze at al., 1997; Doring and Starlinger, 1986; Bennetzen, 2000). By using the transposon probe, the changes can be observed as restriction fragment length polymorphism (RFLPs). These transposon probes can be used to see their activity and integration in various genes as recently investigated in *Arabidopsis* (Marsch-Martinez and Pereira, 2011).

1.10.4 TEs as genetic markers

Because of their mobility and activity, transposons have proved to be valuable markers for genetic diversity and variability. In particular, use of outward facing primers to amplify around pairs of insertion sites, the IRAP technique, (or between retroelements and simple sequence repeats, the ISSR method) first developed by Schulman, (2004) has been useful to provide multi-locus anonymous markers. The detailed protocols, uses and application of SSAP, IRAP, REMAP and RBIP markers were described recently. The SSAP, IRAP and REMAP methods are multiplex and are used to generate several anonymous marker bands, while RBIP targets the individual loci. REMAP markers are also applicable in several phylogenetic and biodiversity studies. Retrotransposon-based insertional polymorphisms (RBIP) have been developed, with the advantages of being a co-dominant markers (Flavell *et al.*, 1998; Schulman *et al.*, 2004; Kalendar and Schulman, 2006; Schulman *et al.*, 2012). These markers are now used in several studies to observe biodiversity and phylogenetic relations of the species and to investigate the retrotransposons. ‘Sequence-specific amplification polymorphism’ (SSAP) markers, first developed to locate the BARE-1 in barley genome (Waugh *et al.*, 1997) and lateral developed methods (Syed and Flavell, 2006) are now extensively used to detect the presence/absence of various transposons at specific loci of the organisms. More recently, SINE mobility has been used to measure relationships of potato varieties (Seibt *et al.*, 2012). MITEs are considered as best molecular markers based on their presence/absence polymorphisms for biodiversity and phylogenetic studies (Lyons *et al*., 2008). The genetic

diversity among *Triticum* and *Aegilops* species was studied by MITE-based markers, which revealed the clustering of the species based on genus, genome composition and ploidy level and a genetic divergence was observed between diploids versus polyploids (Yaakov *et al.*, 2012).

1.11 Aims and objectives of the study

With the advancement in genome sequencing, the knowledge of the structure and composition of genomes is rapidly increasing. Due to the abundance and diversity of TEs in different genomes, the identification, annotation, localization and proper classification of TEs is most important. The aims of the study were

- To identify and characterize all different types of mobile elements: transposable elements from Class I (retrotransposons and non-LTR retrotransposons) and Class II (DNA transposons) elements in two reference genera *Brassica* and *Musa* by using bioinformatics and molecular methodologies.

- To investigate the small non-autonomous transposable elements (TEs), both retrotransposons and DNA transposons with special emphasis on small novel insertions, which are structurally different from known superfamilies, given less importance and not investigated in detail previously due to their small sizes and difficulty in their identification and characterization.

- To identify the autonomous TEs from all major superfamilies of retrotransposons (Copia, Gypsy, LINEs, SINEs) and DNA transposons (Tc1-Mariner, CACTA, hAT, Harbinger, Mutator) in diploid and polyploid *Musa* and *Brassica* genomes. To compare structures of newly discovered transposons with the known elements and study their evolutionary relationships. To identify progenitors of non-autonomous retrotransposons (LARDs, TRIM) and MITEs.

- To develop novel transposon-based molecular markers as previously developed by several research workers such as ‘retrotransposons-based insertion polymorphism’ (RBIP), ‘sequence-specific amplification polymorphisms’ (SSAP) and ‘simple sequence repeat’ (SSR) like molecular markers. The main emphisis was on

developing new co-dominant markers targeting the insertional/empty sites (presence/absence) of TEs. Similarly to develop markers to analyse the insertional polymorphism of MITEs, SINEs, non-autonomous TEs and novel insertions.

- To study the biodiversity, phylogenetic relationship and genetic linkage of various *Brassica/Musa* genomes by using these newly developed genetic markers and to study distribution and abundance of the transposons from a group of genomes in *Brassica* and *Musa*. This will give a new insight into the abundance and distribution of TEs in both genomes by utilizing these genetic markers.

Overall, these aims will allow identification, characterization, naming and annotation of TEs and their use to study the mechanism and pattern of evolution in different diploid and polyploid *Brassica* and *Musa* species and their cultivars.

CHAPTER 2
MATERIAL AND METHODS

2.1 Plant material for *Brassica*

The DNAs from 40 *Brassica* accessions/cultivars were used in the present study. Seeds from 32 *Brassica* accessions were brought from Warwick Research Institute (WRI), Warwick, UK. Two *Brassica juncea* accessions (NARC-1, NARC-II) and one *Brassica carinata* accession (NARC-PK) were brought from Institute of Agri-Biotechnology and Genetic Resources, National Agriculture and Research Center (NARC), Islamabad, Pakistan. Seeds for one commercial variety *Brassica juncea* accession (NATCO) were bought from Asian store at Leicester. The DNA from four synthetic allohexaploids (2n=6x) *Brassica* (Ge *et al.*, 2009) were provided by Xianhong Ge (University of Wuhan, China). The seeds were grown in a green house at Department of Biology, University of Leicester, UK. The names, accessions, genome constituent and ploidy level of all genomes investigated are listed in Table 2.1.

2.2 Plant material for *Musa*

A total of 48 *Musa* accessions (Set A) were maintained in a heated greenhouse at Botanical Garden of University of Leicester, UK. These plants were obtained in tissue culture from the International Transit Centre (ITC) of Bioversity, Leuven, Belgium and transferred to pots in the greenhouse. The DNA was extracted from the young and fresh leaves from these plants. Another collection of 48 *Musa* genomes (Set B) were kindly provided by Professor Ashalatha (Asha) Nair, University of Kerala, India, former research associate at Cytogenetics lab, Department of Biology, University of Leicester, UK. The names, accessions, genome constituent and ploidy level of all *Musa* genomes are listed (Set A: Table 2.2; Set B: Table 2.3).

Table 2.1: List of *Brassica* species and accessions with their accessions and crop names. ND: Not Determine

Sr.No.	Accession No.	Species	Accession Name	Crop name
1	HRIGRU 2488	*B. rapa chinensis*	Pak Choy	Chinese cabbage
2	HRIGRU 2741	*B. rapa pekinensis*	Chinese Wong Bok	Chinese cabbage
3	HRIGRU 7574	*B. rapa chinensis*	San Yue Man	Pak choi
4	HRIGRU 11698	*B. rapa rapa*	Hinona	Turnip
5	HRIGRU 13174	*B. rapa rapa*	Vertus	Turnip
6	ND	*B. rapa*	Suttons	Turnips (Snow balls)
7	HRIGRU011011	*B. nigra*	ND	Wild Species
8	HRIGRU010978	*B. nigra*	ND	Wild Species
9	HRIGRU010919	*B. nigra*	ND	ND
10	PK- 001722	*B. juncea*	NARC-I	ND
11	ND	*B. juncea*	NATCO	ND
12	PK-001325	*B. juncea*	NARC-II	ND
13	HRIGRU 2203	*B. oleracea gemmifera*	De Rosny	Brussels sprout
14	HRIGRU 5108	*B. oleracea*	Kai Lan	ND
15	HRIGRU2859	*B. oleracea*	Early Snowball	Cauliflower
16	HRIGRU 7518	*B. oleracea italica*	Precoce Di Calabria Tipo Esportazione	Broccoli
17	HRIGRU 3211	*B. oleracea capitata*	Cuor Di Bue Grosso	Cabbage
18	GK97361	*B. oleracea*	ND	ND
19	HRIGRU 2487	*B. juncea*	Kai Choy	Mustard cabbage
20	HRIGRU 7563	*B. juncea*	Megarrhiza	Large rooted mustard
21	HRIGRU 11702	*B. juncea*	Tsai Sim	Chinese mustard
22	HRIGRU 11974	*B. juncea*	W3	Indian oilseed
23	HRIGRU 12818	*B. juncea*	Giant Red Mustard	Japanese greens
24	ND	*B. juncea*	Varuna	ND
25	HRIGRU 11967	*B. napus*	New	Hakuran
26	HRIGRU 12685	*B. napus oleifera*	Mar	Oilseed rape
27	HRIGRU 12800	*B. napus biennis*	Last and Best	Kale
28	HRIGRU 13554	*B. napus napoB.*	Fortune	Swede
29	ND	*B. napus*	Drakker	ND
30	ND	*B. napus*	Tapidor	ND
31	HRIGRU 2485	*B. carinata*	Addis Aceb	Ethiopian mustard
32	HRIGRU 6232	*B. carinata*	Patu	Ethiopian mustard
33	HRIGRU 6986	*B. carinata*	Tamu Tex-sel Greens	Ethiopian mustard
34	HRIGRU 13160	*B. carinata*	Mbeya Green	Ethiopian mustard
35	R.G.F 32275	*B. carinata*	Aworks-67	ND
36	PK- 0085490	*B. carinata*	NARC-PK	ND
37	ND	*B. napus* x *B. nigra*	ND	ND
38	ND	*B. carinata* x *B. rapa*	ND	ND
39	ND	*B. napus* x *B. nigra*	ND	ND
40	ND	*B. napus* x *B. nigra*	ND	ND

Table 2.2: Set A. List of *Musa* species and accessions with accession names and numbers. ND: Not Determine.

Sr.No.	Reference	Genome	Accession	Country	ITC Number
1	Eumusa	AAB	Lady Finger	India	ITC.0582
2	Eumusa	AAB	Foconah	Cameroon	ITC.0649
3	Eumusa	AAB	Prata Ana	Brazil	ITC.0962
4	Eumusa	*balbisiana*	P. Klutuk Wulung,	Indonesia	ITC.1063
5	Eumusa	*balbisiana*	P. Batu, IDN 080	Indonesia	ITC.1156
6	Eumusa	*acuminata*	Banksii 623	Papua new guinea	ND
7	Eumusa	*acuminata*	Borneo	Malaysia, Borneo	ITC.0253
8	Eumusa	*acuminata*	Calcutta 4	Calcutta , India	ITC.0249
9	Eumusa	ABB	THA 020	Thailand	ITC.0652
10	Eumusa	AAB	Orishele	Nigeria	ITC.1325
11	Eumusa	ABB	Pelipita	Philippines	ITC472
12	Eumusa	ABB	Dole	ND	ITC.0767
13	Eumusa	AAA	Grande Naine	Guadeloupe	ND
14	Eumusa	AAA	Pisang Kayu, (IDN098)	Indonesia	ITC0420
15	Eumusa	*acuminata*	Agutay	Philippines	ITC.1028
16	Eumusa	*acuminata*	Khae (Phrae), THA 015	Thailand	ITC.0660
17	Eumusa	AAB	Figue Pomme Géante	Guadeloupe	ITC.0769
18	Eumusa	ABB	Saba	Philippines	ITC.1138
19	Eumusa	AAA	Pisang bakar , IDN106	Indonesia	ITC.1064
20	Eumusa	ABB	Monthan	India	ITC0046
21	Eumusa	*balbisiana*	Tani	ND	ND
22	Eumusa	*acuminata*	Long Tavoy pied	ND	ITC.0283
23	Eumusa	AB cv	Safet Velchi	India	ITC.0245
24	Eumusa	AAA	Petite Naine	ND	ITC.0654
25	Eumusa	*acuminata*	Paliama, PNG067	Papua New Guinea	ITC.0766
26	Eumusa	AAA	Poyo	Nigeria	ND
27	Eumusa	AAB	Popoulou	Cameroon	ITC.0335
28	Eumusa	ABB	Simili Radjah	India through Zaire	ND
29	Eumusa	AAA	Gros Michel	Guadeloupe	ND
30	Eumusa	AS	Wompa, PNG063	Papua New Guinea	ITC.1152
31	Eumusa	AB cv	Kunnan	India, Kerala	ITC.1034
32	Eumusa	AAcv (18)	P. Jari Buaya/BS312	Malaysia, Thailand	ITC.0312
33	Eumusa	AAcv (2)	P. mas / Figue Sucrée	Malaysia	ITC.0653
34	Eumusa	AAB	P. Raja Bulu, IDN 093	Indonesia	ITC.0843
35	Eumusa	AAA	Leite	ND	ITC.0277
36	Eumusa	ABB	Ice Cream	ND	ITC020
37	Eumusa	*acuminata*	Zebrina	Indonesia	ITC.1177
38	Eumusa	AAcv	Tomolo, (PNG023)	East New Britain	ITC.1187
39	Eumusa	*balbisiana*	Honduras	seeds from Honduras	ITC.0247
40	Eumusa	*balbisiana*	Lal Velchi	India	ND
41	Eumusa	ABB	Namwa Khom, THA011	Thailand	ITC0659
42	Eumusa	AAA	Mbwazirume	Burundi	ITC.0084
43	Eumusa	AAA	Intokatoke	Burundi	ITC.0082
44	Eumusa	AAA	Yangambi KM5	Cameroon	ITC.1123
45	Eumusa	AAB	Red Yade	ND	ITC.1140
46	Eumusa	AAB	P. Rajah	Brazil	ITC.0243
47	Eumusa	ABBB	Yawa 2, PNG 072	East New Britain	ITC1238
48	Eumusa	AAB	P. Ceylan	Thailand	ITC1441

Table 2.3: Set B. List of various *Musa* species and accessions with their accession names and genome compositions.

Sr. No.	Reference	Genome	Accession Name	Sr. No.	Reference	Genome	Accession Name
1	Eumusa	AA	Calcutta 4	25	Eumusa	AAB	Karimkadali
2	Eumusa	AA	Sannachenkadali	26	Eumusa	AAB	Perumadali
3	Eumusa	AA	Pisanglilin	27	Eumusa	AAB	Kunoor ettan
4	Eumusa	AA	Kadali	28	Eumusa	AAB	Palyamcodan
5	Eumusa	AA	Matti	29	Eumusa	AAB	Mysoreettan
6	Eumusa	AA	Cherukadali	30	Eumusa	AAB	Krisnavazhai
7	Eumusa	BB	PKW1	31	Eumusa	AAB	Poovan
8	Eumusa	BB	PKW2	32	Eumusa	AAB	Doothsagar
9	Eumusa	BB	Javan	33	Eumusa	AAB	Charapadati
10	Eumusa	BB	Klutuk	34	Eumusa	AAB	Kumbillakannan
11	Eumusa	BB	Tani	35	Eumusa	AAB	Velipadati
12	Eumusa	BB	Batu	36	Eumusa	AAB	Vellapalayamcodan
13	Eumusa	AB	Njalipovan	37	Eumusa	AAB	Ettapadati
14	Eumusa	AB	Adukkan	38	Eumusa	AAB	Padati
15	Eumusa	AB	Padalamukili	39	Eumusa	AAB	Chinali
16	Eumusa	AAA	Manoranjitham	40	Eumusa	AAB	Nendran
17	Eumusa	AAA	Grandnain	41	Eumusa	AAB	Poomkalli
18	Eumusa	AAA	Grow-michel	42	Eumusa	AAB	Kamaramasengi
19	Eumusa	AAA	Greenred	43	Eumusa	ABB	Kosta bontha
20	Eumusa	AAA	Red	44	Eumusa	ABB	Peyan
21	Eumusa	AAA	Monsmari	45	Eumusa	ABB	Kanchikela
22	Eumusa	AAA	Robusta	46	Eumusa	ABB	Boothibale
23	Eumusa	AAA	Dwarf cavendish	47	Eumusa	ABB	Monthan
24	Eumusa	AAB	Motta povan	48	Eumusa	ABB	Karpooravali

2.3 Solutions

1L 50X TAE stock

242 g of Tris base

57.1 ml of glacial acetic acid

100 ml of 0.5M EDTA (pH 8.0)

1L 5X TBE stock

54 g of Tris base

27.5g of boric acid

20 ml of 0.5M EDTA (pH 8.0)

6X loading buffers

0.25% Bromophenol Blue

0.25% Xylene cyanol FF

60% Glycerol

10X Enzyme Buffer

40 ml 100 mM Citric acid

60 ml 100 mM Tri-Sodium citrate

Store at 4 ˚C. Dilute 1:10 for 1x solution

10X TE Buffer

100 M Tris-HCl (pH 8.0)

0.5 M EDTA

10 mM EDTA (pH 8.0)

DNA Wash Buffer

76% Ethanol

10 mM NH4ac

Ethidium Bromide (10 mg/ml)

1g Ethidium bromide

100 ml H_2O

EDTA 500 mM (pH 8.0)

186.1g disodium EDTA.$2H_2O$

812.9 ml ml of water

Laboratory chemicals were obtained from Sigma-Aldrich except where noted. Water for buffers and large volumes was obtained from a double-deionization system (building deionization followed by Elgastat in the laboratory). For PCR, restriction digestions, dissolving DNA or other molecular biology, water supplied with reagents or Molecular Biology grade water from Sigma-Aldrich ("Sigma water") was used.

2.4 Extraction of genomic DNA

Genomic DNA was extracted from the young and fresh leaves of *Musa* and *Brassica* plants with a slight modification of the Doyle and Doyle standard cetyltrimethylammonium bromide (CTAB) isolation protocol (Doyle and Doyle, 1990). The frozen leaves were ground with mortar and pestle to powder form adding liquid nitrogen to prevent enzymatic degradation and release of phenolic compounds. The powder was added to a tube containing preheated CTAB buffer (2% CTAB, 20 mM EDTA, 100 mM Tris-Cl at pH: 8.0, 1.4 M NaCl, and 0.2% marcaptoethanol) @ 5ml per gram fresh weight of leaf tissue. After mixing well, the slurry mixture was incubated for 60 minutes at 60 ˚C. An equal volume of chloroform/iso-amyl alcohol (24:1) was added, mixed for 3-5 minutes and all contents were transferred to narrow bore centrifuge tubes. After balancing by adding extra chloroform/iso-amyl alcohol, the mixture was spun at 5,000 rpm for 10 min. The supernatant was removed and chloroform extraction was repeated. DNA was precipitated with 0.66 volume of cold isopropanol, collected by centrifugation or spooled out by glass rod and transferred to DNA wash buffer for 20 minutes. DNA was air dried briefly and 250-500 µl of TE was added and left overnight before adding 1 µl (10ng/ml) RNase to each 1 ml TE/DNA mixture and incubating for 45 minutes at 37˚C. DNA was spooled out; air dried and re-suspended in 0.5 to 1 ml T.E. (8-24 hours; final concentration c. 0.1 to 1 µg/ul) and stored frozen at -20 ˚C.

2.5 DNA quantification

DNA was quantified by using a diode array scanning spectrophotometer (Amersham Biosciences) after dilution with distilled water 1:40 – i.e. 5 µl of DNA + 195 µl of distilled water to make 200 µl final solution. Using the spectrophotometer, 200 µl of de-ionized water was added in black flask and reference was clicked until 0.000 was displayed. Water was removed from the cuvette and a DNA sample was loaded and the absorption readings were taken at 260 nm and 280 nm. The DNA concentration was measured by the formula: DNA ng/µl = A_{260} x 50 x 40, where A_{260} is the absorption reading, 50 is convertion factor (50 ng/µl), and 40 is the dilution factor. The ratio of the absorbance at 260 nm and 280 nm were used to determine the purity of the DNA samples. Samples with ratio of 1.8 or greater were used for PCR amplification. The DNA samples were also run on 1% agarose gel and quality and quantity of DNA was approximately observed by comparing the bands with already known markers.

2.6 Development of new molecular markers for retrotransposons amplification polymorphisms

Several new molecular markers were developed based on the modification of previously described genetic markers. For the amplification of autonomous LTR retrotransposons (Copia, Gypsy) and non-LTR retrotransposons (LINEs), the primers were designed from the most conserved region of their reverse transcriptase (RT) around the D-DD triad (Flavell *et al.*, 1992a) between block III-V by Primer3 (v.0.4.0) (http://frodo.wi.mit.edu/primer3/), which we called 'reverse transcriptase amplification polymorphism' (RTAP) markers (Figure 2.1). For non-autonomous LTR retrotransposons or LARDs elements, 'LARDs amplification polymorphism' (LAP) markers (primers) were designed from 5′ LTRs. Genetic markers to amplify non-autonomous LINEs and SINEs were designed from the regions flanking the TEs, which were named 'transposon insertional polymorphism' (TIP) markers (Figure 2.1). Both RTAP and TIP are co-dominant markers designed to indicate presence/absence polymorphisms of the TEs at specific insertional sites/loci. The degenerate primers to amplify LTR retrotransposons (Copia, Gypsy) and non-LTR retrotransposons (LINEs, SINEs) from *Brassica* and *Musa* are listed in respective chapters.

2.7 Designing PCR primers (markers) for DNA transposons and MITEs

For the autonomous DNA transposons, the primers were designed from the conserved regions of transposase around DDD/E triad by Primer3 (v.0.4.0) (http://frodo.wi.mit.edu/primer3/) and were named 'transposase amplification polymorphism' (TAP) markers (Figure 2.1). For non-autonomous DNA transposons and MITEs, the TIP molecular markers were designed in forward and reverse directions on upstream and downstream of each transposable element from the flanking regions to amplify the insertional or pre-insertional (or empty) sites. The list of primers amplifying various transposon insertions are given in respective chapters.

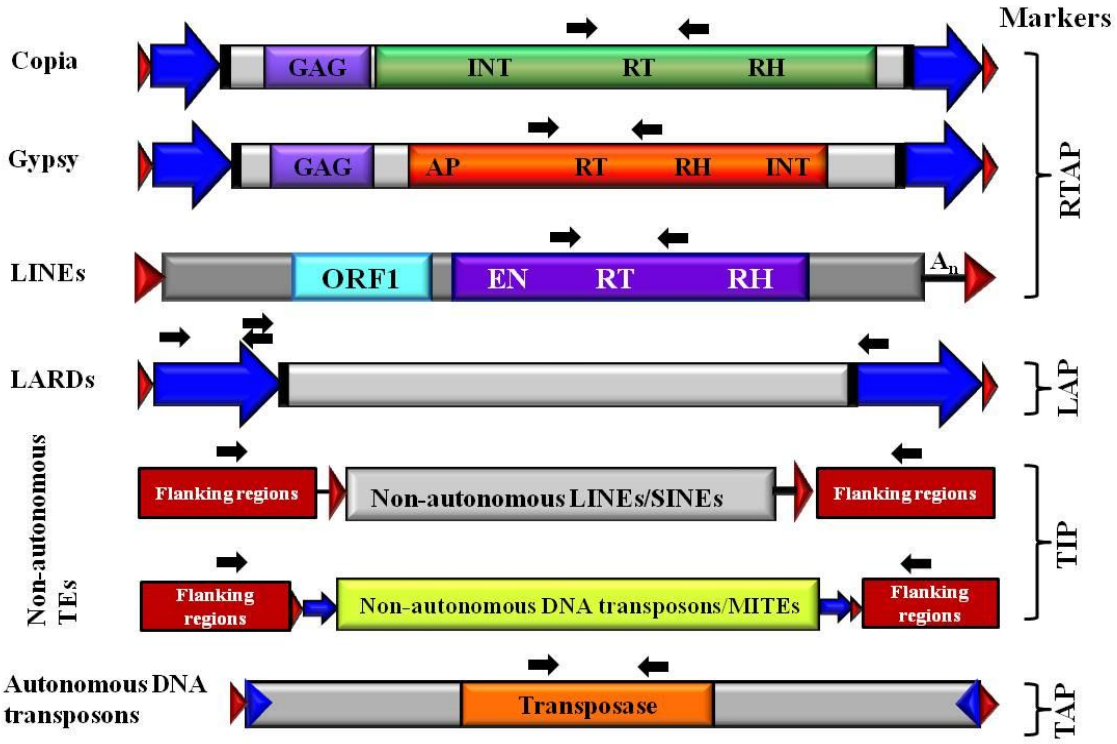


Figure 2.1: Schematic representation of positions of primers designed for different autonomous and non-autonomous transposable elements. The black arrows indicate the positions of primers to amplify various regions of TEs. The primers for autonomous retrotransposons were designed from their most conserved RT region by reverse 'transcriptase amplification polymorphism' (RTAP) markers. For LARDs, the primers were designed from LTRs to amplify the single LTR or whole element by LARDs amplification polymorphism markers (LAP). The 'transposon insertional polymorphisms' (TIP) markers (primers) were developed from the flanking regions of non-autonomous LINEs, SINEs, DNA transposons and MITEs to amplify the insertional sites/loci or empty sites. The primers for autonomous DNA transposons were designed from conserved transposase regions by 'transposase amplification polymorphism' (TAP) markers. In a few cases, alternative domains were used to design the primers to amplify other regions.

2.8 Polymerase chain reactions (PCRs)

Polymerase chain reaction (PCR) was used for the amplification of fragments derived from various transposable elements. Total volume of reaction mixture varied and ranged from 15-25 µl. The genomic DNA was used @ 50-75 ng/µl with 10X Kapa Taq buffer A (Kapa Biosystems, UK), additional 1.0 mM $MgCl_2$, 200-250 µM dNTP (2-2.5 mM; YORKBIO), 10 µM (10 pmoles) of each primer (SIGMA-ALDRICH) and 0.5-1 U of 5U/µl Taq polymerase (Kapa Biosystems, UK). The master mix was mixed well and was kept in ice to keep the DNA and Taq polymerase stable. The PCR conditions were optimized with some minor modifications in time, annealing and extension temperatures and gradient was set in TGradient Thermocycler (Biometra) to gain best amplification. The reaction mixture volume and temperatures are described below.

PCR reaction (15 µl)	**Volume**
DNA (50-75 ng)	1-1.5 µl
Kapa Buffer A (10X)	2 µl
$MgCl_2$ (25 mM)	1 µl
dNTPs (2.5 mM)	1 µl
Forward Primer (10 µM)	0.75-1 µl
Reverse Primer (10 µM)	0.75-1 µl
Kapa Taq (5U/µl)	0.1 µl
Sigma water	8 µl

PCR Program for retrotransposons

Initial denaturation	94˚C	3 min	
Denaturation	94˚C	45 sec-1 min	34 cycles
Annealing	Primer dependent	45 sec-1 min	
Elongation	72˚C	45 sec-1 min	
Final Elongation	72˚C	5 min	
Pause	16˚C	∞	

PCR Program for DNA transposons

Initial denaturation	94˚C	5 min	
Denaturation	94˚C	1 min	34 cycles
Annealing	Primer dependent	1 min	
Elongation	72˚C	1 min	
Final elongation	72˚C	10 min	
Pause	16˚C	∞	

2.8.1 Agarose gel electrophoresis

DNA fragments were separated on the basis of their sizes by agarose gel electrophoresis. A 3-5 µl of 6x loading buffer was added into PCR product depending on the quantity of PCR product (15-25 µl). A 1-1.5% (w/v) agarose gel was prepared according to the size of expected DNA fragments. For the clear visibility of DNA bands, 0.75-1.5 µl ethidium bromide (10 mg/ml) was added according to the volume of 1X TE used. The DNA samples were run and the amplicons were separated by agarose gel electrophoresis typically at c. 5 V/cm. DNA bands were observed and images were captured with the Gene Flash gel documentation system (Syngene, UK).

2.8.2 Isolation and purification of gel bands

The gel bands were isolated and purified by using protocol of MinElute Gel Extraction Protocol from Qiagen Quick Gel Extraction Kit (Qiagen, Hilden, Germany). The sharp bands were cut out with a sterilized and sharp scalpel or disposable blades. The gel slice was weighed in colourless microcentrifuge tubes and 3 volumes of buffer QG was added. Samples were incubated at 50˚C for 10 minutes and were mixed by vortexing the tube every 2-3 minutes during incubation. A 1 gel volume of cold isopropanol was added to the sample and mixed well by inverting the tubes. A MinElute column was placed in 2 ml collection tubes; samples were transferred to these MinElute columns and centrifuged for 1 minute. Flow-through was discarded and MinElute columns were placed back in the same collection tube. A 500 µl of buffer QG was added and centrifuged for 1 minute. After discarding flow-through, 750 µl of buffer PE was added to MinElute columns and centrifuged for 1 minute. This step was repeated again to clean the samples from any contamination of salts or buffers. MinElute columns were placed in clean 1.5 ml

centrifuged tubes, 10-15 µl of Sigma or distilled deionised water was added to the centre of membrane to elute the DNA and centrifuged twice for 1 minute to collect all DNA.

2.8.3 DNA sequencing and analysis

The amplicons after purification by MinElute gel extraction kit protocol were sent to DNA sequencing Enterprise Limited at John Innes Center Genome Laboratory, Norwich by sending the forward primers (1.5 pmol) with samples for sequencing. The resulting DNA sequence chromatograms were opened using the bioinformatics software Chromas version 1.45 (Conor McCarty, Griffith University, Australia). They were exported in FASTA format using sequence export tool present in Chromas version 1.45. The high quality sequences were retained, while sequences with poor quality were removed. The sequences were aligned with the transposon insertions and homology and differences between the query and sequenced element were studied.

2.9 Fluorescent *in situ* hybridization

Geminated root tips from 2-3 days seeds were used for the preparation of mitotic chromosomes. The complete DNA transposons/MITEs or the conserved regions of autonomous retrotransposons (RT) and DNA transposons (transposase) were PCR amplified, gel separated and cleaned with standard Qiagen Gel Extraction Protocol. The DNA was labelled with digoxigenin or biotin with random primers labelling protocol and used as probes. FISH of *Brassica/Musa* chromosomes was performed according to the protocol described by Schwarzacher and Heslop-Harrison, (2000). Chromosomes were counterstained with 0.2 mg/ml DAPI (4', 6-diamidino-2-phenylindole) diluted in McIlvaine's buffer pH7 and mounted in antifade solution (Citifluor). The probe mixture contained 50% (v/v) formamide, 20% (w/v) dextran sulfate, 2 x SSC, 25-100 ng probe, 20 mg of salmon sperm DNA and 0.3% sodium dodecyl sulphate (SDS) as well as 0.12 mM ethylene-diamine-tetraacetic acid (EDTA). Hybridization and washing were carried out at low stringency. Examination of slides was carried out with Zeiss epifluorescence microscope single band pass filters equipped with a CCD camera (Optronics, model S97790). The images were refined using only functions that affect the whole image equally and printed using Adobe Photoshop CS3 software.

2.10 Bioinformatics and computational analysis

Several computation methods were used to identify, characterize and classify novel transposable elements into their respective superfamilies and families and study their evolutionary relationships.

2.10.1 Dot plot analysis for identification of retrotransposons

In the present study, a novel approach was developed for the identification of TEs named 'Dot plot characterization of TEs' (DPCTE). This method is highly effective for the identification and characterization of various types of TEs (autonomous, non-autonomous) as well as small insertions/deletions within the genomic sequences. The approach is based on the dot plot comparison of homoelogous and homologous BAC/genomic sequences, where it indicate the gap-insertion pairs or parallel or vertical lines across the central diagonal line (showing homology between the two sequences). The TEs are indicated by gaps in continuous line showing homology, which are confirmed by analysing their TSDs at flanking ends. The LTR retrotransposons are represented by having two parallel lines (indicating LTRs) across the central diagonal line (Figure 2.2 & 2.3; also see Conclusion; Figure 10.1-10.3).

A total of 90 *Brassica* BAC sequences (Table 2.4), 84 from National Center for Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov/) and 6 from European Bioinformatics Institute (EBI: http://www.ebi.ac.uk) databases were randomly collected to screen various types of retrotransposons. To investigate the retrotransposons among *Musa* genomes, 46 BAC sequences were collected from NCBI (Table 2.5). Initially the candidates of full length LTR retrotransposons were identified by running each BAC genomic sequence against itself in dot plot analysis in the Dotter program (Sonnhammer and Durbin, 1995). The central diagonal line extending from one corner of the dot plot to the diagonally opposite corner represented the homology of the sequence. The LTRs on both termini are represented by 2 small diagonal lines at opposite corners indicating 5 and 3 LTRs (Figure 2.3 & 2.4). The 5′-TG….CA-3′ termini of LTRs were defined by dot plot analysis by scrolling bar to the terminal ends of the line showing the LTRs. The number of nucleotides in LTRs was counted and TSDs were searched by visual inspection.

The initial identification of the novel non-LTR retrotransposons or retroposons (LINEs, SINEs) in *Brassica* species was done by the comparison of homoeologous BAC sequences in Dotter program, where LINEs or SINEs were identified as gap-insertion pairs in the diagonal line indicating the highly homologous region between two sequences. Four pairs of *Brassica rapa* and *Brassica oleracea* homoeologous BAC sequences (AC189298.1 x EU642504.1; AC155341.2 x AC240089.1; AC155344.1 x AC240081.1 and CU984545.1 x EU579455.1), with one additional pair, were plotted against each other to identify novel LINEs, SINEs and several other DNA and novel TEs (see Conclusion; Figure 10.1-10.3).

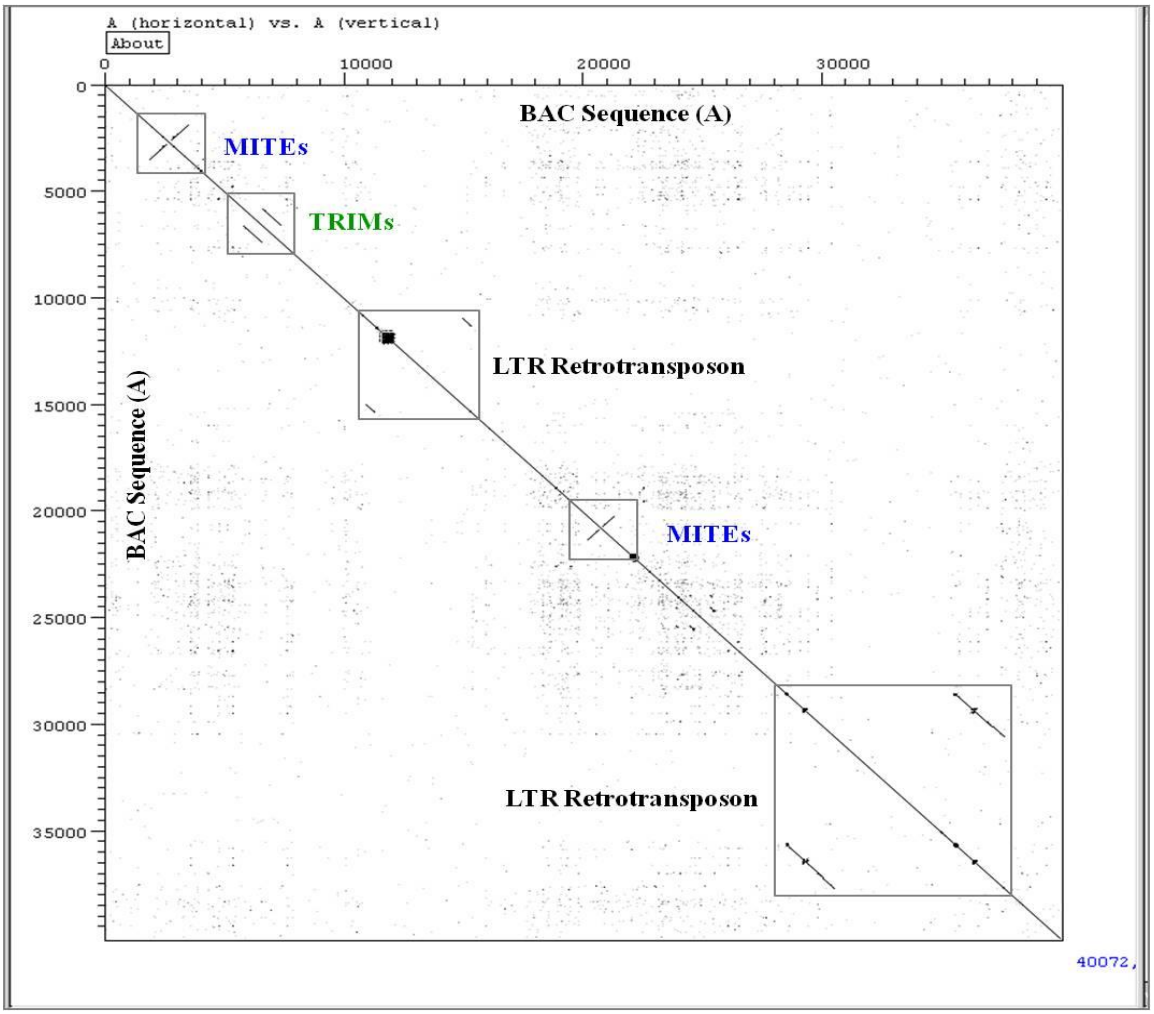


Figure 2.2: Dot plot identification of LTR retrotransposons (Copia, Gypsy, LARDs, TRIM) and MITEs. The (synthetic) BAC sequence is plotted against itself to show a complete line of homology running from one corner to diagonally opposite corner. The parallel lines across the central diagonal lines are indicating the LTRs. The inverted cross lines represent the large TIRs of Mutator-like MITEs. The elements are considered LTR retrotransposons or MITEs, if they possess the TSDs and other structural features.

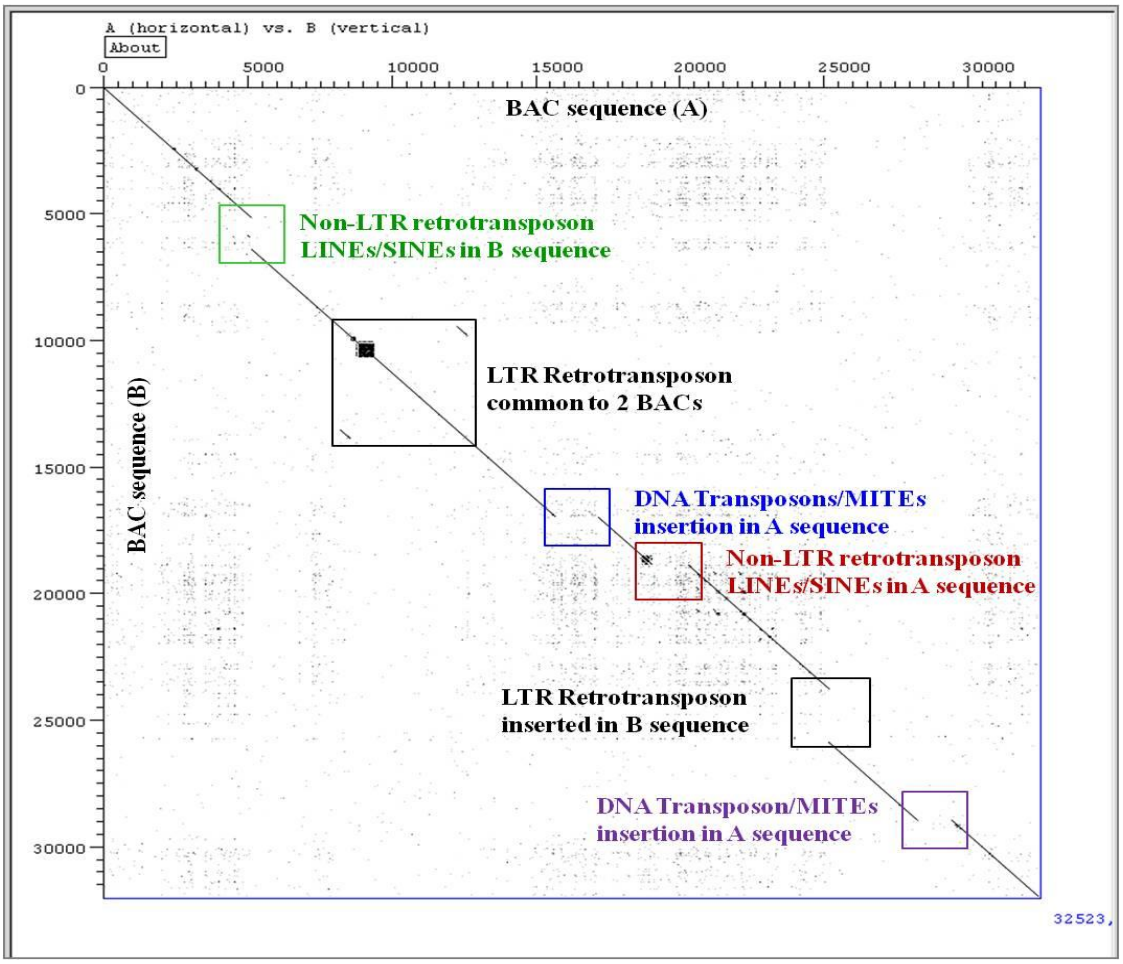


Figure 2.3: Dot plot identification of DNA transposons and MITEs (Stowaway, Tourist). The two (synthetic) homoeologous BACs sequences (A x B) were plotted against each other with a line of homology running from one corner to diagonally opposite corner with gap-insertion pairs. The parallel lines across the central diagonal lines indicate LTRs of retrotransposons. The gaps in the line indicate recent activity of transposon insertion in one sequence.

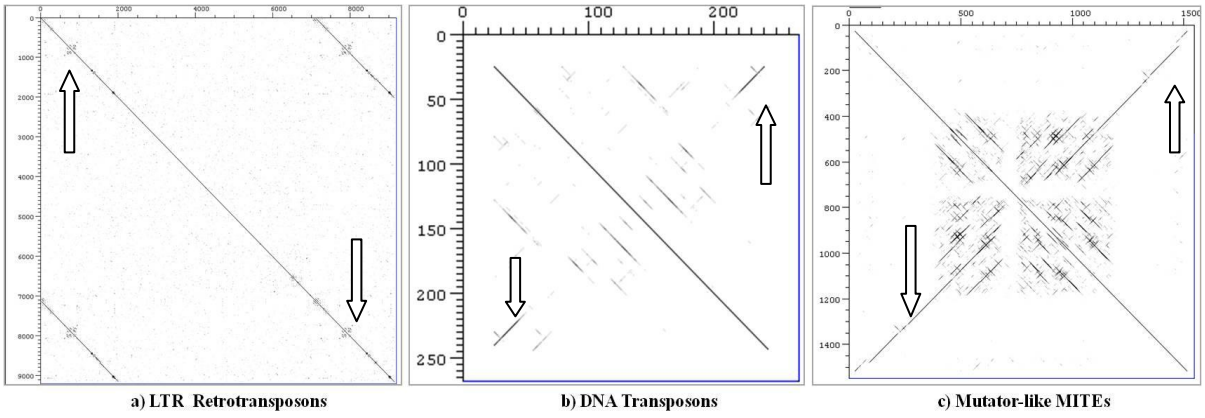


Figure 2.4: Dot plot graphs indicating characteristic features (arrows) of three different types of TEs. a) LTR retrotransposon with small parallel lines indicating LTRs at diagonally opposite corners. b) DNA transposon with TIRs at corners c) MITEs with long TIRs starting from corners to the central diagonal lines.

Table 2.4: List of 90 *Brassica* Bacterial Artificial Chromosomes (BACs) used for the identification of various TEs.

Sr.No.	Species	BAC Accessions	Size	Sr.No.	Species	BAC Accessions	Size
1	*B. oleracea*	AC122543.1	101563	46	*B. rapa*	AC189233.2	128372
2	*B. oleracea*	AC149635.1	96718	47	*B. rapa*	AC189237.1	105067
3	*B. oleracea*	AC152123.1	82329	48	*B. rapa*	AC189263.2	135260
4	*B. oleracea*	AC183492.1	236640	49	*B. rapa*	AC189298.1	133068
5	*B. oleracea*	AC183493.1	284024	50	*B. rapa*	AC189300.2	101741
6	*B. oleracea*	AC183494.1	285752	51	*B. rapa*	AC189364.2	95384
7	*B. oleracea*	AC183495.1	356505	52	*B. rapa*	AC189375.2	151784
8	*B. oleracea*	AC183496.1	385314	53	*B. rapa*	AC189415.2	126831
9	*B. oleracea*	AC183498.1	353037	54	*B. rapa*	AC189430.2	158169
10	*B. oleracea*	AC189656.2	106028	55	*B. rapa*	AC189446.2	126053
11	*B. oleracea*	AC240078.1	86917	56	*B. rapa*	AC189458.2	106248
12	*B. oleracea*	AC240079.1	6684	57	*B. rapa*	AC189472.2	153817
13	*B. oleracea*	AC240080.1	84518	58	*B. rapa*	AC189475.2	159384
14	*B. oleracea*	AC240081.1	108570	59	*B. rapa*	AC189496.2	130819
15	*B. oleracea*	AC240082.1	85043	60	*B. rapa*	AC189529.2	145402
16	*B. oleracea*	AC240083.1	108105	61	*B. rapa*	AC189592.2	133598
17	*B. oleracea*	AC240084.1	114384	62	*B. rapa*	AC232508.1	135661
18	*B. oleracea*	AC240085.1	94292	63	*B. rapa*	AC232512.1	140229
19	*B. oleracea*	AC240087.1	104293	64	*B. rapa*	AC232514.1	147357
20	*B. oleracea*	AC240088.1	84128	65	*B. rapa*	AC232592.1	123616
21	*B. oleracea*	AC240089.1	96237	66	*B. rapa*	AC234770.2	123583
22	*B. oleracea*	AC240090.1	117741	67	*B. rapa*	AC237303.1	110934
23	*B. oleracea*	AC240091.1	77461	68	*B. rapa*	AC237304.1	117696
24	*B. oleracea*	AC240092.1	87435	69	*B. rapa*	AC241035.1	103153
25	*B. oleracea*	AC240093.1	85407	70	*B. rapa*	AC241108.1	86592
26	*B. oleracea*	AC240094.1	96771	71	*B. rapa*	AC241138.1	149767
27	*B. oleracea*	EU568372.1	74376	72	*B. rapa*	AC241191.1	104793
28	*B. oleracea*	EU579454.1	92449	73	*B. rapa*	AC241194.1	119416
29	*B. oleracea*	EU579455.1	104071	74	*B. rapa*	AC241195.1	60300
30	*B. oleracea*	EU581950.1	71205	75	*B. rapa*	AC241196.1	103665
31	*B. oleracea*	EU642504.1	109794	76	*B. rapa*	AC241197.1	177500
32	*B. oleracea*	EU642505.1	86024	77	*B. rapa*	AC241198.1	128488
33	*B. oleracea*	EU642506.1	39495	78	*B. rapa*	AC241199.1	111007
34	*B. rapa*	AC155337.1	125390	79	*B. rapa*	AC241200.1	124586
35	*B. rapa*	AC155338.1	137697	80	*B. rapa*	AC241201.1	134414
36	*B. rapa*	AC155340.2	143518	81	*B. rapa*	CU695254.1	141917
37	*B. rapa*	AC155341.2	106476	82	*B. rapa*	CU695282.1	153759
38	*B. rapa*	AC155342.2	151550	83	*B. rapa*	CU914557.1	120113
39	*B. rapa*	AC166739.1	16947	84	*B. rapa*	CU984545.1	137597
40	*B. rapa*	AC166740.1	100288	85	*B. rapa*	FP340380.1	107018
41	*B. rapa*	AC166741.1	132099	86	*B. rapa*	FP340381.1	146501
42	*B. rapa*	AC189183.2	99407	87	*B. rapa*	FP340382.1	104976
43	*B. rapa*	AC189218.2	128973	88	*B. rapa*	FP340534.1	131483
44	*B. rapa*	AC189222.1	163034	89	*B. rapa*	FP340535.1	128773
45	*B. rapa*	AC189225.2	116064	90	*B. rapa*	FP565592.1	141962

Table 2.5: List of 46 *Musa* Bacterial Artificial Chromosomes (BACs) used for the identification of various TEs.

Sr. No.	Species	BAC Accession	BAC size	Sr. No.	Species	BAC Accession	BAC size
1	*M. acuminata*	AC226031.1	61335	24	*M. acuminata*	AC186752.1	80932
2	*M. acuminata*	AC226032.1	91293	25	*M. acuminata*	AC186753.1	54106
3	*M. acuminata*	AC226033.1	92078	26	*M. balbisiana*	AC186754.1	142973
4	*M. acuminata*	AC226034.1	76327	27	*M. acuminata*	AC186955.1	102232
5	*M. acuminata*	AC226035.1	104637	28	*M. balbisiana*	AC226054.1	124336
6	*M. acuminata*	AC226036.1	37847	29	*M. balbisiana*	AC226055.1	106121
7	*M. acuminata*	AC226037.1	71163	30	*M. acuminata*	AC226196.1	110853
8	*M. acuminata*	AC226038.1	124825	31	*M. balbisiana*	AP009325.2	133047
9	*M. acuminata*	AC226039.1	86265	32	*M. balbisiana*	AP009334.1	131526
10	*M. acuminata*	AC226040.1	108948	33	M. acuminata	AY484588.1	73268
11	*M. acuminata*	AC226041.1	58704	34	*M. balbisiana*	FN396604.1	137100
12	*M. acuminata*	AC226042.1	73023	35	*M. balbisiana*	FN396605.1	141036
13	*M. acuminata*	AC226043.1	95303	36	*M. acuminata*	AC186747.2	141025
14	*M. acuminata*	AC226044.1	74174	37	*M. acuminata*	AC186748.1	113519
15	*M. acuminata*	AC226045.1	95242	38	*M. acuminata*	AC186746.1	105019
16	*M. acuminata*	AC226046.1	180124	39	*M. acuminata*	AC186749.1	29567
17	*M. acuminata*	AC226047.1	87766	40	*M. acuminata*	AC186751.1	96443
18	*M. acuminata*	AC226048.1	134662	41	*M. acuminata*	AC186750.2	148170
19	*M. acuminata*	AC226049.1	92303	42	*M. acuminata*	AY484589.1	82723
20	*M. acuminata*	AC226050.1	177729	43	*M. balbisiana*	AC186755.1	154246
21	*M. balbisiana*	AC226051.1	152711	44	*M. balbisiana*	AP009326.1	119244
22	*M. balbisiana*	AC226052.1	198395	45	*M. balbisiana*	FN396606.1	253366
23	*M. balbisiana*	AC226053.1	135110	46	*M. acuminata*	AC186954.2	144091

2.10.2 Computational analysis and data mining for retrotransposons

The intact or full length elements identified by dot plot analysis are named as reference elements as they are full length elements belonging to different superfamilies and families of retrotransposons. The reference elements were further used to conduct BLASTN searches against the *Brassica* or *Musa* Nucleotide Collection (nr/nt) database using 'somewhat similar sequences' option in NCBI. In the database, the searches for LTR retrotransposons were performed in several steps to identify the intact, truncated, partial elements, solo LTRs and remnants (Figure 2.5). First the LTRs were used as a query to find the solo LTRs, which were counted by any single copy in a BAC sequence or multiple copies without any internal region. The intact elements were counted by having two complete LTRs with internal region >2 kb. In the second step, the complete elements

were used as query to find the full length copies, truncated elements, partial or deleted elements and remnants, which were defined with small modifications according to the recommendations of Ma *et al*., 2004. An intact element is one that is terminated by well characterized TSDs and LTRs, with an internal region encoding one or more protein domains from *gag-pol* genes, and exhibiting the identified PBS and PPT sites. Solo LTR refers to an LTR with TSD, or LTRs truncated with small deletions exhibiting >80% query coverage and homology. Truncated elements are defined as elements having deletions at 5′ or 3′ ends of LTRs. These include elements with one partially or completely deleted LTR, elements with both LTRs partially deleted and the elements with one partially and one completely deleted LTRs. Partial sequences are deletion derivatives showing >40-80% query coverage, with or without LTRs and one or more conserved domains (PBS, AP, RT, RH, INT, PPT) in them. The term remnants describe all the small fragments showing 1-40% query coverage with strong or weak identity to the retrotransposon sequences. The remnants sometimes include the deleted LTRs, any intact domain and internal region from an element (Figure 2.5).

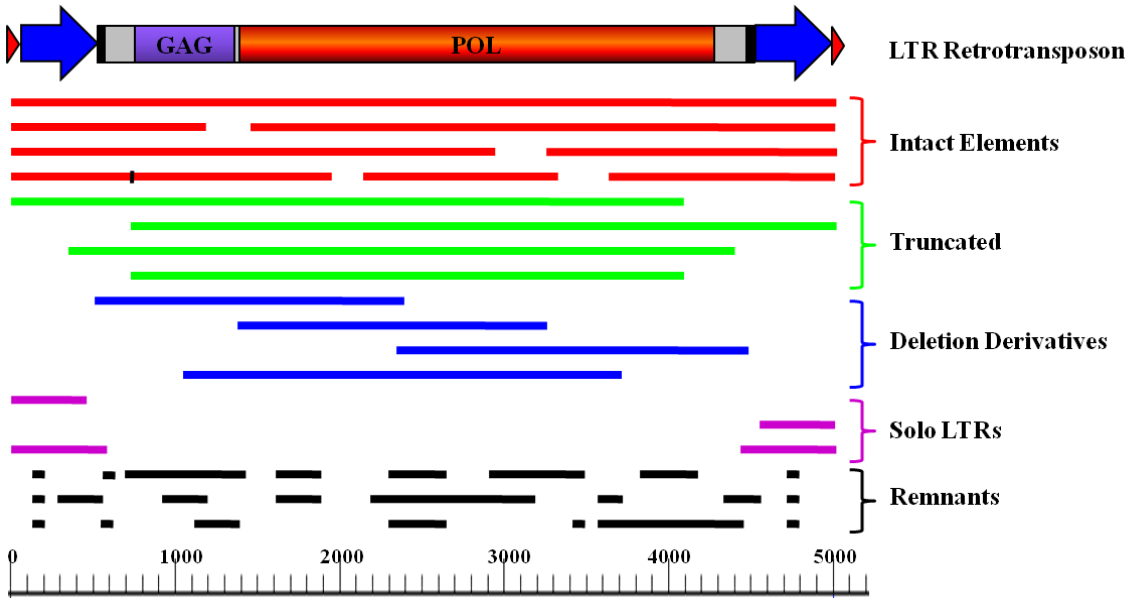


Figure 2.5: Homology matches of an element. Red lines represent intact elements, green lines truncated elements, blue lines indicate deletion derivatives, pink solo LTRs and black represents the remnants. Only complete elements were used to estimate the copy numbers in whole *Brassica*/*Musa* genomes.

The novel non-LTR retrotransposons (LINES, SINEs) identified by dot plot analysis were run against the *Brassica* Nucleotide Collection database in NCBI. The sequences showing >75% of the query coverage with >80% identity in their entire lengths were retrieved and

analysed. The number of the TSDs and the poly(A) tail at 3′ ends were counted manually. Where necessary the TSDs and TIRs were identified by the use of the online dot plot program Dotlet (http://myhits.isb-sib.ch/cgi-bin/dotlet) (Junier and Pagni, 2000).

2.10.3 Characterization, classification and naming of retrotransposons

The online ORF finder program (http://www.ncbi.nlm.nih.gov/projects/gorf/) was used to detect any ORF structure from the identified elements. The Repbase database (http://www.girinst.org/repbase/index.html) (Jurka *et al.*, 2005), Repeat masker of Censor software (http://www.girinst.org/censor/index.php) implemented in Genetic Information Research Institute (GIRI) and Gypsy database (http://gydb.org/index.php/Main_Page) (Llorens *et al.*, 2008; Llorens *et al.*, 2011) were used to characterize the retrotransposons on the basis of homology to the known elements. Elements that failed to be characterized by the above searches against TE databases were characterized by visual inspection on the basis of their hallmark motifs such as TSDs, LTRs, PBS, PPT and organization of their *gag-pol* encoding proteins. The retrotransposons are classified as Copia, if they display *pol* gene as 5 -INT-RT-RH-3 , Gypsy as 5 -RT-RH-INT-3 , LARDs if they exhibit large non-coding internal regions and TRIMs, if they only have LTRs and small internal non-coding region (<1 kb). The families were defined by using the same criteria adopted by other workers for the characterization of families of Copia, Gypsy, Pararetroviruses and Bel-Pao superfamilies. The sequences showing >85% identity at their nucleotide level over at least 1000 bp in their coding regions were considered belonging to the same family. If the homology is >95%, they were considered as copies of single element (Wicker *et al.*, 2007; Minervini *et al.*, 2009).

A novel TE family is declared, when no homology of the family was found with any known LTR retrotransposon, a full set of intact elements were evident with LTRs, internal protein domains for its transposition and the last discriminator was the presence of strong hits to at least another member excluding the reference query (Wang and Liu, 2008). The identified elements were classified into respective superfamilies and families by the recommendations of Wicker *et al*., 2007. The names to the novel elements are given on the recommendations of Capy, 2005. The names are given as *GsXXXN*, where '*G*' represent genus, small letter 's' represent species names, *XXX* indicate first 3 letters of retrotransposons superfamily and '*N'* indicate the number. Thus in *BrCOP1*; *Br* indicate

Brassica rapa, *COP* represent Copia and 1 indicate the number of identified element. *MaGYP1* represent *Musa acuminata* Gypsy element 1 and *MaLAR1* indicate 1st *Musa acuminata* LARDs-like element. In all cases, the names of the elements and their respective families are written in italics.

The LINEs were characterized on the basis of displaying LINE specific EN-RT domains and poly(A) tail at C-terminus. For LINEs homologous copies showing >75% identity at their nucleotide level in their coding regions were collected and considered as members of the same family. For SINEs (small non-coding sequences), the sequences were considered belonging to the same family, if the homology is >80% in their entire lengths. The retroposons (LINEs, SINEs) were named on same pattern, as LTR retrotransposons were named. Thus *BrLINE1-1* is indicating the first member of *Brassica rapa* LINE family 1. Similarly, *BoSINE1-1* represents the first member of *Brassica oleracea* SINE family 1.

2.10.4 Analysis of structural domains in retrotransposons

For the identification of conserved domains in intact elements, the nucleotide sequences were investigated in 'conserved domain database' (CDD) of NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) with default parameters. The *gag-pol* gene encoding proteins (GAG, AP, RT, RH, INT) and additional domains in each element were detected by CDD. The primer binding site (PBS) and polypurine tract (PPT) were detected in the LTR_FINDER by using the parameter 'Predict PBS by using which tRNA database' against *Arabidopsis thaliana* tRNA database in case of *Brassica* LTR retrotransposons, while against *Zea mays* and *Oryza sativa* tRNA database in case of *Musa* LTR retrotransposons. In the first step all elements were screened against the *Zea mays* tRNA database to detect PBS and PPT. If no hits were found for PBS or PPT, than the elements were blast against *Oryza sativa* tRNA database to detect PBS and PPT. The sequence and positions of PBS and PPT is marked and type of tRNA was also detected.

The program ORF finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) was used to identify the open reading frames in the LINE elements. The proteins domains and their patterns were detected by using the 'conserved domain database' (CDD) at NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). For the comparison of tRNA head region of the SINEs, the tRNA sequences were retrieved from *Arabidopsis* Genomic

tRNA Database (http://gtrnadb.ucsc.edu/Athal) (Chan and Lowe, 2009). Frequency plots indicating the insertional preference of SINEs families were generated in WebLogo (http://weblogo.berkeley.edu/logo.cgi).

2.10.5 Multiple sequence alignment and phylogenetic analysis

The reverse transcriptase (RT) sequences from 75 elements belonging to three superfamilies (Copia, Gypsy and Pararetroviruses) of known LTR retrotransposons were collected from Gypsy database (http://gydb.org/index.php/Main_Page) (Llorens *et al.*, 2011), listed in Table 2.6. The conserved (~200 aa) RT regions from identified elements from *Brassica/Musa* were aligned with known elements in CLUSTALW multiple alignments implemented in BioEdit (Hall, 1999). The sequences after alignment were visually inspected and edited manually, if needed. Small insertions and deletions were removed and single deletions were filled by simple majority rule. Frameshifts were introduced to bring the sequences in the same frame. All RT regions were included in alignment, even if they have stop codons or frameshift mutations. In all the cases, the orientations of elements were converted to 5′-3′ for alignment and phylogenetic analysis. The phylogenetic analysis was performed by constructing the tree by the Neighbour-Joining method with 1000 bootstrap replicates. The Tamura-Nei (1973) and Jukes-Cantor models were used to calculate genetic distance for nucleotides and amino acid sequences respectively. The trees were generated in Geneious Pro 5.5.6 program (Drummond *et al*., 2011).

2.10.6 Estimation of copy numbers

The numbers of strong hits against the reference queries with >75% query coverage and identity were extrapolated after getting output from BLASTN searches. Only intact elements were counted, while partial elements and remnants were not included. The following formula was used to estimate the copy number of intact retrotransposons: copy no. = no. in database x genome size/database size as used to estimate TEs (Tu, 2001). The percentage of each TE superfamily in whole genome is calculated as: Estimated copies x average size = N, percentage = N/total of all N values x 100. ‘N’ is the total size in bp of each superfamily identified here.

Table 2.6: List of 75 LTR retrotransposons (Copia, Gypsy and Pararetrovirus (PRV) superfamilies) collected from Gypsy Database (Llorens *et al.*, 2011) for various phylogenetic studies. NG: Not given.

Sr. No.	Super-family	Elements Name	Name in detail	Element Size	Identified from
1	Copia	*SIRE1-4*	NG	9.8 Kb	*Glycine max*
2	Copia	*Opie-2*	NG	11.7 Kb	*Zea mays*
3	Copia	*Endovir1-1*	NG	9.1 Kb	*Arabidopsis thaliana*
4	Copia	*ToRTL1*	NG	9.7 Kb	*Lycopersicon esculentum*
5	Copia	*TSI-9*	NG	9.3 Kb	*Setaria italica*
6	Copia	*Araco*	NG	4.9 Kb	*Arabidopsis thaliana*
7	Copia	*Oryco1-1*	NG	4.9 Kb	*Oryza sativa* ssp. *japonica*
8	Copia	*Vitico1-1*	NG	4.6 Kb	*Vitis vinifera*
9	Copia	*Poco*	NG	4.3 Kb	*Populus trichocarpa*
10	Copia	*Oryco1-2*	NG	4.9 Kb	*Oryza sativa* ssp *japonica*
11	Copia	*Melmoth*	NG	4.8 Kb	*Brassica* spp.
12	Copia	*Vitico1-2*	NG	4.8 Kb	*Vitis vinifera*
13	Copia	*Retrofit*	NG	4.9 Kb	*Oryza longistaminata*
14	Copia	*Koala*	NG	5.0 Kb	*Oryza australiensis*
15	Copia	*Hopscotch*	NG	4.8 Kb	*Zea mays*
16	Copia	*Tto1*	NG	5.3 Kb	*Nicotiana tabacum*
17	Copia	*Batata*	NG	4.2 Kb	*Ipomoea batatas*
18	Copia	*Sto-4*	NG	7.2 Kb	*Zea mays*
19	Copia	*Fourf*	NG	7.0 Kb	*Zea mays*
20	Copia	*Tork4*	NG	4.9 Kb	*Lycopersicon esculentum*
21	Copia	*RTvr2*	NG	7.8 Kb	*Vigna radiata*
22	Copia	*V12*	NG	5.4 Kb	*Vitis vinifera*
23	Copia	*Tnt-1*	NG	5.3 Kb	*Nicotiana tabacum*
24	Copia	*Copia*	NG	5.2 Kb	*Drosophila* spp.
25	Copia	*Ty1B*	NG	6.0 Kb	*Saccharomyces cerevisiae*
26	Copia	*Ty2*	NG	6.0 Kb	*Saccharomyces cerevisiae*
27	Copia	*Ty4*	NG	6.2 Kb	*Saccharomyces cerevisiae*
28	Copia	*Osser*	NG	4.9 Kb	*Volvox carteri*
29	Gypsy	*Ty3-1*	NG	5.5 Kb	*Saccharomyces cerevisiae*
30	Gypsy	*Del*	NG	9.3 Kb	*Lilium henryi*
31	Gypsy	*Galadriel*	NG	6.2 Kb	*Lycopersicon esculentum*
32	Gypsy	*Tntom1*	NG	6.0 Kb	*Nicotiana tomentosiformis*
33	Gypsy	*Cereba*	NG	11.6 Kb	*Hordeum vulgare*
34	Gypsy	*CRM*	NG	7.6 Kb	*Zea Mays*
35	Gypsy	*Beetle1*	NG	6.7 Kb	*Beta vulgaris*
36	Gypsy	*Tma*	NG	7.3 Kb	*Arabidopsis thaliana*
37	Gypsy	*Reina*	NG	5.4 Kb	*Zea Mays*
38	Gypsy	*Gloin*	NG	5.4 Kb	*Arabidopsis thaliana*
39	Gypsy	*Legolas*	NG	7.5 Kb	*Arabidopsis thaliana*
40	Gypsy	*Monkey*	NG	6.3 Kb	*Musa acuminata*
41	Gypsy	*Ifg7*	NG	5.9 Kb	*Pinus radiata*
42	Gypsy	*Peabody*	NG	7.9 Kb	*Pisum sativum*
43	Gypsy	*Retrosat-2*	NG	12.7 Kb	*Oryza sativa*
44	Gypsy	*Athila4-1*	NG	14.0 Kb	*Arabidopsis thaliana*
45	Gypsy	*Cyclops-2*	NG	12.2 Kb	*Pisum sativum*
46	Gypsy	*Diaspora*	NG	11.7 Kb	*Glycine max*
47	Gypsy	*Ogre*	NG	22.7 Kb	*Pisum sativum*
48	Gypsy	*Bagy-1*	NG	14.4 Kb	*Hordeum vulgare*
49	Gypsy	*RIRE2*	NG	11.2 Kb	*Oryza sativa*
50	Gypsy	*RetroSor1*	NG	13.4 Kb	*Sorghum bicolor*
51	Gypsy	*Cinful-1*	NG	8.6 Kb	*Zea mays*
52	Gypsy	*Grande1-4*	NG	13.8 Kb	*Zea diploperennis*
53	Gypsy	*Tat4-1*	NG	11.9 Kb	*Arabidopsis thaliana*
54	Gypsy	*Tft2*	NG	13.2 Kb	*Arabidopsis thaliana*
55	Gypsy	*Gypsy*	NG	7.4 Kb	*Drosophila melanogaster*

Table 2.6: *Continued*

Sr. No.	Super-family	Elements Name	Name in detail	Element Size	Identified from
56	PRV	*CaMV*	*Cauliflower mosaic virus*	8.0 Kb	*Brassicaceae*
57	PRV	*CERV*	*Carnation etched ring virus*	7.9 Kb	*Dianthus caryophyllus*
58	PRV	*MiMV*	*Mirabilis mosaic virus*	7.9 Kb	*Mirabilis* spp.
59	PRV	*FMV*	*Figwort mosaic virus*	7.7 Kb	*Scrophularia californica*
60	PRV	*CSVMV*	*Cassava vein mosaic virus*	8.2 Kb	*Manihot* spp.
61	PRV	*TVCV*	*Tobacco vein clearing virus*	7.8 Kb	*Nicotiana edwardsonii*
62	PRV	*BSVAV*	*Banana streak virus*	7.8 Kb	*Musa* spp.
63	PRV	*BSGFV*	*Banana streak virus*	7.3 Kb	*Musa* spp.
64	PRV	*KTSV*	*Kalanchoë top-spotting virus*	7.6 Kb	*Kalanchoë blossfeldiana*
65	PRV	*BSMyV*	*Banana streak virus*	7.6 Kb	*Musa* spp.
66	PRV	*PVCV*	*Petunia vein clearing virus*	7.2 Kb	*Petunia x hybrida* cv
67	PRV	*RTBV*	*Rice tungro bacilliform virus*	8.0 Kb	*Oryza sativa*
68	PRV	*CSSV*	*Cacao swollen shoot virus*	7.2 Kb	*Theobroma cacao*
69	PRV	*CYMV*	*Citrus yellow mosaic virus*	7.5 Kb	*Citrus* spp.
70	PRV	*DBV*	*Dioscorea bacilliform virus*	7.3 Kb	*Dioscorea* spp.
71	PRV	*TaBV*	*Taro bacilliform virus*	variable	*Colocasia esculenta*
72	PRV	*BSOLV*	*Banana streak virus*	7.4 Kb	*Musa* spp.
73	PRV	*DrMV*	*Dracaena mottle virus*	variable	*Dracaena sanderiana*
74	PRV	*BsCVBV*	*Bougainvillea spectabilis chlorotic vein-banding virus*	8.8 Kb	*Bougainvillea spectabilis*
75	PRV	*ComYMV*	*Commelina yellow mottle virus*	7.5 Kb	*Commelina diffusa*

2.11 Material and Methods for DNA transposons and MITEs

2.11.1 Identification and genome wide analysis of DNA transposons and MITEs

Homoeologous BAC clone pairs from either *Brassica/Musa* species were compared against each other in dot plot analyses. The central diagonal line running from one corner of the dot plot to other represented the homology between the two sequences. The gaps in the homologous line indicated the insertions-deletion pairs, which were characterized (Figure 2.3; see conclusion figures 10.1-10.3). The TSDs and TIRs flanking the insertions were inspected manually and if present, the elements were characterized as mobile transposons and further categorized to define the respective superfamily. The identified transposons were used to query BLASTN searches against *Brassica/Musa* Nucleotide Collection database (nr/nt) available in NCBI by using the 'somewhat similar sequences' parameter. The resultant hits covering >60% of entire query sequences were collected and characterized. The TSDs and TIRs of the elements were defined by manual inspections and their positions in BAC clone sequences were defined by BLAST hits. Protein domain structures and organization of the elements were studied in 'conserved domain database' (CDD: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) available in NCBI.

For identification of MITEs, two different approaches were followed. The small Mariner-like (Stowaway) and Harbinger-like (Tourist) MITEs were identified by the comparison of homoeologous BAC sequences, by the same approach as used for the identification of autonomous and non-autonomous DNA transposons. The other approach was used for the identification of Mutator-like MITEs having long TIRs, where BAC sequence is plotted against itself. The perpendicular line nearly intersecting the diagonal line showed the TIRs (Figure 2.4c), which were further investigated for MITE derived superfamily identification. In any of these approaches, the collected sequences were inspected manually for the presence of TSDs and TIRs and only those sequences were considered as MITEs, which showed well defined TSDs and TIRs. The homologous copies were collected from nucleotide collection database of respective specie (http://www.ncbi.nlm.nih.gov) using BLASTN program (Altschul *et al.*, 1997). Only those sequences were collected which showed high homology with query sequences, having size <2.5 kb and showing precise boundaries with flanking TSDs. The dot plot graphs of the MITEs to represents the TIRs in the sequences were drawn in Dotter (Sonnhammer and Durbin, 1995) or Dotlet software (http://myhits.isb-sib.ch/cgi-bin/dotlet) (Junier and Pagni, 2000) by plotting the sequence against itself.

2.11.2 Sequence analysis and manipulation

The protein domain organizations of autonomous DNA transposons were performed by blast against the known protein database in 'conserved domains database' (CDD; http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). The GC and AT contents of the DNA transposons were calculated using the online program 'GC Calculator' (http://www.genomicsplace.com/gc_calc.html). The sequence logos of TSDs and TIRs were generated online with WebLogo (http://weblogo.berkeley.edu/logo.cgi). For autonomous DNA transposons, the most conserved transposase regions around DDD/E triad ranging in sizes from 630-680 bp (210-228 aa) were collected. The known DNA transposons (hAT, CACTA, Harbinger, Mutator) were collected from Repbase database (Jurka *et al.*, 2005) (http://www.girinst.org/repbase/update/browse.php) and transposase regions were aligned with the transposase regions of our identified DNA transposons. They were trimmed to gain the equal lengths and where necessary frameshifts were introduced to maintain the aligned reading frames. Multiple alignments were generated

with CLUSTALW available in the BioEdit program and small insertions were removed without altering the frame of the elements.

2.11.3 Phylogenetic analysis for autonomous and non-autonomous DNA transposons

The phylogenetic analysis of autonomous DNA transposons was performed by using the Neighbour-Joining (NJ) methods with Janker-Cantor genetic distance model. The highly conserved catalytic triad DDD/E regions (180-210 aa) from the transposase genes were aligned to generate the CACTA and Harbingers transposons phylogeny. The tree was constructed in the Geneious Pro 5.5.6 (Drummond *et al.*, 2011) with a bootstrap value of 1000 repetitions.

The complete sequences from non-autonomous transposons and MITEs were collected from dot plot and blast analysis. They were aligned in CLUSTALW implemented in BioEdit and manually edited. The phylogenetic analysis of various superfamilies of DNA transposons and MITEs were done separately. To generate the tree for non-autonomous hATs, 200 bp from 5´ TIRs were used. For MITEs phylogeny, the TIRs were used for the construction of tree. The tree is generated in Geneious program (Drummond *et al*., 2011) using Neighbour-Joining method with 1000 bootstraps replicates. The Genetic distance was calculated with Tamura-Nei (1973) model.

2.11.4 Copy number estimation of DNA transposons

The copy number estimation was performed in the same way, as done for the copy number estimation of retrotransposons. The numbers of strong hits against the reference queries with >60 query coverage and 70% identity were collected after getting output from BLASTN searches against *Brassica/Musa* Nucleotide Collection database available at NCBI. The following formula was used to estimate the copy number of DNA transposons: copy no. = no. in database x genome size/database size (Tu, 2001). The percentage of each DNA transposon superfamily in whole genome is calculated as: Estimated copies x average size = N, percentage = N/total of all N values x 100. Where ‘N’ is representing the total size (bp) of each superfamily.

2.12 Characterization and nomenclature of DNA transposons and MITEs

For the proper characterization of DNA transposons and MITEs, both homology and structural based approaches were used to characterize the elements. In the homology based approach, the sequences were blasted against the Repeat Masker of Repbase database (http://www.girinst.org/censor/index.php) and Plants Repeats database (http://plantrepeats.plantbiology.msu.edu/) (Ouyang and Buell, 2004) to detect homology of query sequences against any known element. If the element was not characterized on the basis of homology based searches, the structural features and hallmarks (TSDs, TIRs and internal coding regions) of transposons were studied. The numbers of TSDs and TIRs were found to be the best criteria to allocate DNA transposons and MITEs to their superfamily and family, from which they have derived.

The names to the DNA transposons and MITEs were given systematically as recommended by Capy (2005). The first letter indicates genus, second small letter indicate species, 3-4 letters represent superfamily, first digit represent individual insertion or a family and second digit represent its homologue such as *BoCACTA1*, where '*B*' represent genus, small letter (*o*) indicate specie and number after *CACTA* indicate the first identified CACTA in present study. For non-autonomous transposons after genus and species name, letter 'N' indicate non-autonomous element such as *MaN-hAT1*, where '*M*' stands for genus *Musa*, small letter (*a*) indicate *acuminata*, '*N*' indicate the non-autonomous *hAT* and digit '*1*' represent 1st investigated hAT insertion in this study. The MITEs are named in the similar way such as *BrSTOW1-1*, where '*B'* stands for genus *Brassica*, second small letter (*r/o*) represents species name (*rapa/oleracea*), 4 capital letters (STOW/TOUR) indicate the transposon superfamily, from which MITEs are derived, the first number after the superfamily name indicate the family and number followed by hyphen represents the number of the respective member of that family. For Mutator-like MITEs such as *BrMuMITE1-1*, '*Mu'* represents Mutator and in MITES with unknown superfamily such as *BrXMITE1-1*, X indicates unknown superfamily of MITEs, while rest pattern for naming is the same. The family names were given on the basis of first element identified or on the basis of highly abundant element from the respective family. Thus *BrSTOW1* is the family representing *BrSTOW1-1* to *BoSTOW1-5* elements. Similarly, *BrMuMITE5-1* and *BoMuMITE5-9* are treated in the same family *BrMuMITE5*. The names of the elements and their families were represented by italics.

CHAPTER 3
IDENTIFICATION AND CHARACTERIZATION OF NOVEL LTR RETROTRANSPOSON FAMILIES FROM *BRASSICA*

Summary

By using computational and molecular methods, 280 intact LTR retrotransposons were identified from *Brassica* by dot plot analysis and data mining. The Copia elements were dominant (206), followed by Gypsy (56), while non-autonomous retrotransposons LARDs (16) and TRIM (1) were much less. Around 1596 Copia, 540 Gypsy, 110 LARDs and 25 TRIM elements were estimated from *Brassica rapa* and 7540 Copia, 780 Gypsy, 760 LARDs with no TRIM from *Brassica oleracea* whole genomes. The results indicated that Copia outnumbered Gypsy by a ratio of 4:1. Several truncated or partial homologues of the elements were found dispersed in the genomes. PCR amplification based on conserved RT regions revealed their abundance and distribution among *Brassica* species and cultivars. The evolutionary relationship of *Brassica* TEs with other known elements clearly splits them into three main lineages; Copia, Gypsy and Pararetrovirus. *Brassica* elements clustured into 41 families, of which 35 are Copia and 6 are Gypsy. The analysis also confirmed that the majority of the families are novel, as no significant homology was observed with other known elements in other species. The detailed analysis of the reverse transcriptase region of *Brassica* and several other known LTR retrotransposons revealed few conserved regions among all elements investigated.

3.1 Introduction

Long terminal repeat retrotransposons (LTR retrotransposons) are characterized by 4 to 6 bp TSDs, 100-5000 bp LTRs, internal regions encoding *gag-pol* protein domains and PBS and PPT at 5′LTR and 3′LTR respectively. The LTRs are homologous and generally have conserved termini (5′-TG-----CA-3′). The LTRs carry the promoter elements, TATA box, polyadenylation signals and enhancers, which regulate the transposition mechanism of LTR retrotransposons. The *gag-pol* encodes protein domains necessary for transposition and integration mechanisms, while PBS and PPT act as minus and plus priming sites for RNA transcription (Kumar and Bennetzen, 1999; Wicker and Keller, 2007; Wicker *et al.*, 2007; Vukich *et al.*, 2009).

Copia and Gypsy are two major superfamilies of LTR retrotransposons dispersed in plants that differ in order of protein domains encoded by *pol* gene. The canonical Ty1/copia exhibit TSDs, LTRs, display PBS and PPT and has internal *gag-pol* genes which encode the proteins as 5´-GAG-INT-RT-RH-3´ (Flavell *et al.*, 1992b; Flavell *et al.*, 1998; Kumar and Bennetzen, 1999; Hansen and Heslop-Harrison, 2004; Wicker *et al.*, 2007). Few elements encode additional domains of known or unknown nature in their *pol* gene. Ty3/gypsy elements constitute a superfamily of LTR retrotransposons. They display 5 bp TSDs, LTRs and internal region encoding *gag-pol* protein domains as 5´-GAG-RT-RH-INT-3´, or have some additional domains. On the basis of presence or absence of chromodomain, they are further divided into chromodomain bearing Gypsy and non-chromodomain Gypsy elements. The chromodomain bearing Gypsy are common in several plants (Novikova *et al.*, 2008; Novikova, 2009). The gypsy elements are diverse and abundant group of retrotransposons dispersed in several plants.

LARDs are non-autonomous LTR retrotransposons, which mobilized in *trans* by using the proteins from the autonomous elements. The internal regions of these elements contain conserved non-coding DNA segments that may provide the important secondary structure to mRNA, although it is not clear how these non-coding sequences work in the life cycle of the elements (Kalendar *et al.*, 2004). Due to lack of internal *gag-pol* coding genes and large sizes (5.5 to 8.5 kb), the non-autonomous *Dasheng* elements from maize genome were named as 'Large Retrotransposons Derivatives' (LARDs). The non-autonomous *Dasheng* and *Zeon-1* elements from maize genome are represented by around 1000 copies each (Hu *et al.*, 1995; Jiang *et al.*, 2002). Another group of small sized elements with flanking LTRs are described from several plants and are named 'Terminal-Repeat Retrotransposons in Miniature' (TRIM). TRIM are small in sizes, have TSDs, LTRs, short non-coding internal regions, PBS and PPT motifs. They are studied in many monocot and dicot plant families including *Gramineae*, *Brassicaceae*, *Solanaceae*, and *Fabaceae* (Witte *et al.*, 2001), apple (Antonius-Klemola *et al.*, 2006) and *Brassica* (Yang *et al.*, 2007).

With the aim of studying the identification of novel retrotransposons, their genetic diversity, distribution, activity and evolutionary impacts on *Brassica* genomes, bioinformatic and molecular approaches were used to characterize the mobile elements in the genome.

3.2 Results

3.2.1 Strategy for mining and characterizing LTR retrotransposons in *Brassica*

Ninety *Brassica* BACs (Table 2.4) were screened by dot plot analysis to identify repetitive elements by plotting each BAC against itself. An unbroken diagonal line crossing from one corner to the other shows the homology of the sequence. Two small lines drawn parallel on both sides of the central diagonal line indicates 5′ and 3′ LTRs (Figure 3.1) and their associated TSDs were identified by visual inspection. The termini of LTRs were identified, with most having 5′-TG….CA-3′ termini observed in 95% elements investigated in present work. Total sizes of full length elements and the size of 5′/3′ LTRs were counted and tabulated (Table 3.1). The elements were characterized by structural features and homology basis with known elements. For homology based characterization, the elements were BLAST searched against transposon databases such as Repbase, Gypsy database and Plant Repeat database (TIGR) to characterize them on the basis of homology to the known elements. The Repeat masker of CENSOR software implemented in 'Genetic Information Research Institute' (GIRI) was used to see the occupancy of elements with known elements.

Very few elements were characterized by homology searches against TE databases and were characterized on the basis of their TSDs, LTRs, organization of PBS and PPT and protein domain organizations. The elements were defined by the same criteria as adopted by several other workers for the characterization of Copia, Gypsy, Pararetroviruses and LARDs-like LTR retrotransposons. The sequences showing >85% identity at their nucleotide level over at least 85% in their coding regions were considered belonging to the same family. If the homology was >95%, they were considered as copies of single element (Wicker *et al*., 2007; Minervini *et al*., 2009). A novel family was defined when no homology was identified with known elements, there were complete LTRs, internal protein domains and there was homology to at least another (uncharacterized) sequence (Wang and Liu, 2008). For the nomenclature of identified LTR retrotransposons and new families, the recommendations of Capy, (2005) were followed. Thus *BoCOP1* indicates Copia 1 member from *Brassica oleracea*, *BrGYP5* indicates the 5th Gypsy element from *Brassica rapa* and *BoLAR3* represents 3rd *LARD* element from *Brassica oleracea*.

Seventy seven full length (intact) retroelements (Table 3.1) from *Brassica rapa* and *Brassica oleracea* BAC clones were identified by dot plot analysis belonging to Copia (55), Gypsy (15), LARDs (6) and TRIM (1). These reference elements were used to conduct BLAST searches against the *Brassica* Nucleotide Collection (nr/nt) database in NCBI before January, 2012. In the database, the searches were performed in several steps to identify the intact, truncated, partial elements, solo LTRs and remnants. First the LTRs were used as a query to find the solo LTRs, which were counted by any single copy in a BAC or multiple copies without any internal region. The intact elements were counted by having two complete LTRs with internal region >2 kb. In the second step, the complete elements were used as a query to find the full length copies, truncated elements, partial or deleted elements and remnants, which were defined according to the recommendations of *Ma et al*., (2004). An intact element is one that is terminated by well characterized TSDs and LTRs, with an internal region encoding one or complete protein domains from *gag-pol* genes, and exhibiting identified PBS and PPT sites. Solo LTR refers to an LTR with TSD, or LTRs truncated with small deletions exhibiting >80% query coverage and homology. Truncated elements are defined as elements having deletions at 5′ or 3′ ends or both ends of LTRs. Partial sequences are the deletion derivatives showing 40-70% query coverage, with or without LTRs and one or more conserved domains. The term remnants describe the small fragments showing 1-40% query coverage with strong or weak identity to the retrotransposon sequences. The remnants sometimes include the deleted LTRs, any intact domain or internal region from an element and sometimes many pseudo copies (Figure 2.3). The copy numbers for each superfamily were estimated by the formula. Copy no. = no. in database x genome size/database size (Tu, 2001).

3.2.2 Distribution of LTR retrotransposons in *Brassica* BACs

The dot plot analysis revealed that some BAC sequences have shown a high activity of LTR retrotransposons, while others have only one or two copies in them or even no copies. It depends on the region of the chromosomes, from where the BAC clone is sequenced. LTR retrotransposons are present both in *Brassica rapa* and *Brassica oleracea* BAC clones but maximum activity is seen in *Brassica oleracea*. The maximum number of elements were observed in BAC AC240090.1, where five Copia and one Gypsy element was detected. The total size of the AC240090.1 is 117.7 kb, while the total size of five elements harbouring in this BAC is 33.3 kb, 28.5% of total BAC sequence (Figure 3.1).

Another BAC clone AC183496.1 contains 4 individual copies of elements, 3 Copia (5063 bp, 4616 bp and 4001 bp) and a Gypsy element (11275 bp). The total size of the BAC clone sequence is 385 kb and the total lengths of elements are 25 kb, covering 15.5% of total BAC sequence.

The analysis of each retrotransposon sequence against itself in the dotplot indicated the LTRs, which are variable in sizes in various supefamilies (Copia, Gypsy and LARDs) (Figure 3.2).

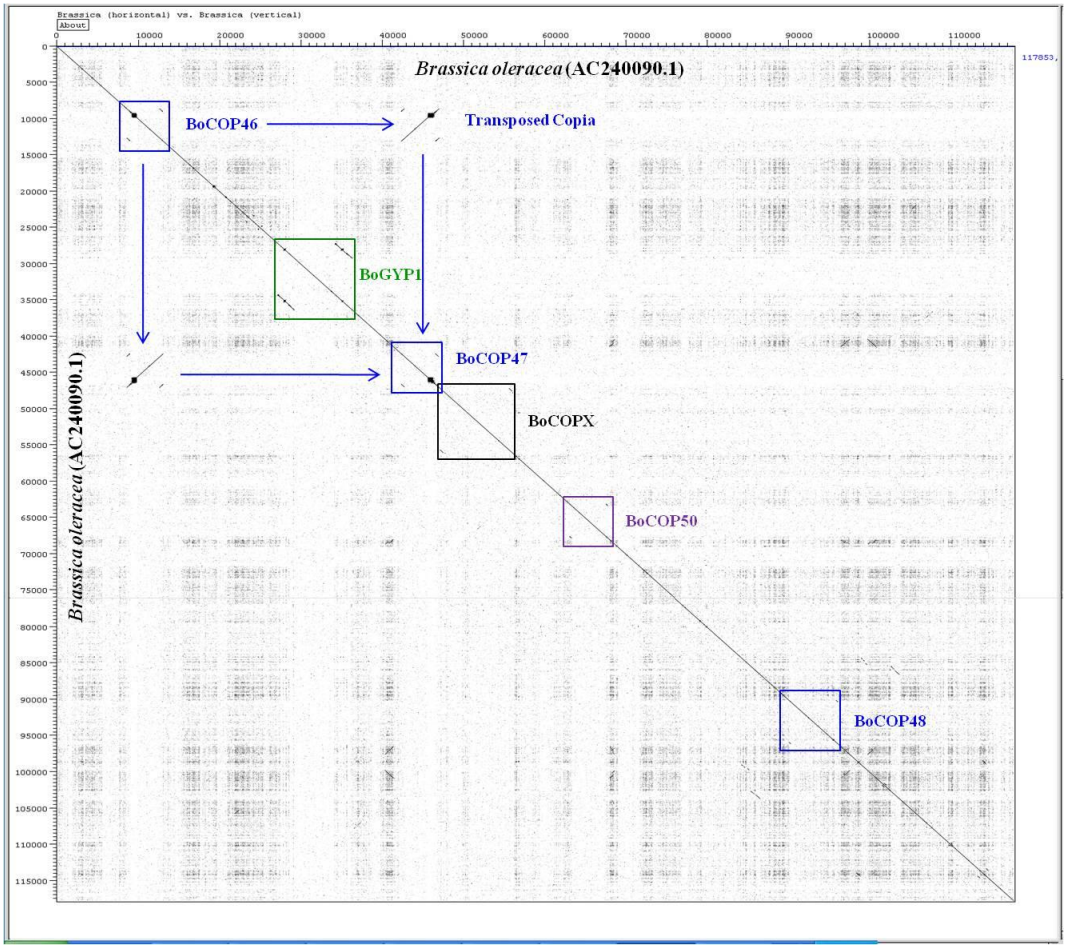


Figure 3.1: Dot plot of *Brassica oleracea* (AC240090.1) BAC sequence against itself to identify LTR retrotransposons. The central diagonal line running from one corner to other shows the homology of the sequence to itself. The boxes on the diagonal line show the position of LTR retrotransposon insertions with LTRs. Five Copia and one Gypsy elements are inserted, with a total size of 33.3 kb out of 117.7 kb BAC size covering 28.5% of total BAC sequence (scales indicate base numbers).

3.2.3 Diversity and abundance of LTR retrotransposons in *Brassica* genomes

The diversity, evolution and abundance of LTR retrotransposons in *Brassica* genomes was studied by a combination of bioinformatics and molecular genetics approaches. The reference full length elements (77) and their solo LTRs were used as query against the *Brassica* Nucleotide Collection (nr/nt) database from NCBI before January 2012, and all full length, truncated, partial elements and their remnants were counted. Around 14904 copies of elements and their fragments belonging to Copia, Gypsy, LARDs and TRIM-like elements were counted. Out of 14904 elements, only 280 are intact elements, of which 206 elements belongs to Copia, 56 from Gypsy, 16 from LARDs and 1 TRIM superfamily (Figure 3.3). A total of 178 truncated elements, 857 partial copies, 101 solo LTRs and 13488 remnants were counted. The ratio of intact elements to solo LTRs in *Brassica* BAC sequences is ~2:1. The remnants covered more than 90% of the copies identified in this study, but due to small sizes they cover less percentage of the genome as compared to intact or truncated copies. A total of 857 partial copies of LTR retrotransposons are present, which approximately cover the same size as full length elements. A total of 16 full lengths and 8 truncated copies of LARDs were collected from blast searches. We have not any precise sequence alignment for the partial copies, truncated elements and the remnants due to their high numbers and degraded sequences but have an approximate estimation of partial copies and fragments based on the length of retrieved sequences against the reference element.

The copy numbers of intact elements from *Brassica rapa* and *Brassica oleracea* whole genomes were estimated. A total of 1596 Copia, 540 Gypsy, 110 LARDs and 25 TRIM were estimated for *Brassica rapa*, while 7540 Copia, 780 Gypsy and 760 LARDs with no TRIM were estimated for *Brassica oleracea* whole genomes (Figure 3.3). Collectively, 11351 copies of LTR retrotransposons including LARDs and TRIM were estimated from *Brassica rapa* and *Brassica oleracea*. The results indicated that Copia superfamily is more diverse and abundant in *Brassica* as compared to Gypsy followed by LARDs and TRIM (Figure 3.3).

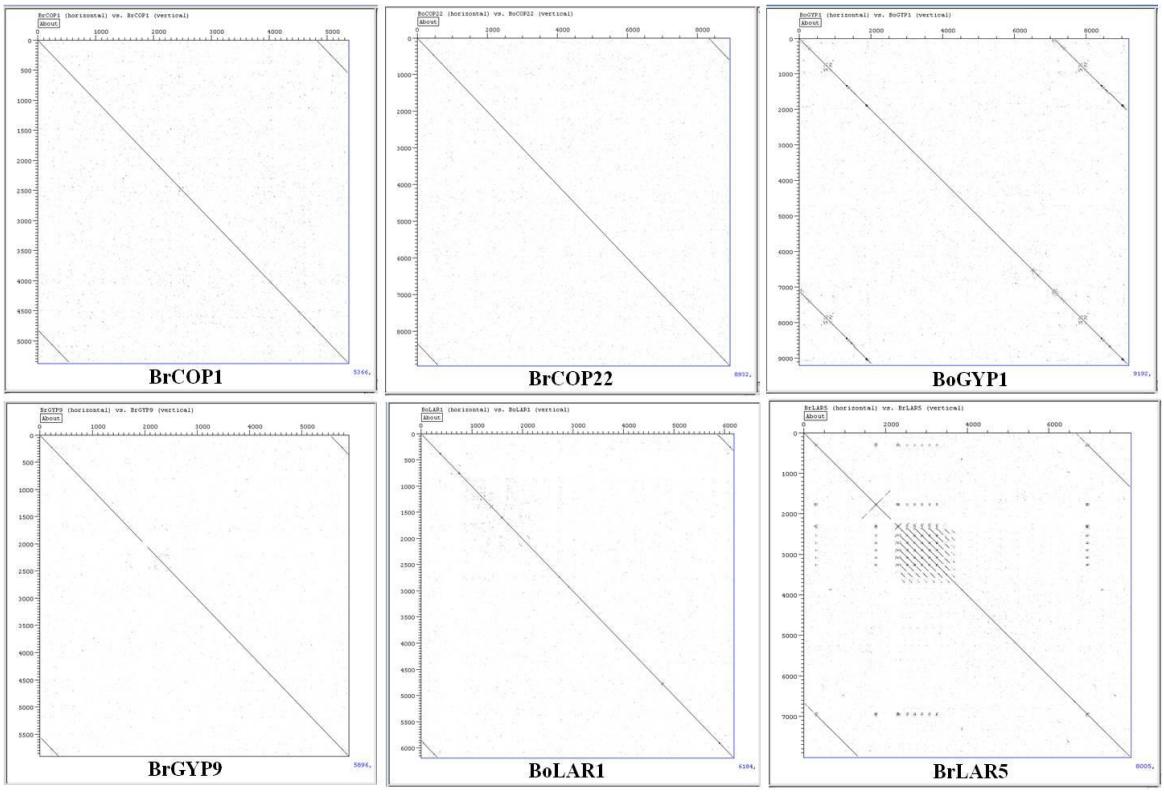


Figure 3.2: Dot plot graphs showing the LTRs in Copia, Gypsy and LARDs. The sequences from elements are plotted against themselves. The central diagonal line showed the homology of the sequences. The parallel lines at the corners are indicating the LTRs. The parallel diagonals line in *BrLAR5* indicates ~200bp tandem repeats in that specific region.

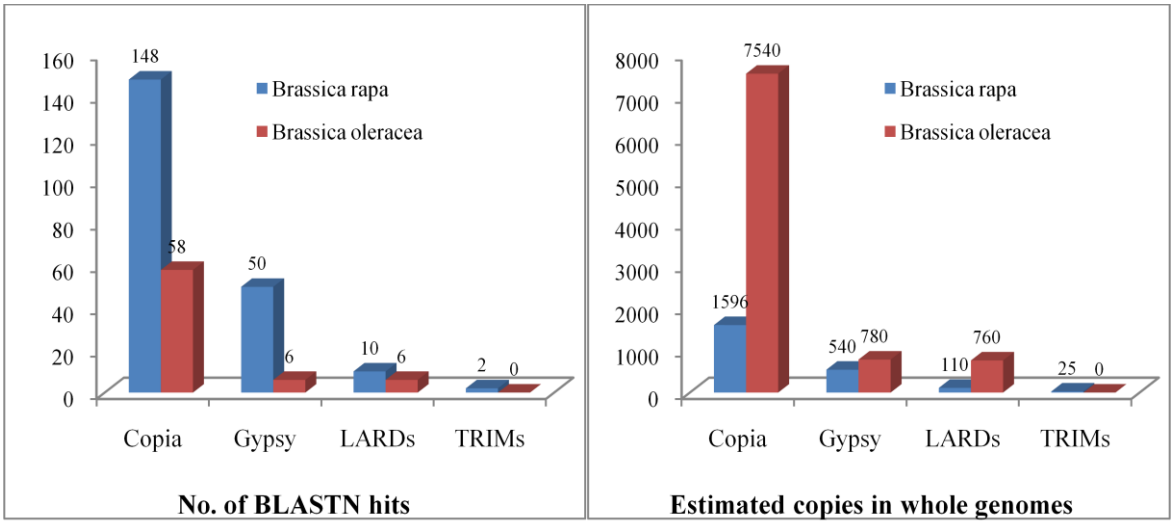


Figure 3.3: The graphic representation of copy numbers for each group of LTR retrotransposons identified in *Brassica*. The strong BLASTN hits for intact elements against *Brassica* Nucleotide Collection database in NCBI (left) and their estimated copy numbers in whole *Brassica rapa* and *Brassica oleracea* genomes (right) are represented.

3.2.4 Phylogenies of LTR retrotransposons from *Brassica* and other plant genomes

The phylogenetic relationship of 110 reverse transcriptase (RT) domains from *Brassica* and other known retrotransposons from plants were analysed. Out of 110 RT sequences, 63 were from *Brassica* LTR retrotransposons identified in this study. The other 47 RT sequences belonging to well known Copia, Gypsy and Caulimovirideae (Pararetovirus-like) superfamilies were collected from Gypsy database having sequences from all types of retrotransposons (Table 2.6). Of 63 *Brassica* RT, 53 sequences were from Copia and 10 from Gypsy superfamily. The protein sequences from RT regions of all these elements were aligned in CLUSTALW, which have shown some conserved amino acid motifs (Figure 3.13). The tree showed two main lineages separating the Copia and Gypsy superfamilies. A total of 76 families from 110 RT domains were observed. A family is defined when two or more members share >85% RT region. The 63 *Brassica* elements fall into 41 families, of which 35 are Copia and 6 are Gypsy based families. The known elements from different plants assemble into 35 families. The tree generated 9 clades, 6 from Copia and 3 from Gypsy superfamily. The members in a clade range from 5 to 22 elements. The members of *Brassica* Gypsy are found distributed within the Gypsy clade, with one *Brassicaceae* specific group and others making family specific subgroups. The *Brassica* gypsy *BoGYP1* share a family with *Arabidopsis* '*Legolas*' while the non-chromodomain elements (*BrGYP3, BrGYP4, BrGYP5, BrGYP6, BrGYP7*) clustered together with *Arabidopsis* '*Tat4-1*' and '*Tf2*' elements (Figure 3.4) suggesting the origin of these elements predating the separation of two genera ~19-20 Mya.

In contrast to Gypsy, the Copia elements are highly abundant in *Brassica* genomes. The evolutionary study of Copia RT sequences suggested the clustering of *Brassica* specific groups. One group of 10 related *Brassica* (9 *B. oleracea*, 1 *B. rapa*) elements were identified in a deep branch, forming a sister group to *Vitis vinifera* '*Vitico1*' element. Another group of 12 elements share a *Brassica* specific clade, where *Oryza sativa* '*Oryco1-2*' make an out group. *BrCOP4*, *BrCOP5*, *BrCOP17* and *BoCOP32* share a same group, where *BrCOP4* and *BoCOP32* share a single family, while *BrCOP5* and *BrCOP17* share another family. Another *Brassica* specific group is observed, where *Solanum* (*Lycopersicon*) *esculentum* '*Tork-4*' makes a sister family with *BoCOP36* (Figure 3.4). Two *Brassica* Copia elements share the family with known elements; *BrCOP20* shares a family with *Brassica* Copia '*Melmoth*' and *BoCOP45* shares a family with '*Araco*', a 4.8

kb element from *Arabidopsis thaliana*. The evolutionary analysis indicated that despite of homology in RT regions, the *Brassicaceae* members showed genera or species specific clusters (Figure 3.4). The tree also revealed that the known retrotransposons from other plants mostly formed outgroups from *Brassica*, suggesting only ancient common ancestors.

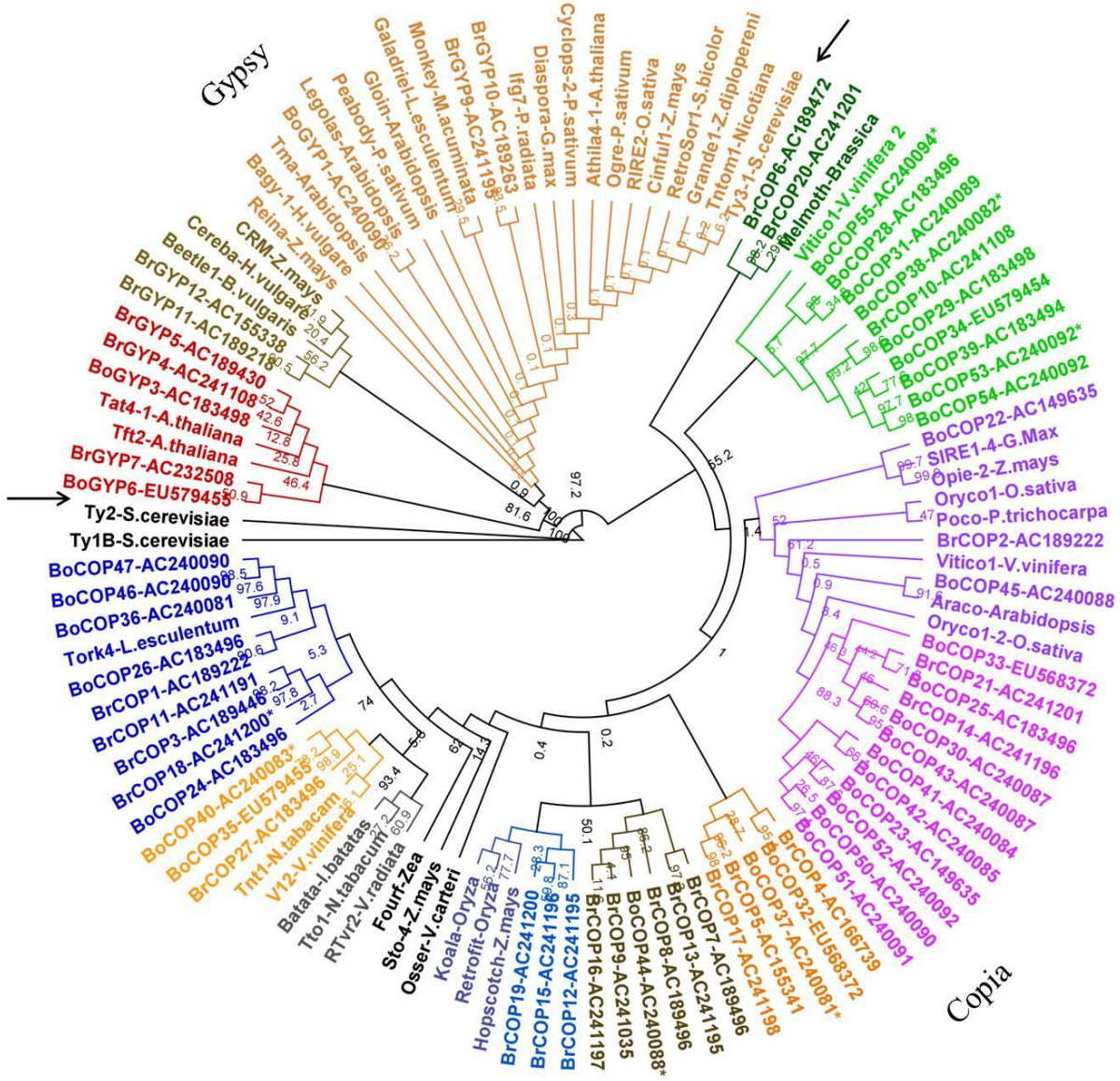


Figure 3.4: Phylogenetic relationships of LTR retrotransposon families identified in *Brassica* and other plant LTR retrotransposons. The RT sequences of 110 individual elements were used to construct the phylogenetic tree, which was rooted using the RT sequences of Copia element *Ty1B* of *Saccharomyces cerevisiae.* Out of 110, 63 RT sequences are from *Brassica* and the remaining 47 are from known retrotransposons collected from Gypsy database (Table 2.6). Neighbour-Joining tree was constructed with 1000 bootstrap replicates in Geneious Pro5.5.6 program. The consensus percentages are given at each node. Two main lineages separate the Gypsy and Copia, where 9 clades are represented and sub-clades are shown by different colours. About 75 families were identified; of which 49 are Copia and remaining 26 are Gypsy. The arrows mark the separation of Copia and Gypsy lineages. Br: *Brassica rapa.* Bo: *Brassica oleracea*. Bn: *Brassica napus.* COP: Copia. GYP: Gypsy. The details for the known elements are given in table 2.6.

Table 3.1: List of Copia, Gypsy, LARDs, and TRIM with their sizes, TSDs, TIRs, positions and orientations in BAC clone sequences. ND: not determined. O.I BAC: Orientation in BAC

Element Name	Super-family	Accession	Species	Size	TSDs	LTRs	Position in BACs	O.I. BACs
BrCOP1	Copia	AC189222.1	*B. rapa*	5366	GTGAA	539/541	54707-60072	3′-5′
BrCOP2	Copia	AC189222.1	*B. rapa*	4828	ATAAT	312/312	96814-101614	5′-3′
BrCOP3	Copia	AC189446.2	*B. rapa*	5778	CCTTT	493/493	74000-79760	5′-3′
BrCOP4	Copia	AC166739.1	*B. rapa*	6020	GTCAT	599/599	2956-8975	3′-5′
BrCOP5	Copia	AC155341.2	*B. rapa*	4807	CCGTC	180/180	67278-72084	5′-3′
BrCOP6	Copia	AC189472.2	*B. rapa*	5029	AGTTG	159/159	51849-56877	5′-3′
BrCOP7	Copia	AC189496.2	*B. rapa*	4481	ATTAG	152/152	72529-77009	3′-5′
BrCOP8	Copia	AC189496.2	*B. rapa*	4971	CCCTG	385/385	86234-91204	3′-5′
BrCOP9	Copia	AC241035.1	*B. rapa*	5313	GGATG	407/488	77808-83120	3′-5′
BrCOP10	Copia	AC241108.1	*B. rapa*	6489	AACCT	306/299	74968-81456	5′-3′
BrCOP11	Copia	AC241191.1	*B. rapa*	5630	ATTAA	304/304	60038-65667	3′-5′
BrCOP12	Copia	AC241195.1	*B. rapa*	4672	TATCT	147/147	5590-10261	5′-3′
BrCOP13	Copia	AC241195.1	*B. rapa*	4117	GTAAG	127/127	54558-58674	3′-5′
BrCOP14	Copia	AC241196.1	*B. rapa*	4595	AACTT	228/230	2514-29738	5′-3′
BrCOP15	Copia	AC241196.1	*B. rapa*	4585	CTCTA	172/172	80837-85421	3′-5′
BrCOP16	Copia	AC241197.1	*B. rapa*	4940	CTCTT	345/345	134939-139878	5′-3′
BrCOP17	Copia	AC241198.1	*B. rapa*	5010	GAACC	170/170	17376-22385	5′-3′
BrCOP18	Copia	AC241200.1	*B. rapa*	6096	AAAGT	399/399	46476-52571	3′-5′
BrCOP19	Copia	AC241200.1	*B. rapa*	4196	CACAA	121/121	61155-65350	5′-3′
BrCOP20	Copia	AC241201.1	*B. rapa*	4838	GAGGT	182/182	35112-39949	3′-5′
BrCOP21	Copia	AC241201.1	*B. rapa*	5089	ATAAT	266/266	95924-101012	3′-5′
BoCOP22	Copia	AC149635.1	*B. oleracea*	8922	TAGCT	579/582	23364-32285	3′-5′
BoCOP23	Copia	AC149635.1	*B. oleracea*	3757	GACTA	296/296	71762-75458	5′-3′
BoCOP24	Copia	AC183496.1	*B. oleracea*	5063	GAAGT	429/425	34468-39530	5′-3′
BoCOP25	Copia	AC183496.1	*B. oleracea*	4616	TCC	221/221	146660-151275	3′-5′
BoCOP26	Copia	AC183496.1	*B. oleracea*	4001	GTGTA	425/425	251315-255315	3′-5′
BoCOP27	Copia	AC183492.1	*B. oleracea*	4790	CCCCC	368/368	38224-43014	3′-5′
BoCOP28	Copia	AC183492.1	*B. oleracea*	6395	CATAC	333/333	50944-57338	5′-3′
BoCOP29	Copia	AC183498.1	*B. oleracea*	6576	ATATT	288/318	162553-169128	5′-3′
BoCOP30	Copia	AC240087.1	*B. oleracea*	4682	AGTTT	268/253	71136-75817	5′-3′
BoCOP31	Copia	AC240089.1	*B. oleracea*	6230	ACAAT	249/249	11346-17575	3′-5′
BoCOP32	Copia	EU568372.1	*B. oleracea*	6160	TGAAC	577/587	31626-37785	5′-3′
BoCOP33	Copia	EU568372.1	*B. oleracea*	4660	ACTTT	201/252	56936-61595	5′-3′
BoCOP34	Copia	EU579454.1	*B. oleracea*	6060	ATTAT	233/244	48881-54940	5′-3′
BoCOP35	Copia	EU579455.1	*B. oleracea*	4769	ACTAA	392/392	61558-66325	3′-5′
BoCOP36	Copia	AC240081.1	*B. oleracea*	5108	GCACT	366/366	41065-46172	5′-3′
BoCOP37	Copia	AC240081.1	*B. oleracea*	4879	TTGTA	170/170	59406-64283	3′-5′
BoCOP38	Copia	AC240082.1	*B. oleracea*	7097	TAAAT	313/313	2322-9418	3′-5′
BoCOP39	Copia	AC240082.1	*B. oleracea*	5371	TACAG*	304/293	61467-66837	3′-5′
BoCOP40	Copia	AC240083.1	*B. oleracea*	4778	AAGAG	370/370	43143-47920	3′-5′
BoCOP41	Copia	AC240084.1	*B. oleracea*	4690	CCTTA	300/303	66766-71455	5′-3′
BoCOP42	Copia	AC240085.1	*B. oleracea*	4656	GAACA	264/264	71673-76328	5′-3′
BoCOP43	Copia	AC240087.1	*B. oleracea*	4682	AGTTT	268/253	71136-75817	5′-3′
BoCOP44	Copia	AC240088.1	*B. oleracea*	4802	CATTG	321/320	48706-53507	3′-5′
BoCOP45	Copia	AC240088.1	*B. oleracea*	4706	GACAT	400/400	57933-62638	3′-5′
BoCOP46	Copia	AC240090.1	*B. oleracea*	4450	CTTTT	366/366	8583-13032	3′-5′
BoCOP47	Copia	AC240090.1	*B. oleracea*	4616	CTATA	366/366	42364-46979	5′-3′
BoCOP48	Copia	AC240090.1	*B. oleracea*	6096	TAAAT	257/248	90035-96130	3′-5′
BoCOP49	Copia	AC240091.1	*B. oleracea*	6096	ATTTA	248/257	28774-34869	5′-3′
BoCOP50	Copia	AC240090.1	*B. oleracea*	4748	AAGCA	263/263	63073-67820	5′-3′
BoCOP51	Copia	AC240091.1	*B. oleracea*	4748	TGCTT	263/263	57085-61832	3′-5′
BoCOP52	Copia	AC240092.1	*B. oleracea*	4763	GAGAC	288/288	15999-20762	3′-5′
BoCOP53	Copia	AC240092.1	*B. oleracea*	5887	AATAG	200/198	71126-77012	3′-5′
BoCOP54	Copia	AC240093.1	*B. oleracea*	4703	TATCG	273/273	41973-46475	5′-3′
BoCOP55	Copia	AC240094.1	*B. oleracea*	6131	AATTA	251/250	36442-41571	3′-5′
BoGYP1	Gypsy	AC240090.1	*B. oleracea*	9161	CAAAA	2004/2035	27208-36368	3′-5′
BoGYP2	Gypsy	AC183496.1	*B. oleracea*	11275	GCTGA	1140/1272	283163-294437	3′-5′
BoGYP3	Gypsy	AC183498.1	*B. oleracea*	11845	GTGTT	471/476	257711-269554	5′-3′
BrGYP4	Gypsy	AC241108.1	*B. rapa*	11744	GATTC	480/480	31686-43429	5′-3′
BrGYP5	Gypsy	AC189430.2	*B. rapa*	11872	CTAGG	480/480	107900-119771	5′-3′
BoGYP6	Gypsy	EU579455.1	*B. oleracea*	11576	ATGGC	508/509	13914-25488	3′-5′

Table 3.1: *Continued*

Element Name	Super-family	Accession	Species	Size	TSD	LTR	Position	O.I. BAC
BrGYP7	Gypsy	AC232508.1	*B. rapa*	11664	ATCTT	506/506	118772-130435	3′-5′
BrGYP8	Gypsy	AC241108.1	*B. rapa*	5094	TGGGG	331/331	74345-79439	5′-3′
BrGYP9	Gypsy	AC241195.1	*B. rapa*	5900	GATTG	346/339	43731-49630	3′-5′
BrGYP10	Gypsy	AC189263.2	*B. rapa*	5221	CAAGA	346/346	38008-43228	5′-3′
BrGYP11	Gypsy	AC189218.2	*B. rapa*	5173	CTCTA	340/343	68590-73762	5′-3′
BrGYP12	Gypsy	AC155338.1	*B. rapa*	5163	CTTAA	360/360	110515-115677	5′-3′
BoGYP13	Gypsy	AC240081.1	*B. oleracea*	4168	TGCGC	199/200	89533-93700	3′-5′
BrGYP14	Gypsy	AC189233.2	*B. rapa*	7195	ATCAT	1553/1553	66972-74166	3′-5′
BrGYP15	Gypsy	CU984545.1	*B. rapa*	5140	GGGAA	369/369	74632-79771	5′-3′
BoLAR1	LARDs	AC149635.1	*B. oleracea*	6183	TAATA	313/322	8183-14365	ND
BoLAR2	LARDs	AC183498.1	*B. oleracea*	6008	TTGTC	231/231	301636-307643	ND
BoLAR3	LARDs	AC183498.1	*B. oleracea*	5816	CAAAC	720/707	337911-343726	ND
BrLAR4	LARDs	AC189415.2	*B. rapa*	5670	CATAT	666/666	59818-65487	ND
BrLAR5	LARDs	AC241138.1	*B. rapa*	7991	GGTGA	1319/1319	34978-42986	ND
BrLAR6	LARDs	AC241195.1	*B. rapa*	3819	CGATG	347/347	36606-40424	ND
BrTRI1	TRIM	AC155342.2	*B.rapa*	1323	GAAAT	257/262	10377-11697	ND

3.2.5 Structural features of Ty1/copia elements identified from *Brassica rapa*

The dot plot analysis of BAC sequences against themselves led to the identification of 55 Copia elements. The Copia are generally smaller than Gypsy elements with a size of 3.7-8.9 kb and average of 4.5-5.5 kb. *BrCOP1* was identified from *Brassica rapa* accession 'AC189222.1'. It is 5.3 kb element generating 5 bp TSDs, flanked by 5′-541/539-3′ bp LTRs (Table 3.1) and has both PBS and PPT down and upstream to 5′LTR and 3′LTR respectively. The genomic sequence of the element is AT rich (57%), where internal regions are more AT rich as compared to its LTRs (54%). The *gag-pol* genes display the 5′-GAG-INT-RT-RH-3′ structure (Figure 3.5). A 4.8 kb large element *BrCOP2* was also identified from the *Brassica rapa* accession 'AC189222.1' flanked by 312 bp LTRs, and displays 5′-GAG-AIR1-ZK-INT-RT-RH-3′, where an Arginine methyltransferase-interacting protein and Zinc knuckle (ZK) motif is present.

BrCOP3 is 5.7 kb in size displaying PBS and PPT sites and domain organization as 5′-GAG-ZK-INT-RT-RH-3′, where an extra ZK like motif is incorporated before the integrase (Figure 3.5). *BrCOP4* is 6.0 kb in size, including the 599 bp LTRs and a PBS on downstream of 5′LTR and PPT motif towards the upstream of 3′LTR. The *pol* gene encodes domains (5′-INT-RT-RH-3′), which clearly characterize this element to Copia superfamily. *BrCOP5* is 4.8 kb in size, flanked by 180 bp LTRs, exhibit the PBS and PPT motifs and a typical Copia domain organization (5′-GAG-INT-PRK-RT-RH-3′). Additionally to typical *gag-pol* coding domains, *BrCOP5* and its homologues show an

extra protein domain between INT and RT called PRK, which is a 2′-phosphodiesterase/3′-nucleotidase precursor protein (Table 3.2). *BrCOP6* is similar to *BrCOP5* but larger in size with small LTRs and lacking additional PRK domain.

BrCOP7 and *BrCOP8* are 4.4 and 4.9 kb in size and flanked by 152 and 385 bp LTRs respectively (Table 3.1). *BrCOP7* exhibit both PBS and PPT motifs, while PBS is not detected in *BrCOP8*. Their domain organization is 5′-GAG-INT-PRK-RT-RH-3′. *BrCOP9* is 5.3 kb in size, flanked by 488 bp 5′LTR and 407 bp 3′LTR. It displays the PBS and a PPT with typical Copia *gag-pol* genes encoding protein domains as 5′-GAG-HVE-INT-RT-RH-3′ (Table 3.2). It harbours a Herpes virus envelop-like (HVE) protein domain, which is not observed in any of the other Copia elements investigated (Figure 3.5). The genome of *BrCOP10* is 6.5 kb in size including the 306 bp 5′LTR and 299 bp 3′LTR. It exhibit PBS and PPT motifs and *gag-pol* genes encoding the sequences as 5′-GAG-INT-RT-RH-3′. The two elements *BrCOP11* and *BrCOP18* showed >90% similarity in their RT-domains and are 5.6 and 6.1 kb in size respectively. Both display the PBS and PPT towards downstream and upstream of 5′LTR and 3′LTR and have the similar protein domain organization of 5′-GAG-INT-RT-RH-3′. *BrCOP12* and *BrCOP13* display a genome of about 4.6 and 4.1 kb and are flanked by 147 and 127 bp respectively. Both display typical *gag-pol* Copia-like polyprotein domains but an extra Phage virion morphogenesis (PVM) protein is incorporated between RT and RH domains in *BrCOP13* (5′-GAG-INT-RT-PVM-RH-3′). *BrCOP14* and *BrCOP15* are same sized Ty1/copia elements identified in *Brassica rapa* accession ‘AC241196.1’. They are 4.6 kb in size, includes the 230 and 172 bp LTRs respectively (Table 3.1).

BrCOP16 and *BrCOP17* share few structural features. The sizes of the elements are 4.9 and 5.0 kb, including LTRs of 345 and 177 bp respectively. They have PBS next to 5′LTR complimentary to $tRNA_{Met}$, and 15 bp PPT adjacent to 3′LTR. *BrCOP19* is a 4.2 kb long element including short LTRs of 121 bp. The *gag-pol* gene (5′-GAG-INT-ETS-RT-RH-3′) showed the presence of an extra transcription factor (ETS) domain. Two elements *BrCOP20* and *BrCOP21* were identified in *Brassica rapa* accession ‘AC241201.1’. They are phlogenetically distinct, with different mode of *gag-pol* domain organization, different types of PBS and PPT motifs and LTRs of variable sizes. *BrCOP20* is 4.8 and *BrCOP21* in 5.0 kb in size and are flanked by 182 and 266 bp LTRs respectively (Table 3.1).

3.2.5.1 Structural features of Ty1/copia elements identified from *Brassica oleracea*

Interestingly, the largest (*BoCOP22*) and smallest (*BoCOP23*) Copias were identified from the same *Brassica oleracea* accession 'AC149635.1'. *BoCOP22* is 8.9 kb, including flanking LTRs of 5′-582/579-3′ bp. It displays a PBS next to 5′LTR and a PPT adjacent to 3′LTR with an extra AIR1 domain (5′-GAG-AIR1-INT-RT-RH-3′). An additional unrelated sequence/insertion was found towards the C-terminus of RT and RH domains (Figure 3.5). *BoCOP23* is 3.7 kb in size, flanked by 296 bp LTRs and 5′-RT-RH-3′ domains in it. Three retrotransposons were identified from *Brassica oleracea* accession 'AC183496.1 named *BoCOP24*, *BoCOP25* and *BoCOP26*, which are 5.0, 4.6 and 4.0 kb in size, flanked by LTRs of 525, 221 and 525 bp respectively. The PBS of *BoCOP24* and *BoCOP26* are similar and complementary to $tRNA_{Met}$ but *BoCOP25* exhibits $tRNA_{Trp}$. The typical *gag-pol* gene polyproteins domain organization (5′-GAG-INT-RT-RH-3′) is observed in all except *BoCOP25* (Table 3.2). The *BoCOP27* have a genome of 4.8 kb including 368 bp LTRs and terminated by a perfect 5′-CCCCC-3′ TSDs. *BoCOP28* and *BoCOP29* are 6.4 and 6.6 kb in sizes, flanked by 333 bp and 5′-288/318-3′ bp LTRs respectively. They exhibit PBS and PPT motifs in their structures and have *pol* protein gene encoding domains as 5′-INT-RT-RH-3′. The genome of *BoCOP30* is also similar to *BoCOP29*, but it is a member of another family. *BoCOP31* is 6.3 kb large, terminated by 249 bp LTRs, and displays the PBS and PPT downstream and upstream of 5′LTR and 3′LTR respectively (Figure 3.5; Table 3.2).

Brassica oleracea BAC clone 'EU568372.1' harbour *BoCOP32* and *BoCOP33*, which are 6.1 and 4.6 kb in sizes terminated by 5 bp TSDs. Their 5′ and 3′LTRs are variable in numbers, which are 5′-577/587-3′ bp in *BoCOP32* and 5′-201/252-3′ bp in *BoCOP33*. The difference in sizes of 5′LTR and 3′LTR in both elements is due to short insertions/deletions in their other LTRs. Both display the similar type of PBS and PPT motifs with similar *gag-pol* proteins organization (5′-GAG-INT-RT-RH-3′) (Table 3.2). A similar Copia element *BoCOP34* was studied from *Brassica oleracea* accession 'EU579454.1'. The genomic organization of the element display 6.0 kb long element, flanked by 5′-233/244-3′ bp LTRs. *BoCOP35* is a Copia element from *Brassica* with an average size of 4.8 kb, including the perfect 392 bp LTRs. No identifiable PBS and PPT motifs were detected in this element with 5′-GAG-DUF-ZK-INT-RT-RH-3′ domain organization (Table 3.2).

Two retroelements *BoCOP36* and *BoCOP37* were found in *Brassica oleracea* accession 'AC240081.1', which are 5.1 and 4.9 kb in sizes, flanked by 366 and 170 bp LTRs respectively. The *Brassica oleracea* accession 'AC240082.1' harbours two elements named *BoCOP38* and *BoCOP39*, which are 7.1 and 5.3 kb long. The elements have LTRs of 313 bp and 5′-293/304-3′ bp respectively. No PBS was detected in *BoCOP38*, while *BoCOP39* exhibit both PBS and PPT motifs (Table 3.2). *BoCOP41*, *BoCOP42* and *BoCOP43* are around 4.6 kb in sizes, flanked by 300, 264 and 268 bp LTRs respectively. They have *gag-pol* protein domains 5′-GAG-INT-RT-RH-3′ and exhibit PBS next to 5′LTR and PPT adjacent to 3′LTR, but PBS is lacking in *BoCOP43*. While screening the *Brassica oleracea* BAC clone accession 'AC240088.1', two Copia elements named *BoCOP44* and *BoCOP45* were identified displaying 4.8 and 4.7 kb sizes, flanked by 320 and 400 bp LTRs respectively. The LTRs flanked the internal region displaying a typical PBS and a PPT strand towards the downstream of 5′LTR and adjacent to 3′LTR. They exhibit a typical Copia-like *pol* domain structures 5′-INT-RT-RH-3′, where an additional unknown protein (UKP) is present in *BoCOP44* and a ZK motif in *BoCOP45* (Table 3.2).

The elements *BoCOP46*, *BoCOP47*, *BoCOP48* and *BoCOP50* are detected in *Brassica oleracea* accession AC240090.1 (Figure 3.1). *BoCOP46* and *BoCOP47* are about 4.5 and 4.6 kb in size, belonging to the same family and are flanked by 366 bp LTRs. *BoCOP48* and *BoCOP49* are similar elements present in opposite orientations in two different *Brassica oleracea* accession (AC240090.1, AC240091.1). These elements represent a genomic size of about 6.1 kb flanked by LTRs of 5′-248/257-3′ bp. They display a similar PBS and PPT motifs with *pol* region encoding RH domain only (Table 3.1 & 3.2). This indicates that the other protein domains were lost by during the rearrangement of the element in the evolutionary phases. The members of *BoCOP50*, *BoCOP51* and *BoCOP52* were identified from *Brassica oleracea* accessions 'AC240090.1', 'AC240091.1' and 'AC240092.1' respectively. *BoCOP50* and *BoCOP51* are homologous and share same family, while *BoCOP52* makes a sister family. The genome of these elements is 4.7 kb in size, including LTRs of 263-288 bp and showing the typical PBS and PPT motifs. The PBS used the $tRNA_{Met}$ in all the three elements. The *gag-pol* domain organization is 5′-GAG-INT-RT-RH-3′. The genome of *BoCOP53*, *BoCOP54* and *BoCOP55* is about 5.9, 4.7 and 6.3 kb in sizes, including LTRs of about 200, 273 and 251 bp respectively with an internal region having *gag-pol* gene polyproteins (5′-GAG-INT-RT-RH-3′) with additional 'DUF' in *BoCOP55* (Figure 3.5; Table 3.2).

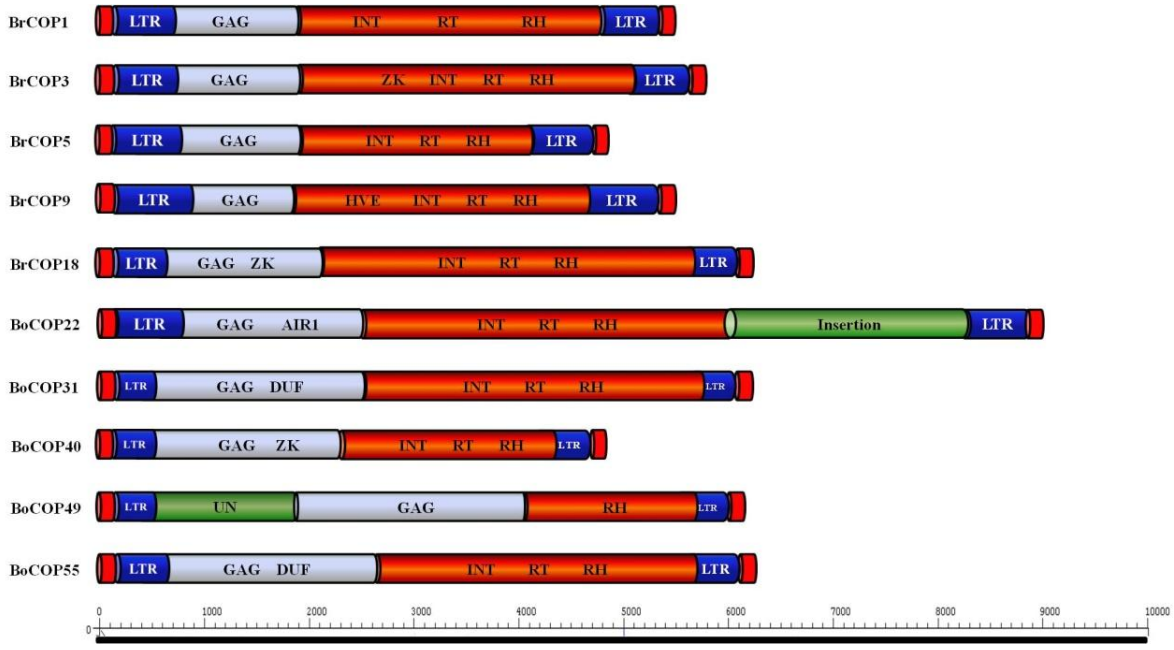


Figure 3.5: Structures of Copia elements in *Brassica*. The red discs at the ends represent the TSDs. LTRs are shown in blue. The *gag* and *pol* regions are drawn with their protein domains. The scale below is measuring the lengths of the elements (bp). Additional insertions are highlighted by green. AP: Aspartic protease. RT: Reverse transcriptase. INT: integrase. GAG: *gag*-nucleocapsid. ZK: zinc knuckle. DUF: domain of unknown function. AIR1: Arginine methyltransferase-interacting protein. UN: unknown.

3.2.5.2 Protein domain organization of *gag-pol* genes in *Brassica* Copias

The genome organization of *gag-pol* proteins were analyzed in 55 intact Copia elements (Table 3.2) identified here. Two major types of *gag/pol* domain organization were observed with other 8 sub-patterns with one or other additional or deleted domains in *pol* regions. The canonical Copia elements have 5′-GAG-INT-RT-RH-3′ domain organization. Out of 55 elements, 50 showed this arrangement with or without other additional domains. The other main type of domain organization is observed in elements, which lack *gag* gene but exhibit *pol* only. The sub-patterns include an additional domain in *pol* or lack of one or more domains. The domain organization such as 5′-RT-3′, 5′-RH-3′, 5′-RT-RH-3′ and 5′-INT-RT-3′ was shown by few elements. The elements displaying this kind of domain organizations are considered as deleted elements, with deletion of one or more domains by chromosomal rearrangement or other factors. If the elements have lost their RT region, then they are considered as non-autonomous as they cannot further replicate and move within the genome. Around 16% of elements have shown an extra ZK domain in *gag* gene upstream to INT domain (5′-GAG-ZK-INT-RT-RH-3′), while the remaining elements have shown extra protein domain within *pol* polyprotins as 5′-INT-ETS-RT-RH-3′, 5′-DUF-INT-RT-RH-3′ and 5′-INT-UTP-RT-RH-3′ (Table 3.2).

Table 3.2: List of *Brassica* retrotransposons with PBS, PPT motifs and *gag-pol* gene protein domains. AP: Aspartic protease. RT: reverse transcriptase. INT: integrase. ZK: zinc knuckle. ZF: zinc finger. CHR: Chromodomain. HVE: Herpes virus envelop. CHR: Chromatin organization modifier. PVM: Phage virion morphogenesis. ETS: ETS-domain transcription factor. UKP: Unknown protein. DUF: Protein of Unknown function. AIR1: Arginine methyltransferase-interacting protein. NAD: NADH dehydrogenase subunit. PRK: bifunctional 2', 3'-cyclic nucleotide 2'-phosphodiesterase/3'-nucleotidase precursor protein. HVW: Herpes virus major outer envelope glycoprotein (BLLF1); TLC: TLC domain. CL: Copia-like.

Element Name	tRNA type	PBS (5´-3´)	Position	PPT (5´-3´)	Position	Domain Structure (5´-3´)
BrCOP1	Met	TATCAGAGCCAGGTT	572-586	AGAGAAAGATGGAAG	4803-4817	GAG,INT,RT,RH
BrCOP2	Thr	GCTTTACGTTTGAGAG	328-347	ATGATTAAGGAGGAG	4497-4511	GAG,AIR1,ZK,INT,RT,RH
BrCOP3	Met	TATCAGAGCACAGTTGATCG	504-524	GAGAGACGAAGTAGA	5219-5233	GAG,ZK,INT,RT,RH
BrCOP4	Met	TATCAGAGCCAGGTT	608-622	AAGCTTGAGGGGGAG	5402-5416	GAG,INT,RT,RH
BrCOP5	Tyr	TCCGCTACCAAAAGTTCG	236-255	GGAGTATTAGGAAAG	4634-4802	GAG,INT,PRK RT,RH
BrCOP6	Met	GTATCAGAGCATTTCTTT	267-284	CATCTTGAGGGGGGG	4851-4865	GAG,INT,RT,RH
BrCOP7	Thr	AGACTGTTCTTGAATGAGTTG	195-216	AGAAGAGCAGAGAAG	4236-4250	GAG,INT,RT,RH
BrCOP8	ND	----	----	AGAGATGGAGGAGCG	4537-4551	GAG,INT,RT,RH
BrCOP9	Gln	AGGTCTTCACCGGTAAGGATT	262-282	GGTTGAGAGTATAGA	4544-4558	GAG,HVE,INT,RT,RH
BrCOP10	*Trp	TAAATCCCTGAGACCTAAATC	333-353	GAATGTTATAAAGAA	6182-6196	GAG,INT,RT,RH
BrCOP11	*Pro	TATAGTTGATAGAATCTTG	310-327	AGAGAGGTGAAGACA	5233-5247	GAG,ZK,INT,RT,RH
BrCOP12	Met	AACCTCTCTCCCGTGCCCA	212-220	CCTCCACCCCTTCTC	4444-4458	GAG,INT,RT,RH
BrCOP13	*Thr	TGCCTCCAAGCTAAAACGAT	170-190	AAGACTGCGGGGGAG	3971-3985	GAG,INT,RT,PVM,RH
BrCOP14	Leu	GAGCATTCTATTGAATT	247-264	TAAGGGGGAGAATGT	4349-4363	GAG,INT,RT,RH
BrCOP15	Gln	AGCGTTCCAAACCGAGTCCTT	225-245	ATGGATCGAAAGGTG	4383-4397	GAG,INT,RT,RH
BrCOP16	*Met	TATCAGAGCTCAGCAAGT	354-371	GAGTTTGCGAGGGGA	4576-4590	GAG,INT,RT,RH
BrCOP17	Met	TATCAGAGCACAAAATTC	179-196	CAACTTGAGGGGGAG	4821-4835	GAG,INT,RT,RH
BrCOP18	Met	TATCAGAGCCAGGTT	410-424	AGAGAGACGGAGAAG	5644-5658	GAG,ZK,INT,RT,RH
BrCOP19	Val	GGCTTCGTCATGGTGTCG	201-218	GGTCTAGGAGCAAAG	4045-4059	GAG,INT,ETS,RT,RH
BrCOP20	Arg	ATCTTGCCAATGAGTGCG	224-241	AGCGAGAAAAAGAAA	4590-4604	GAG,INT,RT,RH
BrCOP21	Met	TATCAGAGCCAGGTT	277-292	TATCAGAGCCAGGTT	4751-4765	GAG,INT,RT,RH
BoCOP22	Leu	GACAGCTACAGTGAGATGTT	652-672	TAAAAAGGGGGAGAT	8324-8334	GAG,AIR1,INT,RT,RH
BoCOP23	ND	----	----	ND	----	RT,RH
BoCOP24	Met	TATCAGAGCCTGAGTTACG	440-458	AAGACAGAAGACAGA	4593-4607	GAG,INT,RT,RH
BoCOP25	Trp	CATCTCTTTGAATTTG	284-301	GATATCAATAAGAAG	4375-4389	GAG,ZK,INT,RT,RH
BoCOP26	Met	TATCAGAGCTGAGGTT	437-452	AGGACAAGGAGGAGA	3555-3569	RT,RH
BoCOP27	ND	----	----	GGGAAGGGGGAGATT	4404-4418	GAG,ZK,INT,RT,RH
BoCOP28	Arg	CGGTCCCCAAGGAGAGT	378-394	CCTCTACTATTATTT	5964-5978	GAG,INT,RT,RH,
BoCOP29	Ser	CGTTATCAGCACGATCG	294-311	GCATCAAAGGGGGAG	6239-6253	GAG,INT,RT,RH
BoCOP30	ND	----	----	GAAGTAAAGGAAGAA	4678-4682	GAG,INT,RT,RH
BoCOP31	Lys	ATCACTCTGCGATTCG	268-284	GAGAGCGGATAGTGA	5942-5956	GAG,DUF,INT,RT,RH
BoCOP32	Met	TATCAGAGCCAGGTT	586-600	AAGCTTGAGGGGGAG	5554-5568	GAG,INT,RT,RH
BoCOP33	Met	TATCAGAGCAAAATCT	262-277	AAGGAGATGCGAGAG	4670-4674	GAG,INT,RT,RH
BoCOP34	Thr	CGTTATCAGCACGATT	234-254	ACATCCAAGGGGGAG	5797-5811	GAG,INT,RT,RH
BoCOP35	ND	----	----	ND	----	GAG,DUF,ZK,INT,RT,RH
BoCOP36	Met	TATCAGAGCTTCGGGTT	378-394	AGTCAAGGTGGGGAG	4722-4736	GAG,INT,RT,RH
BoCOP37	Met	TATCAGAGCAGAAAGATTC	179-197	CAACTTGAGGGGGAG	4689-4703	GAG,INT,RT,RH
BoCOP38	ND	----	----	AGGTGGAGAGCACAA	6751-6765	GAG,INT,RT,RH
BoCOP39	Ser	CGTTGTCAGCACGATTACG	299-317	GCATCCAAGGGGGAG	5048-5062	GAG,INT,RT
BoCOP40	Met	TATCAGAGCCAGGTT	381-395	GGGAAGGGGGAGATT	4389-4403	GAG,ZK,INT,RT,RH
BoCOP41	Met	TATCAGAGCCTGAGTT	301-316	AAGGAAATGAGAGAC	4324-4338	GAG,INT,RT,RH
BoCOP42	Met	TATCAGAGCGTTAGGTTACG	275-294	AGCTCAAGAGAGAGA	4390-4404	GAG,INT,RT,RH
BoCOP43	ND	----	----	GAAGTAAAGGAAGAA	4387-4401	GAG,INT,RT,RH
BoCOP44	ND	----	----	GGAAAGGGATAAGGG	4416-4430	GAG,INT,UKP,RT,RH
BoCOP45	Met	TATCAGAGCTACAAGTTCC	409-427	AAGTTTAAGAGGGGG	4284-4298	GAG,ZK,INT,RT,RH
BoCOP46	Met	TATCAGAGCTTCGGTTT	378-395	AGTCAAGGTGGAGAA	4064-4078	RT
BoCOP47	Met	TATCAGAGCTTCGGGTT	378-394	AAGTCAAGATGGAGA	4229-4243	GAG,ZK,RT
BoCOP48	Leu	TGTCATAACCATATAGGGTTT	275-295	AAGGGCCGGAAGAGA	5761-5775	RH
BoCOP49	Leu	TGTCATAACCATATAGGGTTT	275-295	AAGGGCCGGAAGAGA	5761-5775	RH
BoCOP50	Met	TATCAGAGCCATTCA	274-290	AAAGAGATGAGAGAC	4413-4427	GAG,INT,RT,RH
BoCOP51	Met	TATCAGAGCCATTCA	274-290	AAAGAGATGAGAGAC	4413-4427	GAG,INT,RT,RH
BoCOP52	Met	TATCAGAGCTCCAGGTTTCG	298-317	AATTAAGGGGGAGAA	4457-4471	GAG,INT,RT,RH
BoCOP53	*Met	TGTCATAACCATACAGGGATT	218-238	AAACATAAAGAGTCA	5659-5673	GAG,INT,RT,RH
BoCOP54	Met	TATCAGAGCAACTAGGT	284-300	AAAGAAGATATGAAG	4397-4411	GAG,INT,RT,RH
BoCOP55	Pro	TATCATGTTATAATTG	313-331	AAGAGCGGATAGTGA	5880-5854	GAG,DUF,INT,RT,RH

Table 3.2: *Continued*

Element Name	tRNA type	PBS (5´-3´)	Position	PPT (5´-3´)	Position	Domain Structure (5´-3´)
BoGYP1	Met	TATCAGAGCGGGTTCCG	2107-2124	ATTAGTGGGGGAGAA	7138-7152	GAG,TLC,AP,RT,RH,INT
BoGYP2	Cys	AGGTCCCAATGCGTGGT	1185-1201	ND	----	GAG,AP
BoGYP3	Lys	CGCCCATCGTGGGGCT	486-501	GTGAACTGGAGGGGA	11345-11359	GAG,AP,RT,RH,INT
BrGYP4	Lys	CGCCCACCGTGGGGCT	491-506	GAACTGGGGGGGGAC	11241-11255	GAG,AP,RT,RH,INT
BrGYP5	Lys	CGCCCACCGTGGGACCG	491-508	GAACTGGGGGGGGAC	11369-11383	GAG,AP,RT,RH,INT
BoGYP6	Lys	CGCTCACCGTGGGATCA	520-537	ACTGGGGGGGGGGGG	11044-11058	GAG,RT,RH,INT
BrGYP7	Lys	CGCCCACCGTGGGGC	517-531	GATGGACTGGGGGGA	11134-11148	GAG,AP,RT,RH,INT
BrGYP8	Phe	TGCGGTGACTCGATCG	343-360	AAGCTTGAGGACAAG	4722-4736	GAG,AP,RH,INT,CHR
BrGYP9	Tyr	TTCGAACCTCGGAATC	361-378	GGGAGAAGAAGAAGC	5452-5466	GAG,AP,RT,RH,INT,CHR
BrGYP10	Tyr	TTCGAACCTCGGAATC	368-385	GGGAGAAGAAGAAGC	4773-4787	GAG,AP,RT,RH,INT,CHR
BrGYP11	Arg	CGATTCTACTCGTGATC	371-387	GTACGGGAGGGGACC	4814-4828	GAG,AP,RT,RH,INT,CHR
BrGYP12	Met	TATCAGAGACCTTTAAATTA	371-390	GTACGGGAGGGGACC	4784-4798	GAG,ZK,AP,RT,RH,ZF,INT
BoGYP13	Tyr	CGGATGAGCAGCGGCTGTG	196-214	AAGTAAAAGAATAAG	3939-3953	GAG,AP
BrGYP14	ND	----	----	AAAAGAAAATAAAAA	5552-5566	GAG,AP
BrGYP15	-Ser	CGAATCCTTCT-CACCCG	4742-4759	GCTTTGCTACGCTCC	388-402	GAG,AP,RT,RH,INT
BoLAR1	ND	----	----	GCATCCAAAGGAGAG	5842-5856	UD
BoLAR2	ND	----	----	ND	----	UD
BoLAR3	Met	TATCAGAGCGCTGGTT	718-733	AAAGGAAGGTAGAGA	5047-5061	UD
BrLAR4	Met	TATCAGAGCGCTGGTT	677-692	AAGGGAAGGTAGAGC	4955-4969	UD
BrLAR5	ND	----	----	ND	----	UD
BrLAR6	Cys	GGTCCCTCCGGGTTTG	353-369	AAGACACACAAATAA	3377-3391	UD
BrTRI1	Lys	TGTTCATTGGTGGTG-TTG	331-349	GGGGAGTATTAGAGA	1056-1070	UD

3.2.5.3 PBS and PPT of *Brassica* Copia elements

The PBS towards the downstream of 5´LTR and PPT located adjacent to 3´LTR were investigated in all Copia elements (Table 3.2). The PBS and PPT were identified by scanning each intact element with parameter 'Predict PBS by using *Arabidopsis thaliana* tRNA database' in LTR_FINDER. Out of all the elements investigated for the presence of PBS and PPT, 85% have shown the presence of both PBS and PPT, 12 % have shown no PBS and only 3% have shown no PPT. The PBS and PPT from *BoCOP23* and *BoCOP35* was not detected. No PBS was detected from *BrCOP8*, *BoCOP27*, *BoCOP30*, *BoCOP38*, *BoCOP43* and *BoCOP44* by scanning them against *Arabidopsis thaliana*, *Zea mays and Oryza sativa* tRNA database'. Eleven different tRNA types were used by PBS of Copia elements. The most frequently use type was $tRNA_{Met,}$ which was present in 45% of the elements. The second important primer type was $tRNA_{Thr,}$ which was present in 9% of the elements. The nucleotide sequences and positions of both PBS and PPT in all elements were defined and enlisted. It was observed that in majority of the elements, the PBS and PPT starts immediately after ending or before starting the 5´LTR and 3´LTR respectively, while in few cases, they are few bp to ~200 bp apart from the end and start of LTRs (Table 3.2).

3.2.5.4 Diversity and distribution of Copia retrotransposons in *Brassica* accessions

The diversity and distribution of various Copia elements among 40 *Brassica* cultivars/accessions were studied by PCR analysis using newly developed markers (reverse transcriptase amplification polymorphism; RTAP), where RT-specific primers were designed to amplify the elements. Eleven sets of primers (Table 3.3) were designed to amplify the RT regions of respective Copia families among *Brassica* crops. Primer set BrCOP2F and BrCOP2R was designed to amplify a 710 bp RT region of *BrCOP2* family. The results showed the amplification of RT regions from 37 cultivars from six *Brassica* diploid and allotetraploid species from the 'Triangle of U' and hexaploid *Brassica* (Table 2.1). The amplicons were observed in *Brassica rapa* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons), *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Cuor Di Bue Grosso, Precoce Di Calabria, GK97361), *Brassica juncea* (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna), *Brassica napus* (New, Mar, Fortune, Drakker, Tapidor), *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and 4 synthetic hexaploids *Brassica*. No amplification was seen in *Brassica nigra* except HRIGRU010919, showing absence of these elements and a separate history (Figure 3.6a).

The amplification of *BrCOP5* revealed the A-genome specificity of the element. The primer pair BrCOP5F and BrCOP5R amplified 709 bp RT products from 23 *Brassica* accessions: the A-genome diploids and polyploids (AABB, AACC, AABBCC). All 6 *Brassica rapa* and 9 *Brassica juncea* amplified bands, while only 5 *Brassica napus* amplified the expected bands. No amplification from *Brassica oleracea* and *Brassica carinata* except 'NARC-PK' suggests its absence in C-genome. Two hexaploids out of 4 also generated the bands (Figure 3.6b). The amplification polymorphisms of *BrCOP9* and *BrCOP11* showed their A-genome specificity, where *Brassica rapa* and its polyploids amplified the products while C-genome lack these elements. Both elements amplified 26 products from *Brassica rapa* and its polyploids, with no amplification observed in most *Brassica oleracea* and *Brassica carinata* (Figure 3.6c & d). The amplification of 690 bp RT-products of *BrCOP12* revealed its diversity and abundance in almost all *Brassica* genomes including *Brassica nigra* (BB) genome. The expected product was amplified from all *Brassica* diploids and polyploids (40 cultivars), except *Brassica nigra* 'HRIGRU010978' (Figure 3.6e).

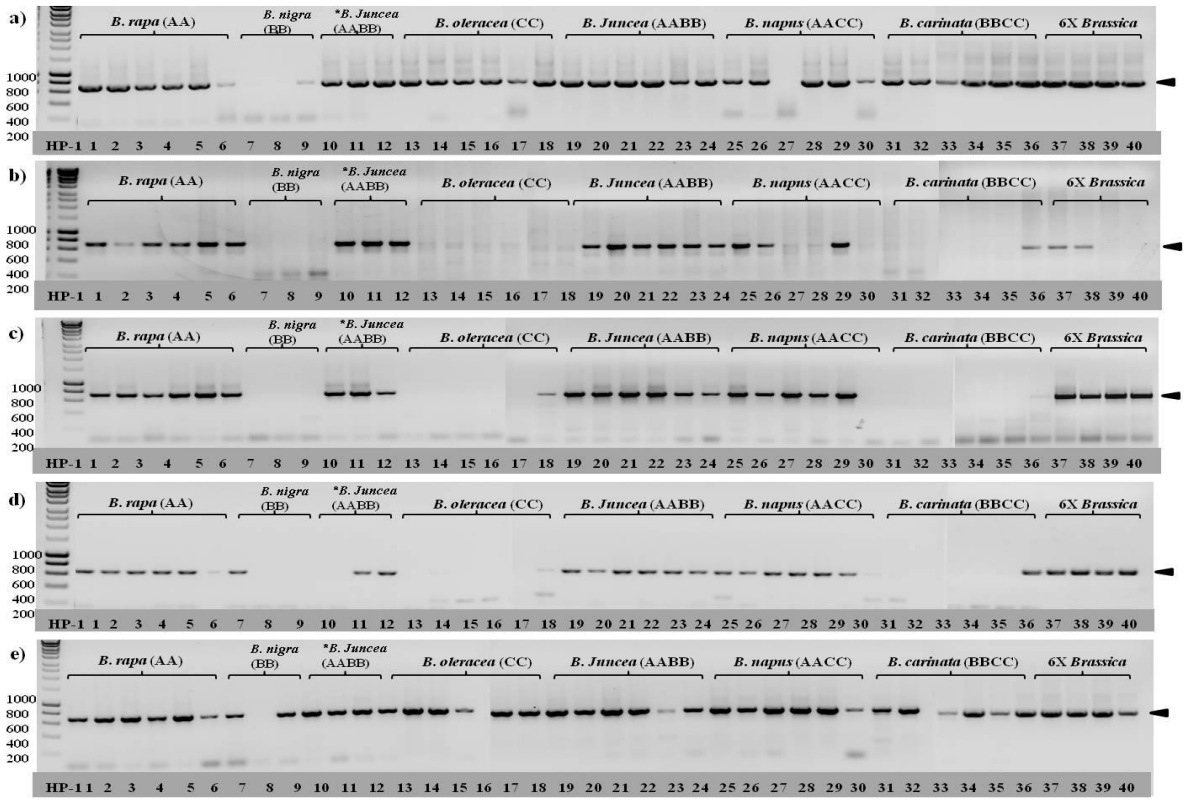


Figure 3.6: PCR analysis for the detection of Copia RT polymorphisms across 40 cultivars in *Brassica*. Dark arrow heads at right are indicating the expected product sizes. The amplification of a) *BrCOP2* b) *BrCOP5* c) *BrCOP9* d) *BrCOP11* e) *BrCOP12*. (PCR figures show reversed images of size-separated ethidium bromide-stained DNA on agarose gels after electrophoresis; ladders show fragments sizes in base pairs; numbers at the base indicate accessions of the species indicated from Table 2.1).

The C-genome specific Copia elements were also observed in our study. The amplification of 703 bp RT region of *BoCOP25* revealed its C-genome specific nature and was amplified by primer pair BoCOP25F and BoCOP25R. Out of 40 *Brassica* lines tested, the product was amplified from 30 lines. All the *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Cuor Di Bue Grosso, Precoce Di Calabria, GK97361), *Brassica napus* (New, Mar, Fortune, Drakker, Tapidor), *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and 4 hexaploids *Brassica* amplified the element. Weak bands were amplified from *Brassica rapa* (Hinona, Suttons) and *Brassica juncea* (NARC-I, Tsai Sim, W3, Giant Red Mustard) cultivars. No amplification in *Brassica nigra* suggests separate line of evolution (Figure 3.7a). The PCR amplification of *BoCOP27* and *BoCOP37* revealed their high abundance and distribution among *Brassica* crops. The elements amplified 39 and 36 products from a collection of 40 *Brassica* genomes. The *BoCOP27* is distributed in all *Brassica* except 2 *Brassica nigra* and a *Brassica napus* cultivar. The *BoCOP37* amplified the RT regions from all *Brassica* species except *Brassica nigra* (Figure 3.7b & c).

The amplification polymorphisms of *BoCOP44* by primers BoCOP44F and BoCOP44R (Table 3.3) showed its abundance and distribution in all *Brassica* species. This suggests that the elements are very ancient, which were present in a common ancestor before the separation of B and A/C genomes. Out of 40 cultivars tested, 38 yielded the product while only two (*Brassica rapa*; 'Hinona' and *Brassica nigra*; 'HRIGRU011011') failed to generate the products (Figure 3.7d). The amplification of 711 bp RT domain of *BoCOP51* indicated its random and patchy distribution among *Brassica* crops, with a high diversity and proliferation in *Brassica oleracea* and C-allele containing polyploids. All except 1 *Brassica oleracea* (Cuor Di Bue Grosso) amplified the expected products. Eight *Brassica juncea*, four *Brassica napus*, five *Brassica carinata* and two synthetic hexaploid *Brassica* cultivars also generated the RT bands suggesting the abundance and distributions among *Brassica* genomes. Only 2 *Brassica rapa* accessions (Chinese Wong Bok, Vertus) amplified the *BoCOP51* RT-domain showing weak product signals (Figure 3.7e).

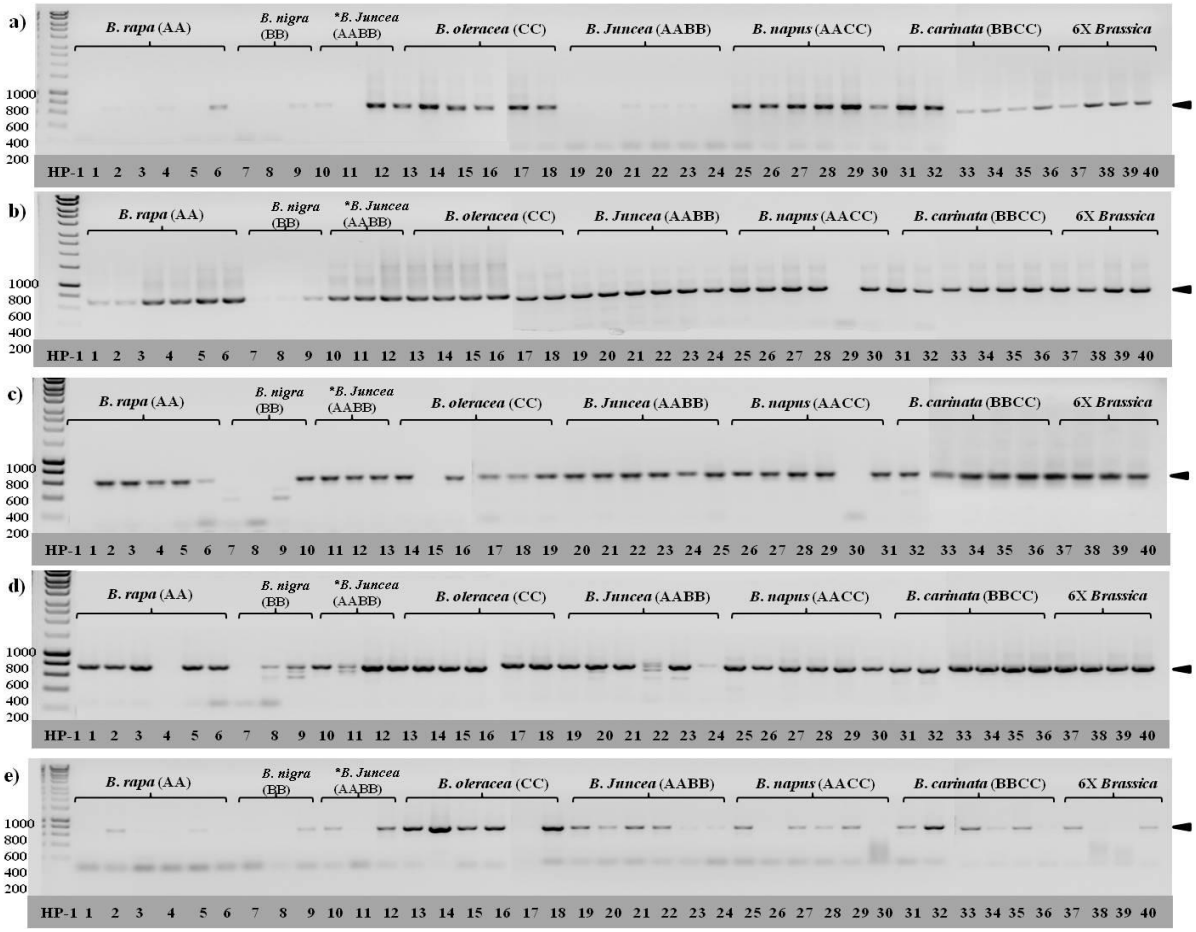


Figure 3.7: PCR analysis for the detection of Copia RT polymorphisms across 40 cultivars in *Brassica*. Dark arrow heads at right are indicating the expected product sizes. The numbers at the base are indicating the cultivars (Table 2.1). The amplification of a) *BoCOP25*; b) *BoCOP27*; c) *BoCOP37* d) *BoCOP44*; e) *BoCOP51*.

Table 3.3: List of Primers to amplify the RT regions of *Brassica* Copia, Gypsy and LARDs-like LTR retrotransposons. The expected product sizes and primers sequences are also given.

No.	Super-family	TE family	Product size	Primer name	Primer Sequence
1	Copia	*BrCOP2*	710	BrCOP2F	GACGTGGGAACTAGTGGAC
				BrCOP2R	CACTCTTGCTGTCTCGCATC
2	Copia	*BrCOP5*	709	BrCOP5F	TCTCTCTTCCGAAAGGCAAG
				BrCOP5R	TCCATGGGAATAGTGACTGG
3	Copia	*BrCOP9*	715	BrCOP9F	AGGGGAGTGGAGACAGGAG
				BrCOP9R	CCTTGGTGCCATATCAACCT
4	Copia	*BrCOP11*	650	BrCOP11F	CAGCTTTGCAATCTGTCATG
				BrCOP11R	GGGAATTCCAGGAGTTGAAG
5	Copia	*BrCOP12*	690	BrCOP12F	CATTGTTGGTTGCAGGTGGA
				BrCOP12R	CACATGGGTGTTGGCATAGG
6	Copia	*BoCOP25*	703	BoCOP25F	CATTGCACGATCCCATTCCG
				BoCOP25R	TGGGATCTCGTTGAACTACC
8	Copia	*BoCOP27*	720	BoCOP27F	ATGTCCACCAAGTGGAGTGC
				BoCOP27R	CAAAAGGAAAGAGAGCCTT
9	Copia	*BoCOP37*	722	BoCOP37F	TGAGCTCCACTGGTACATAG
				BoCOP37R	GGAGGTTGCTACTCTTCCTC
10	Copia	*BoCOP44*	715	BoCOP44F	AGGCAGAGGAGTAGGCATTG
				BoCOP44R	GGTGCCACCAACTGAAGATA
11	Copia	*BoCOP51*	711	BoCOP51F	GGATTACATTCTGCCATTCC
				BoCOP51R	CAGAACATGGGATCTCGTTG
12	Gypsy	*BoGYP1*	521	BoGYP1F	AATCACATGGCCCAAAAATC
				BoGYP1R	GGCCGAGTACTTCACTGTGG
13	Gypsy	*BrGYP5*	562	BrGYP5F	AGGTTACTCGGTGCAGGTTC
				BrGYP5R	TTCCTCGCTGTGTGACAATG
14	Gypsy	*BrGYP9*	598	BrGYP9F	AACCGCTTTAACCTTGTTAG
				BrGYP9R	GGTTCAAAGTCTGTTGGATG
15	Gypsy	*BrGYP12*	770	BrGYP12F	CCCCCTTCGAGATATACAGC
				BrGYP12R	AGAAAGAGGCAAGTCCGTGA
16	Gypsy	*BrGYP15*	421	BrGYP15F	CGAGCAATCAACAAGATAAC
				BrGYP15R	GTACTTCTGAAGCGCCGAAC
17	LARDs	*BoLAR3*	680	BoLAR3F	TCTATCGGTTTCCTGCAAGC
				BoLAR3R	TCTCTCAGCCAAGGAGAAAG
18	LARDs	*BrLAR5*	1295	BrLAR5F	CACGACGGAATCAATGTTTG
				BrLAR5R	GAACCGAAATTCGCACTGTC

3.2.5.5 Evolutionary relationship of *Brassica* Copia elements

The phylogenetic relationship of 138 Copia-RT sequences was performed by aligning the most conserved regions (Figure 3.8 & 3.14). The tree was generated by Neighbour-Joining method with 1000 bootstrap values and genetic distance was calculated with Jukes-Cantor model implemented in Geneious software. *Arabidopsis thaliana* Copia '*Araco*' was used to root the tree. Two major clades were found representing 33 and 105 sequences, which further showed 18 sub-clades or groups. Each sub-clade represents the clustering of sister families. The representatives of *BrCOP2* family clustered in one sub-clade. *BrCOP7* and *BrCOP13* clustered together in the same group. *BrCOP8*, *BrCOP9* and *BrCOP16* make sister families. *BrCOP1*, *BrCOP3*, *BrCOP11*, *BrCOP18* and *BoCOP26* clustered in the same group representing their respective families. *BrCOP4*, *BrCOP6* constitute sister families with *BoCOP32* and sharing the same group. *BrCOP5*, *BrCOP17* makes sister

families with *BrCOP37* family. The elements *BoCOP30* and *BoCOP43* share the same family. The largest group is represented by 15 retroelements, where *BrCOP14*, *BoCOP25/BrCOP25*, *BoCOP30*, *BoCOP33* and *BoCOP43* grouped together. *BoCOP23*, *BoCOP41*, *BoCOP42* and *BoCOP52* clustered in one group. *BoCOP45* and its homologues from *Brassica rapa BrCOP45* out grouped from other *Brassica* families and come closer to the *Arabidopsis* '*Araco*' element (Figure 3.8). Due to high homology in the RT regions of various Copia families, few families make their family specific group, while others were distributed in their respective clade. No species specific group was observed, indicating the presence of these elements in A and C-genome predating their separation.

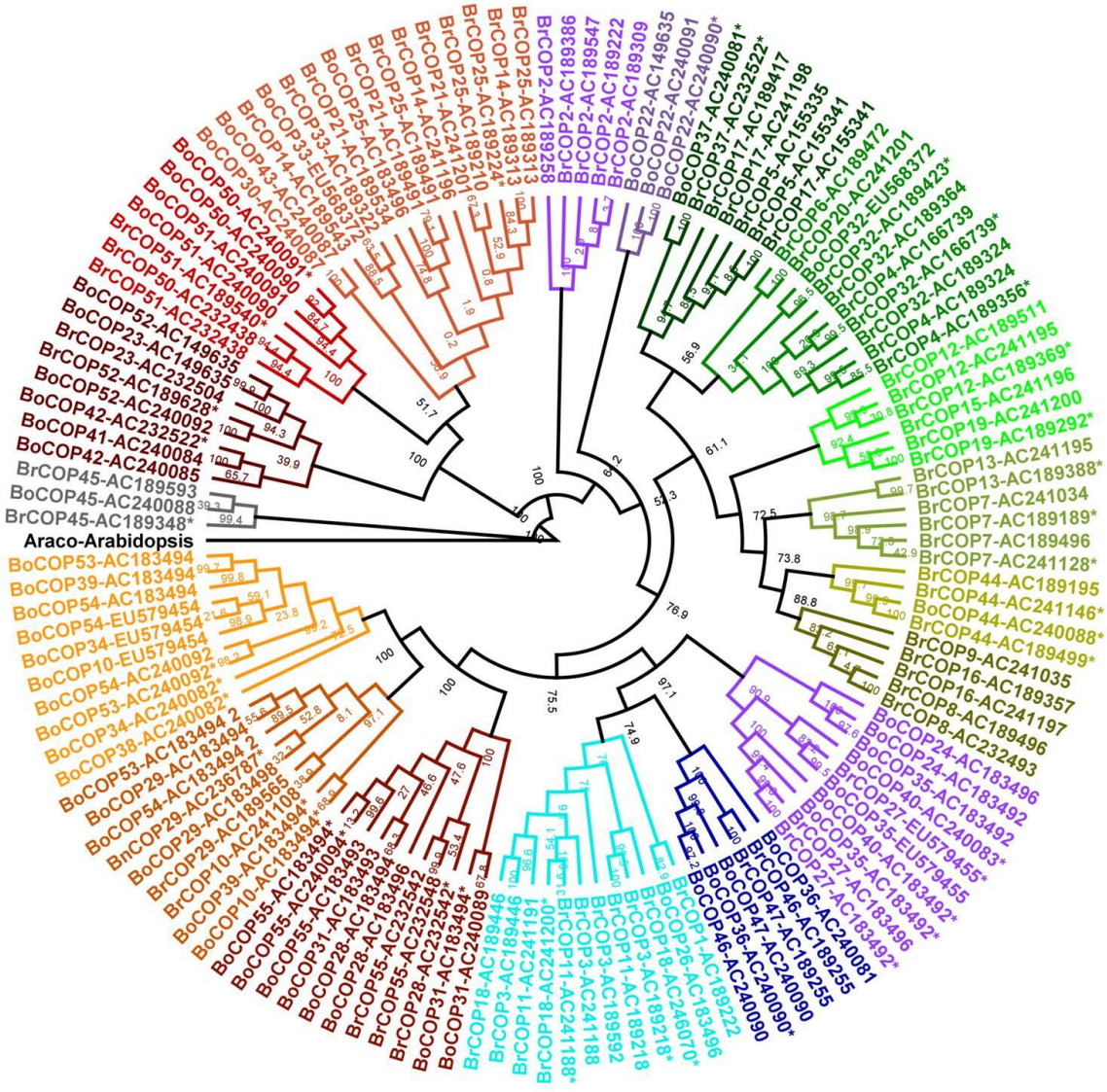


Figure 3.8: Phylogenetic analysis of 138 *Brassica* Copia-RT sequences. The tree was generated with Neighbour-Joining method implemented in Geneious Pro5.5.6. The tree is based on 1000 bootstrap values (% value shown at nodes) and a Jukes-Cantor model is used to calculate genetic distance. *Arabidopsis thaliana* Copia *Araco* was used to root the tree. Two major lineages split the elements into 18 clades shown by different colours. Br: *Brassica rapa*. Bo: *Brassica oleracea*. Bn: *Brassica napus.* COP: Copia.

3.3 Overview of Gypsy retrotransposons

Fifteen full lengths Gypsy retroelements were detected in *Brassica*. The elements range in sizes from 4.1 kb to 11.9 kb, with flanking LTRs ranging from 199 to 2035 bp. The 4.1 kb *BoGYP13* is a non-autonomous element, while the autonomous Gypsy elements range in sizes from about 5.0 kb up to 11.9 kb. Two major groups of elements can be distinguished on the basis of their sizes, one group representing the small sized elements (5.0-5.9 kb) and the other group represents large sized Gypsy (11.2-11.9 kb) elements (Table 3.1). Most elements have generated perfect and equally sized LTRs but in a few (*BoGYP1* and *BoGYP2*) variable sized LTRs were detected. This unequal size is due to the uneven activity of small repeat sequences in their 5´LTR. Almost all the elements have shown the perfect 5 bp TSDs, which in most cases are GC rich in contrast to AT rich Copia TSDs. With the exception of *BoGYP2, BoGYP13* and *BrGYP14*, all other are complete autonomous elements, showing the *gag-pol* protein domains. Around 95% of the Gypsy elements encode the PBS and PPT in their internal sequences downstream and upstream to 5´LTRs and 3´LTRs respectively.

The knowledge about the diversity of Gypsy in *Brassica* was further extended by using initially identified 15 reference elements as query in blast searches to find the total numbers of full length elements and their copies. Around 2324 hits were received in BLASTN searches, of which 56 were intact elements, 17 were truncated copies, 103 partial segments, 39 solo LTRs and 2109 remnants. The copy numbers of intact elements for *Brassica rapa* and *Brassica oleracea* Gypsy elements were estimated, which were 540 and 780 respectively in the total genomes (Figure 3.3).

3.3.1 Characterization and structural features of Gypsy superfamily

Although less abundant in comparison to Copia, Gypsy elements make a major proportion of *Brassica* genomes. The sizes of Gypsy elements were found to be 2 fold larger than the Copia elements with largest elements 11.8 kb (*BoGYP3*) in size while the non-autonomous (*pol* region deleted) *BoGYP13* is only 4.1 kb large in size. A Gypsy was identified from *Brassica oleracea* accession ‘AC240090.1’, named as *BoGYP1*. The structure of *BoGYP1* is about 9.1 kb in size, flanked by 2035 bp 5´LTR and 2004 bp 3´LTR. The LTRs from this element are considered to be the largest LTRs in *Brassica* genome in present study

(Figure 3.9; Table 3.1). *BoGYP1* displays a PBS complimentary to $tRNA_{Met}$ towards the downstream of 5′LTR and a PPT adjacent to 3′LTR, typical Gypsy-like *gag-pol* polyproteins structures 5′-TLC-GAG-AP-RT-RH-INT-3′, where an additional TLC domain is integrated downstream to PBS. A defective element (*pol* region deleted) *BoGYP2* is about 11.3 kb in size, flanked by 1272 bp 5′LTR and 1140 bp 3′LTR. The element has shown the deleted *pol* gene and lacking RT, RH and INT domains (Table 3.1). *BoGYP3*, *BrGYP4* and *BrGYP5* are 11.8, 11.7 and 11.8 kb in size with 471-480 bp LTRs flanking the internal region. The internal regions from the elements display PBS with complimentary to $tRNA_{Lys.}$ The typical *gag-pol* organization of non-chromodomain bearing Gypsy (5′-GAG-AP-RT-RH-INT-3′) is studied. All the elements display a PPT of 15 bp adjacent to their 3′LTRs (Table 3.2). *BoGYP6* and *BrGYP7* are about 11.5 and 11.6 kb long elements, flanked by 509 and 506 bp LTRs respectively. *BoGYP6* is identified from a *Brassica oleracea* accession, while *BrGYP7* from a *Brassica rapa* accession. They have a PBS which use $tRNA_{Lys}$ for RNA replication next to their 5′LTRs and a PPT of 15 bp upstream to the 3′LTRs. The typical *gag-pol* organization of non-chromodomain Gypsy 5′-GAG-RT-RH-INT-3′ is observed in the elements, where an unknown domain region was identified from *BoGYP6* (Figure 3.9; Table 3.2).

The structural features of chromodomain (CHR) bearing elements showed relative similarity in their structural features. *BrGYP1, BrGYP8, BrGYP9, BrGYP10, BrGYP11* and *BrGYP12* belong to the chromoviral branch of Gypsy LTR retrotransposons. *BrGYP8* is a 5.1 kb element, flanked by 331 bp LTRs and an internal domain displaying typical PBS complementary to $tRNA_{Phe}$, Open Reading Frames (ORFs) for *gag-pol* polyproteins as 5′-GAG-AP-RH-INT-CHR-3′, where RT is lost during a recent rearrangement phase. The genome of *BrGYP9* and *BrGYP10* are 5.9 and 5.2 kb, flanked by 346 bp LTRs. They represent similar PBS, typical chromoviral *gag-pol* genes organization 5′-GAG-AP-RT-RH-INT-CHR-3′ and a 15 bp homologous PPT adjacent to 3′LTR. Their similar structural features suggest that they are sister elements belonging to the same family. *BrGYP11* and *BrGYP12* have shown homologies in their genomic structures. They are 5.1 kb large is size, including the LTRs of 340-360 bps. They are characterized by the presence of a PBS complementary to $tRNA_{Arg}$ and $tRNA_{Met}$ respectively, with the ORFs for the canonical *gag-pol* genes presenting 5′-GAG-AP-RT-RH-INT-CHR-3′, where a CHR is absent in *BrGYP12* during the evolutionary scenario. A PPT strand composed of similar nucleotides in both elements indicates their close relationship and a common ancestor. Two non-

autonomous Gypsy *BoGYP13* and *BrGYP14* were identified from *Brassica oleracea* and *Brassica rapa* respectively. They are about 4.1 and 7.2 kb large elements including LTRs of 200 and 1553 bp respectively. The internal region of *BoGYP13* represents PBS and a PPT downstream and upstream to the 5´LTR and 3´LTR respectively, but no recognizable PBS is detected in *BrGYP14* (Figure 3.9). The PPT of both elements is a 15 bp segment highly rich in AT contents. Although they have typical Ty3/gypsy-like ORFs for the *gag-pol* genes but their *pol* polyproteins lost the RT, RH and INT domains in rearrangements during the ancient evolutionary period (Table 3.2).

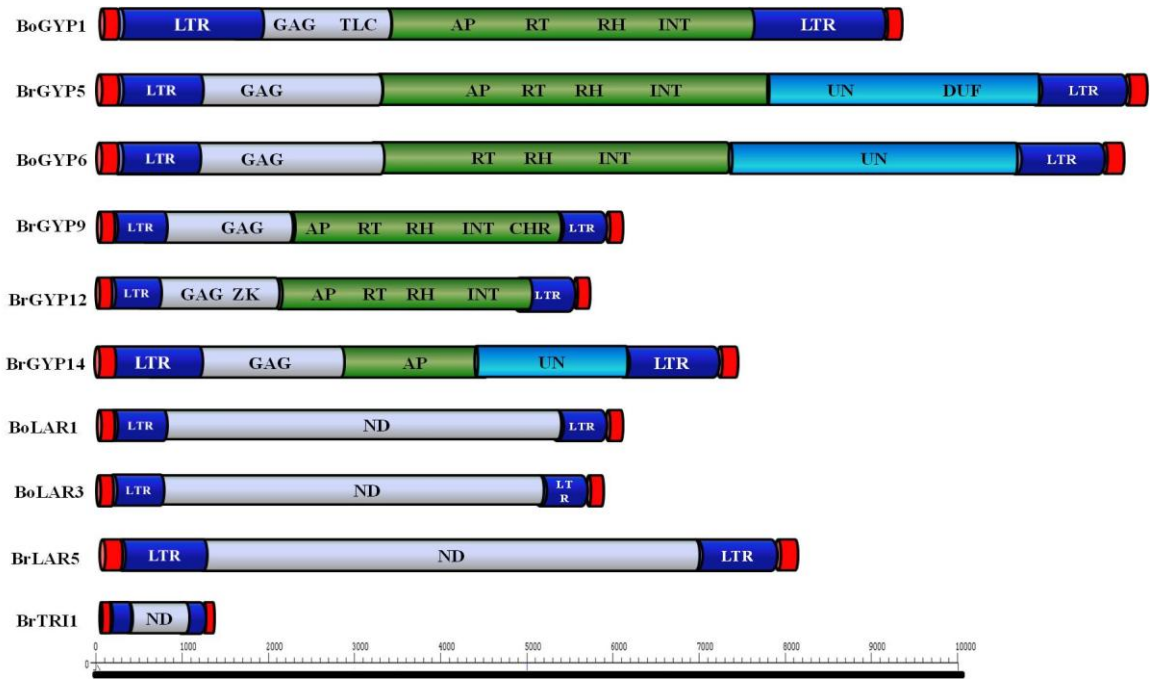


Figure 3.9: Schematic representation of structures of Gypsy, LARD and TRIM example elements in *Brassica*. The red discs at the ends represent the TSDs, dark blue indicates LTRs. The *gag* and *pol* regions are drawn with their protein domains. The scale below measures lengths of the elements (bp). Additional insertions or unknown sequences are highlighted by light blue. AP: Aspartic protease. RT: reverse transcriptase. INT: integrase. GAG: *gag*-nucleocapsid. ZK: zinc knuckle. DUF: domain of unknown function. CHR: Chromatin organization modifier. UN: unknown. ND: not detected.

3.3.2 Domain organization in intact Gypsy elements

The organization of *gag* and *pol* genes coding protein domains were studied in all the 15 Gypsy elements. All the Gypsy elements (100%) have encoded the *gag* gene. Five different types of domain organizations were observed in Gypsy elements. The canonical Ty3/gypsy *gag-pol* organization is 5´-GAG-AP-RT-RH-INT-3´, which was observed in

35% of the elements investigated. Gypsy having a chromodomain was the second abundant group covering 28.5% of the total elements with 5′-GAG-AP-RT-RH-INT-CHR-3′ domain organization. The non-autonomous Gypsy elements were also identified encoding a *gag* protein but only AP domain from *pol* gene as 5′-GAG-AP-3′. Remaining elements incorporate one or other extra protein in *gag* or *pol* genes. In *BoGYP1*, an extra protein motif called TLC domain is present upstream to the *gag* protein as 5′-TLC-GAG-AP-RT-RH-INT-3′. Two extra domains such as ZK and Zinc finger (ZF) are present after and before GAG and INT domains respectively (5′-GAG-ZK-AP-RT-RH-ZF-INT-3′) in *BoGYP12* element (Table 3.2).

3.3.3 PBS and PPT motifs of Gypsy elements

The PBS and PPT primers necessary for RNA amplification for Gypsy retrotransposons were detected by scanning them against tRNA database using parameter 'Predict PBS by using *Arabidopsis thaliana* tRNA database' in LTR_FINDER. A total of 93% elements showed the presence of 14-21 bp PBS downstream to the 5′LTR. *BoGYP14* have shown no signs of PBS with *Arabidopsis thaliana* and *Zea mays* tRNA database'. Six different tRNA types were observed in all Gypsy elements investigated in the present study. The most frequent tRNA type in Gypsy elements was $tRNA_{Lys,}$ detected in 35% of the elements; the second important type was $tRNA_{Tyr,}$ observed in 20% of the investigated elements. Generally the $tRNA_{Met}$ is the most frequent type present in LTR retrotransposons but here only 15% of the elements showed this tRNA type. All the other 3 types of tRNA contributed only 7% each of the tRNA type. PPT adjacent to the 3′LTR was detected in 93% of all Gypsy elements investigated in this study. *BoGYP2* was the only element, where no PPT was detected indicating its deletion. All the other elements have 15 bp PPT sequence towards the upstream of 3′LTR. The PBS and PPT sequences and their positions in the retrotransposons at 5′LTR and 3′LTR respectively are tabulated (Table 3.2).

3.3.4 Analysing diversity and distribution of Gypsy elements by RTAP markers

The diversity and distribution of Gypsy retrotransposons in *Brassica* genomes was investigated by RTAP method using 5 primers pairs (Table 3.3). Although less in numbers compared to Copia, Gypsy elements showed high diversity and distribution among

Brassica genomes. The primer pair BoGYP1F and BoGYP1R was designed to amplify the conserved 521 bp RT region of *BoGYP1*. The RT regions were amplified from all the 40 *Brassica* cultivars including *Brassica rapa*, *Brassica nigra*, *Brassica oleracea*, *Brassica juncea*, *Brassica napus*, *Brassica carinata* and four synthetic hexaploids *Brassica* (Figure 3.10a). The amplification of *BoGYP1* family from A, B, and C-genome *Brassica* suggests its common ancestry. The insertional polymorphism of *BrGYP5* also showed same pattern, where it is amplified from all 40 *Brassica* cultivars (Figure 3.10b).

The amplification of chromodomain containing Gypsy elements were also investigated among *Brassica* species. Using BrGYP9F and BrGYP9R, a 598 bp amplicon was amplified from 36 out of 40 *Brassica* lines tested. All *Brassica rapa* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons), *Brassica juncea* (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna), *Brassica napus* (New, Mar, Last And Best, Fortune, Drakker, Tapidor), *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and hexaploid *Brassica* cultivars amplified the expected product. *Brassica nigra* also amplified the product except accession 'HRIGRU010978', whereas three *Brassica oleracea* (De Rosny, Precoce Di Calabria, Cuor Di Bue Grosso) accessions amplified the *BrGYP9* RT regions (Figure 3.10c).

The polymorphisms of *BrGYP12* revealed its distribution among all the six diploids and polyploids *Brassica* species from 'triangle of U' and their cultivars used in present study. By using BrGYP12F and BrGYP12R, 770 bp RT regions were amplified from all *Brassica* cultivars tested. *Brassica nigra* also amplified the products, although the signals were weak as compared to A and C-genome *Brassica* (Figure 3.10d). This confirmed the ancient nature of element and predicts their presence before the separation of B and A/C-genomes (~9 Mya). Similarly *BrGYP15* yielded the 421 bp RT domains from all *Brassica* except *Brassica nigra* (HRIGRU011011) genomes (Figure 3.10e). The RTAP method used for the amplification of Gypsy elements revealed their diverse nature and distribution among all *Brassica* species and also indicated their ancient nature and common phylogeny predating the separation of A, B and C-genomic *Brassica*.

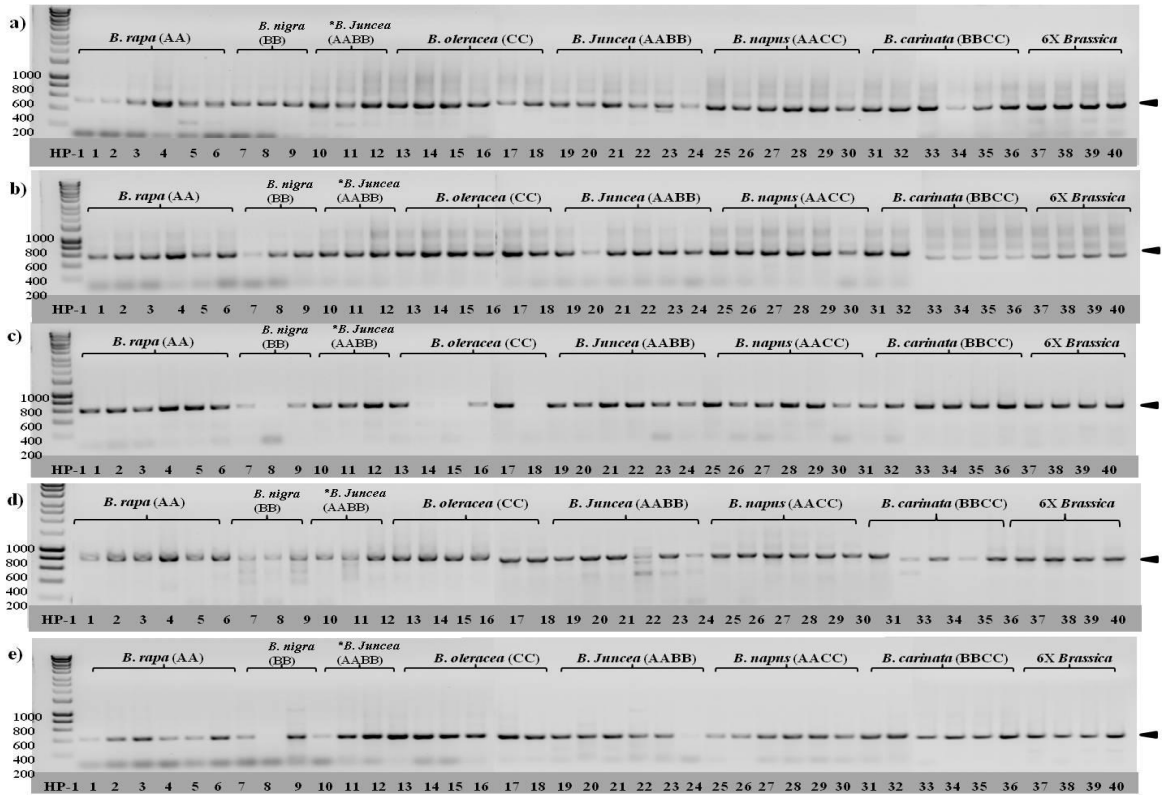


Figure 3.10: PCR analysis showing fragments with and without Gypsy RT regions between the primers. DNA samples were obtained with primers hybridizing to conserved RT regions of various Gypsy families. Dark arrow heads (right) indicate expected product sizes. Numbers underneath indicate accessions (Table 2.1). The amplification of a) *BoGYP1*; b) *BoGYP5*; c) *BrGYP9*; d) *BrGYP12*; e) *BrGYP15*.

3.3.5 Phylogenetic analysis of *Brassica* Gypsy RT segregated two major groups

The phylogenetic analysis of 40 RT domains from *Brassica* Gypsy elements were performed by Neighbor-Joining method with 1000 bootstrap replicates. *Arabidopsis* Gypsy element 'Tat4-1' was used to root the tree. The tree is generated by (~180 amino acids residues) from most conserved D-DD triad of RT (block3-5). Clustering of *Brassica* Gypsy into two major clades were observed, clearly separating the chromodomain bearing Gypsy (Chromoviruses) and non-chromodomain Gypsy elements. Three clades from chromodomain bearing and 4 from non-chromodomain Gypsy are distinct. The members from *BoGYP1*, *BrGYP9*, *BrGYP10*, *BrGYP11* and *BrGYP12* share one clade representing the Chromoviruses-like elements, while *BrGYP3*, *BrGYP4*, *BrGYP5*, *BrGYP6* and *BrGYP7* come together in other group making the large clade of non-chromodomain holding Gypsy. In first major clade, *BrGYP9* and *BrGYP10* clustered in one, *BrGYP11* and *BrGYP12* in other and *BoGYP1* in a third sub-clade. The second major clade have also shown 3 sub-clades and groups, where *BrGYP3*, *BrGYP4*, *BrGYP5*, *BrGYP6* and *BrGYP7*

elements are dispersed suggesting their common ancestry. *BrGYP3*and *BrGYP7* are phlogenetically close to each other representing the same family. *BrGYP4*, *BrGYP5* and *BrGYP6* develop sister families, sharing lot of homology in their coding regions (Figure 3.11).

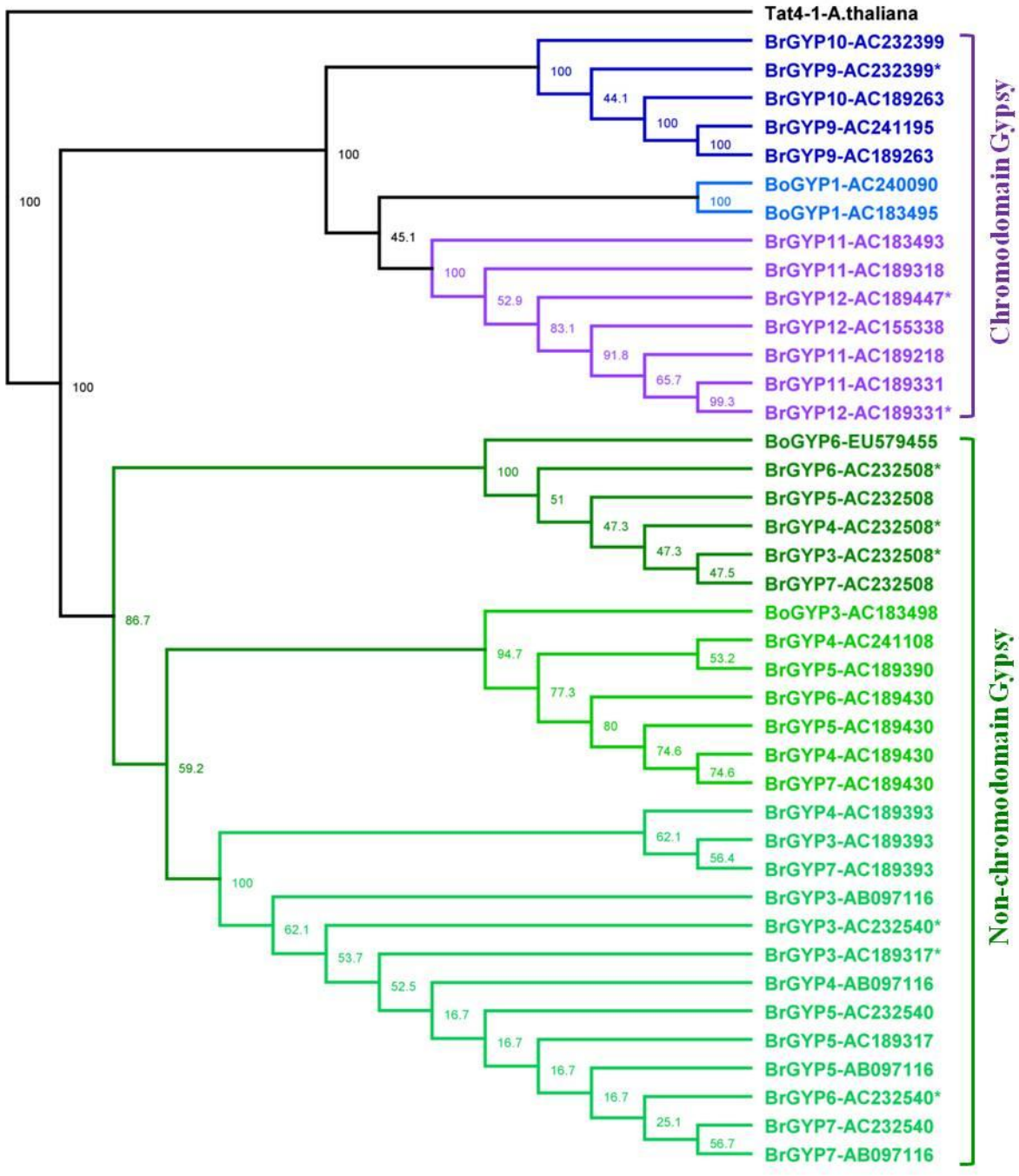


Figure 3.11: Phylogenetic tree of 40 *Brassica* Gypsy elements. Phylogenetic relationships of the Gypsy retrotransposons based on the amino acid alignment of the conserved RT domains (~180 aa). The two main lineages separate the chromodomain containing group from non-chromodomain group. Three clades from chromodomain bearing and four from non-chromodomain Gypsy are distinct represented by different colours or shades. The N-J bootstrap values supporting the internal branches are indicated at the nodes. The tree is out-grouped with *Arabidopsis thaliana* Tat4-1 element. Br: *Brassica rapa*. Bo: *Brassica oleracea*. GYP: Gypsy.

3.4 Diversity of Large Retrotransposon Derivatives (LARDs) in *Brassica*

The identification of the LTR retrotransposons led to the detection of elements that lack coding capacity for *gag-pol* genes, but acquired the LTRs, PBS and PPT. They range in size from 3.8 to 8.0 kb, flanked by the LTRs ranging from 231 to 1319 bp. Due to structural similarities with LARDs-like elements studied in other plants; these elements were considered as members of the LARDs. Six intact (reference) elements from LARDs family were identified by dot plot analysis and investigated their copy number by BLASTN searches in NCBI database. A total of 1007 copies were found against the *Brassica* Nucleotide Collection (nr/nt) database, out of which 16 are full length elements, 08 truncated copies, 39 partial elements, 12 solo LTRs and 932 remnants. The small dispersed fragments (remnants) cover 92% in number but in size they are less than the size of intact copies. The copy numbers of full length elements were estimated from whole genome of *Brassica rapa* and *Brassica oleracea*, which were 110 and 760 respectively (Figure 3.3). The *gag-pol* protein coding regions were investigated, with no recognizable domain. The elements were also investigated for any PBS and PPT towards the downstream and upstream of 5´LTRs and 3´LTRs respectively. PBS motif was detected in 50% and PPT motif in 65% of the elements.

3.4.1 Structural features of LARDs-like elements

A 6.2 kb long element was identified from the *Brassica oleracea* accession 'AC149635.1'. The element is named *BoLAR1*, which generates 5´-313/322-3´ bp LTRs, and has perfect 5 bp TSDs (Figure 3.9). No recognisable PBS was observed by scanning the sequence against *Arabidopsis thaliana* and *Zea mays* tRNA databases. A 15 bp PPT was detected adjacent to the 3´LTR from 5842-5856 bp within the element (Table 3.2). A similar element *BoLAR2* was detected from *Brassica oleracea* accession 'AC183498.1'. *BoLAR2* is a 6.0 kb large element, generates 5 bp TSDs and flanked by 231 bp LTRs. The internal region shows no identifiable *gag-pol* polyproteins, and no sign of any PBS or PPT. Two similar elements *BoLAR3* and *BrLAR4* were detected in *Brassica oleracea* and *Brassica rapa* accessions. *BoLAR3* and *BrLAR4* are 5.8 and 5.6 kb large elements, flanked by 5´-707/720-3´ bp and 666 bp respectively. They both exhibit PBS motif complementary to $tRNA_{Met}$ towards the downstream of 5´LTR and a 15 bp PPT region upstream to 3´LTR. The PBS and PPT sequences in both elements are exactly similar indicating them the

sister elements. The screening of *Brassica rapa* accession 'AC241138.1' for LTR retrotransposons led to the identification of a nearly 8.0 kb large element designated as *BrLAR6*. It is flanked by large LTRs of 1319 bp with non-coding internal region of 5.4 kb, typical characteristics of LARDs-like elements. The smallest LARDs-like element was identified from the *Brassica rapa* 'AC241195.1', which is only 3.8 kb in size including the LTRs of 347 each on both terminals and a internal non-coding region of 3.1 kb.

3.4.2 PCR amplification of LARDs elements

The diversity of LARDs elements were studied among *Brassica* genomes, where they were found abundantly distributed in *Brassica* species. The degenerative primer pair BoLAR3F and BoLAR3R (Table 3.3) was designed from LTRs to amplify 680 bp LTR regions. The amplification was achieved in *Brassica rapa* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons), *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361), *Brassica juncea* (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3), *Brassica napus* (New, Mar, Last And Best, Fortune, Drakker), *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and 4 hexaploid *Brassica* cultivars (Figure 3.12a). The amplification of 1295 bp LTR regions from *BrLAR5* showed high diversity and distribution among *Brassica* species. All 40 cultivars yielded the products with two bands of the same size indicating the multiple copies on both alleles. The three *Brassica nigra* also yielded the product with one additional band of ~1200 bp (Figure 3.12b).

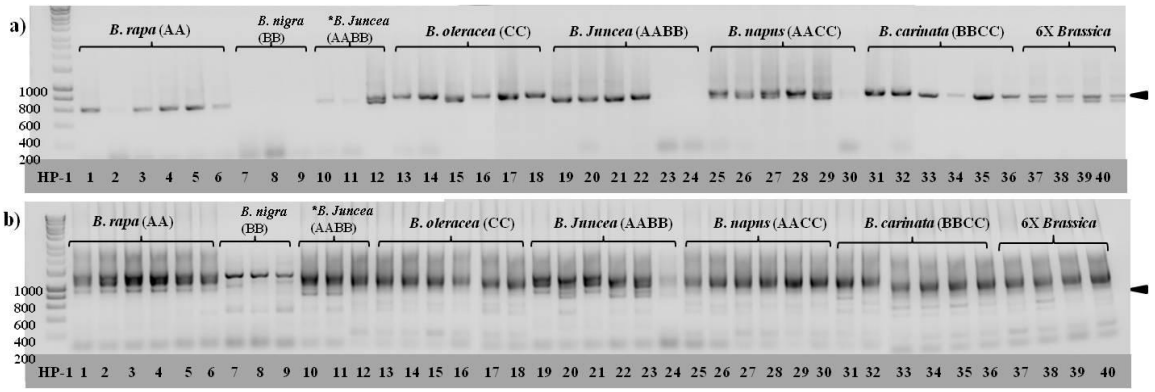


Figure 3.12: PCR amplification of LARD-like elements. DNA samples were obtained with primers hybridizing the conserved LTR regions of LARD-like elements. Dark arrow heads at right indicate expected product sizes. The numbers underneath indicate cultivars (Table 2.1). Amplification of a) *BoLAR3*; b) *BoLAR5*.

3.5 Terminal-Repeat Retrotransposons in Miniature (TRIM)

TRIM are small elements with TSDs, LTRs and short internal regions (Witte et al., 2001). They can be differentiated from LARDs on the basis of their small sizes. A family of TRIM was identified from *Brassica rapa* and named *BrTRI1*. Two complete and 1 truncated copies were retrieved from GenBank database and 25 copies were estimated in *Brassica rapa* genome. *BrTRI1* is 1323 bp in size including 257 bp 5′ and 362 bp 3′-LTRs. The internal region is 800 bp only which are highly AT rich. The element in general in AT rich (66%) and contain poly(T) repeats dispersed in its central region. It posses PBS and PPT motifs but lack internal *gag-pol* protein domains (Figure 3.9).

3.6 Discussion

3.6.1 LTR retrotransposons are highly diverse and abundant in *Brassica* crops

The LTR retrotransposons are highly abundant in plants including *Brassica*. A total of 206 Copia, 56 Gypsy, 16 LARDs and 2 TRIM were collected from A (51.3 Mbp) and C-genome (4.7 Mbp) *Brassica* from available GenBank database. From *Brassica rapa*, 148 Copia, 50 Gypsy, 10 LARDs and 25 TRIM were retrieved, while 58 Copia, 6 Gypsy and 6 LARDs were collected from *Brassica oleracea*. A total of 1596 Copia, 540 Gypsy, 110 LARDs and 25 TRIM were estimated for *Brassica rapa*, while 7540 Copia, 780 Gypsy and 760 LARDs with no TRIM were estimated for *Brassica oleracea* whole genomes. Collectively, 11351 intact copies of LTR retrotransposons including LARDs and TRIM were estimated from *Brassica rapa* and *Brassica oleracea* (Figure 3.3).

3.6.2 LTR retrotransposon landscape in different plant genomes

The PCR analyses revealed the distribution of elements among various *Brassica* species. The majority of the elements were amplified from all *Brassica* species including *Brassica nigra*, while a few elements were found proliferating in A or C-genome alleles. (There are few sequenced BACs from the B-genome to analyse, so the method would not be expected to find A-genome specific sequences). The abundance and diversity of Copia retrotransposons are studied in several plants genomes as conifers (Friesen *et al*., 2001), wheat, barley, rice and *Arabidopsis* (Wicker and Keller, 2007; Tsukahara *et al*., 2009),

wheat (Tomita *et al.*, 2010), rice (Vicient and Schulman, 2002), sugarcane (Muthukumar and Bennetzen, 2004), oil palm (Price *et al.*, 2002), sunflower (Kawakami *et al.*, 2010), sugar beet (Schmidt *et al.*, 1995), jute (Ahmed *et al.*, 2011), grapevine (Moisy *et al.*, 2008), melon (Ramallo *et al.*, 2008), tomato (Tam *et al.*, 2007; Cheng *et al.*, 2009), *Medicag*o (Wang and Liu, 2008), cassava (Gbadegesin *et al.*, 2008) and several other plants. The Gypsy elements are also actively proliferating in plant genomes and showed their diversity and abundance in several plants like wheat (Tomita *et al.*, 2010; Salina *et al.*, 2011), sorghum (Muthukumar and Bennetzen, 2004), jute (Ahmed *et al.*, 2011), citrus (Bernet and Asins, 2003), soybean (Du *et al.*, 2010), pepper and tomato (Park *et al.*, 2011), tomato (Peters *et al.*, 2009), chickpea (Rajput and Upadhyaya, 2009), *Glycine max* (Yano *et al.*, 2005) and *Arabidopsis* (Tsukahara *et al.*, 2009). This suggests the diversity, abundance and distribution of retrotransposons and their role in genome size duplication and diversification of plant genomes.

A small number of inconsistent results were found in the RTAP retrotransposon insertion assays, where one or other accession did not include an element present in many other accessions. This could result from mutation in the primer sites, or excision of this genomic region in some accessions. It would be interesting to explore the genomic structure further in the accessions showing no amplification using more distal primers to the insertion to see if there was a different structure in these accessions, which may have arisen as a consequence of the transposon's presence. A few *Brassica* accessions showed unexpected amplification of transposons (Figure 3.6b & d; 3.7b & c), where most did not amplify or the ancestral diploids did not include the element. For example, one *Brassica oleracea* and a *Brassica carinata* NARC-PK amplified one element (Figure 3.6d). This could be because of the contrasting origin of the accessions showing phytogeographical polymorphisms, with some regions including lines with elements. For example, *Brassica carinata* 'NARC-PK' originated from Pakistan around 4,500 miles from the European accessions. It is also notable that the genomic and chromosomal constitution of the diverse *Brassica* accessions has not been studied in detail.

3.6.3 Reverse transcriptase is the most conserved region in LTR retrotransposons

The reverse transcriptase (RT) of 137 *Brassica* and 23 other plants Copia were collected and investigated for their most conserved regions. A 'YVDD' motif is found as the most

conserved motif in 98% of Copia elements as previously studied in plants (Flavell *et al*., 1992b). The region around this 'YVDD' signature is most conserved among all Copia elements, with few other conserved regions dispersed in RT. In Gypsy elements the 'YVDD' motif is observed in nearly half the elements, while others have 'YNDD' signature, where N is any other amino acid. Other conserved regions were found in Gypsy elements from *Brassicaceae* and other plants. The aspartic acid residue (DD) is most conserved motif observed in almost 100% LTR retrotransposons aligned (Figure 3.13 & 3.14). The detail investigations of *gag-pol* internal domains of various superfamilies of LTR retrotransposons indicated that the RT is most conserved region among all retrotransposons and DD motif is shared by all superfamilies (Hansen and Heslop-Harrison, 2004). The most conserved nature of RT among various superfamilies is confirmed by several other workers. The analysis of 82 RT sequences from various organisms confirmed the conserved nature of RT, where seven common blocks were observed suggesting the highly conserved nature of RT. The analysis also showed that aspartic acid (DD) motif is present in all the sequences aligned (Xiong and Eickbush, 1990).

3.6.4 LARDs lack internal coding regions but are active elements

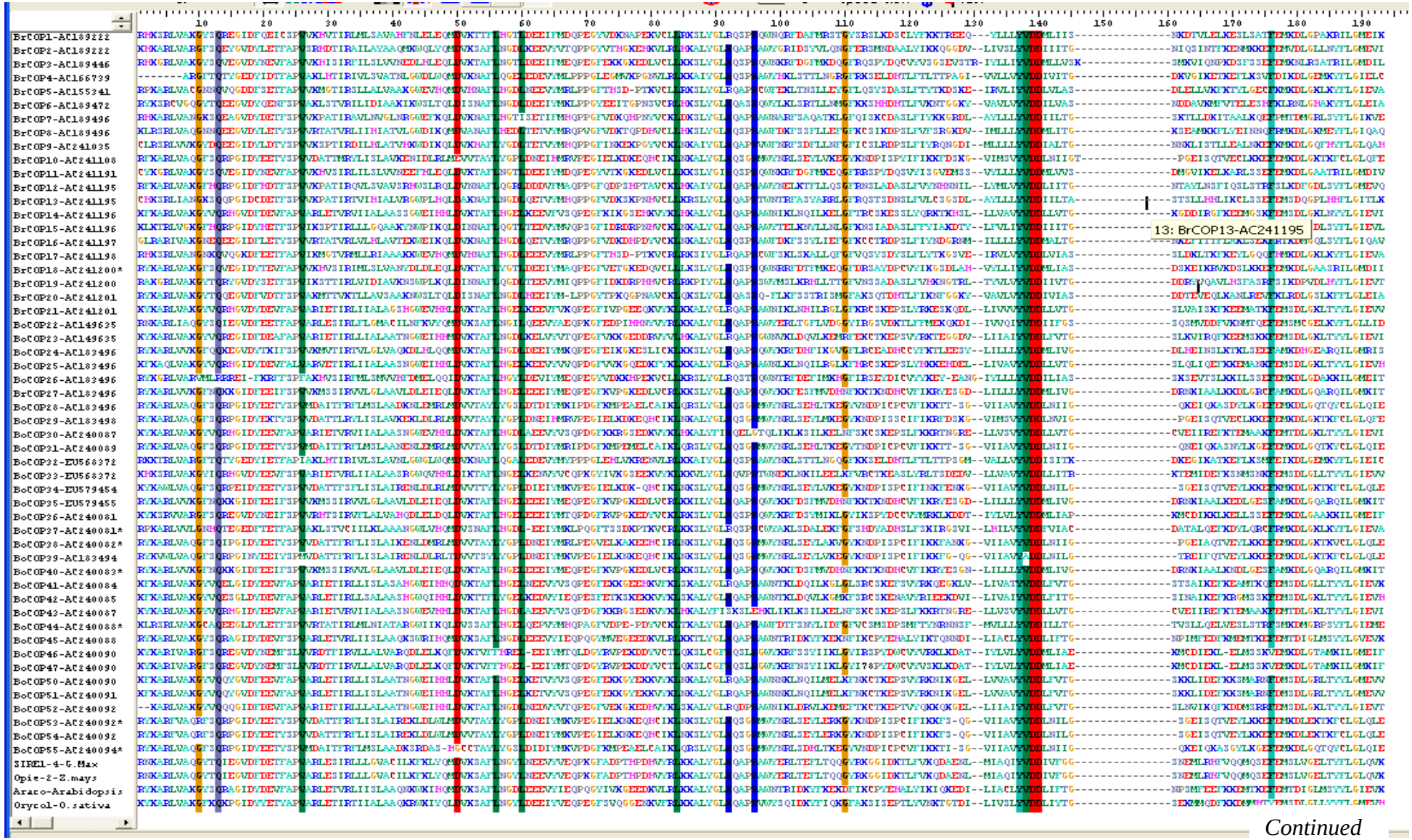
Several copies of LARDs-like elements were detected in *Brassica* and amplified bp PCR. No single element has shown any *gag-pol* protein domains in their internal regions. The structural analysis in the members of *Triticaceae* revealed this fact that LARDs are non-autonomous retrotransposons (Kalendar *et al.*, 2004). Despite of lacking their internal coding domains, many active copies were found in *Brassica* genomes. We cannot fully resolve the question which LTR retrotransposon class these LARDs belong to and which superfamily or family is borrowing them their coding domains for transposition and integration to a new site. But the comparison of the elements with known TE sequences in Repbase database indicate that *BoLAR1* and *BoLAR2* have shown ~40% homology to the *Arabidopsis thaliana* Copia elements. In other LARDs, more homology was observed with Gypsy elements on the basis of LTR sizes, PBS and PPT sequences. Around 110 and 760 elements from *Brassica rapa* and *Brassica oleracea* respectively (Figure 3.3) were estimated. The study revealed that LARDs-like elements are actively proliferating in plant genomes as identified in barley and members of *Triticaceae* (Kalendar *et al.*, 2004).

3.6.5 TRIM are less active LTR retrotransposons

Only 2 copies of a TRIM family were identified and estimated 25 copies from *Brassica rapa* whole genome. No TRIM-like elements were detected from *Brassica oleracea* suggesting their less abundance in comparison to Copia, Gypsy and LARDs investigated in present study. Only 43 TRIM-like elements belonging to three groups (Katydid-AT1, Katydid-AT2 and Katydid-AT3) were identified from *Arabidopsis* (Witte *et al.*, 2001), whereas, only three TRIM-like elements were identified from apple (Antonius-Klemola *et al.*, 2006). The estimated copy numbers for 4 families (*Br1*, *Br2*, *Br3* and *Br4*) of TRIM in *Brassica rapa* and *Brassica oleracea* are 530 and 660 respectively (Yang *et al.*, 2007).

3.7 Conclusion

The present study has increased the knowledge about the characterization, diversity, distribution, mobilization and evolutionary impacts of *Brassica* retrotransposons. Studying the characteristics of the different families, it was observed that several families are still autonomous and active with 1–16 copies encoding a single putatively functional *gag-pol* polyprotein. To our knowledge, this study is the first extensive and detailed compilation of LTR retrotransposons landscape of the *Brassica* genome. The results enable identification and understanding of the structure and nature of full length elements and their derivatives, including TSDs. The BAC-based approach does not rely only on conserved protein domains most often analysed, and it also ensures that all the families studied have shown activity during their recent evolutionary history within the *Brassica* genus. The markers derived here will be useful for examining chromosome and genome evolution in *Brassica.* In the future, it will be important to study B-genome derived BACs in a similar way to identify elements in this genome. It will also be valuable to examine many of the 'wild' *Brassica* species outside the U triangle, and other related genera, to see the value of the RBIP-type insertional polymorphism markers for identifying alien chromosome and alien genome introgression. These lines are being exploited to transfer new variation into crop *Brassicas* (Ge *et al.*, 2009) and the identification of the alien chromosomes and particular of introgressed segments using robust and potentially genome-wide markers is critical to directing the exploitation of these valuable lines.



Continued

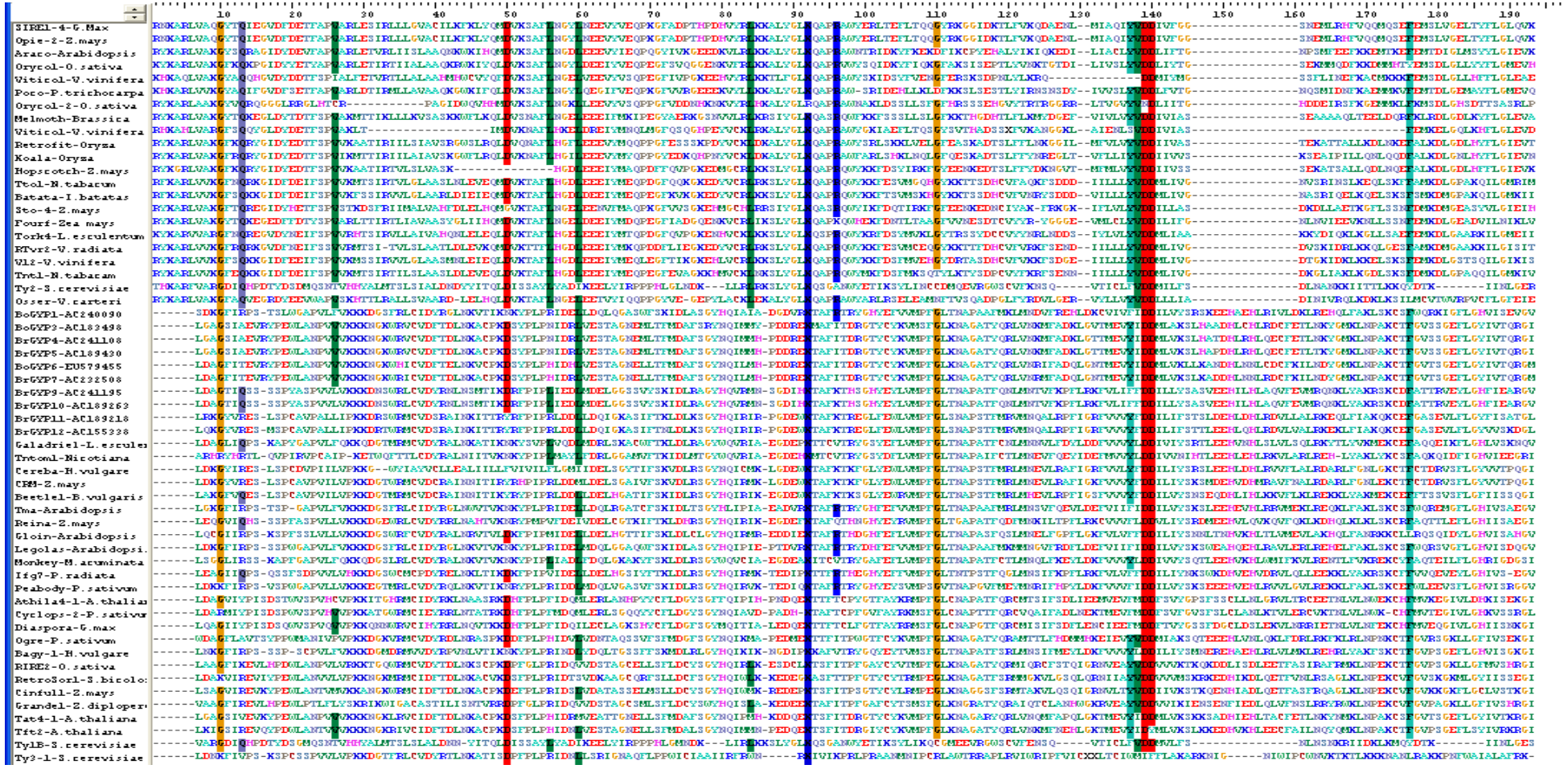
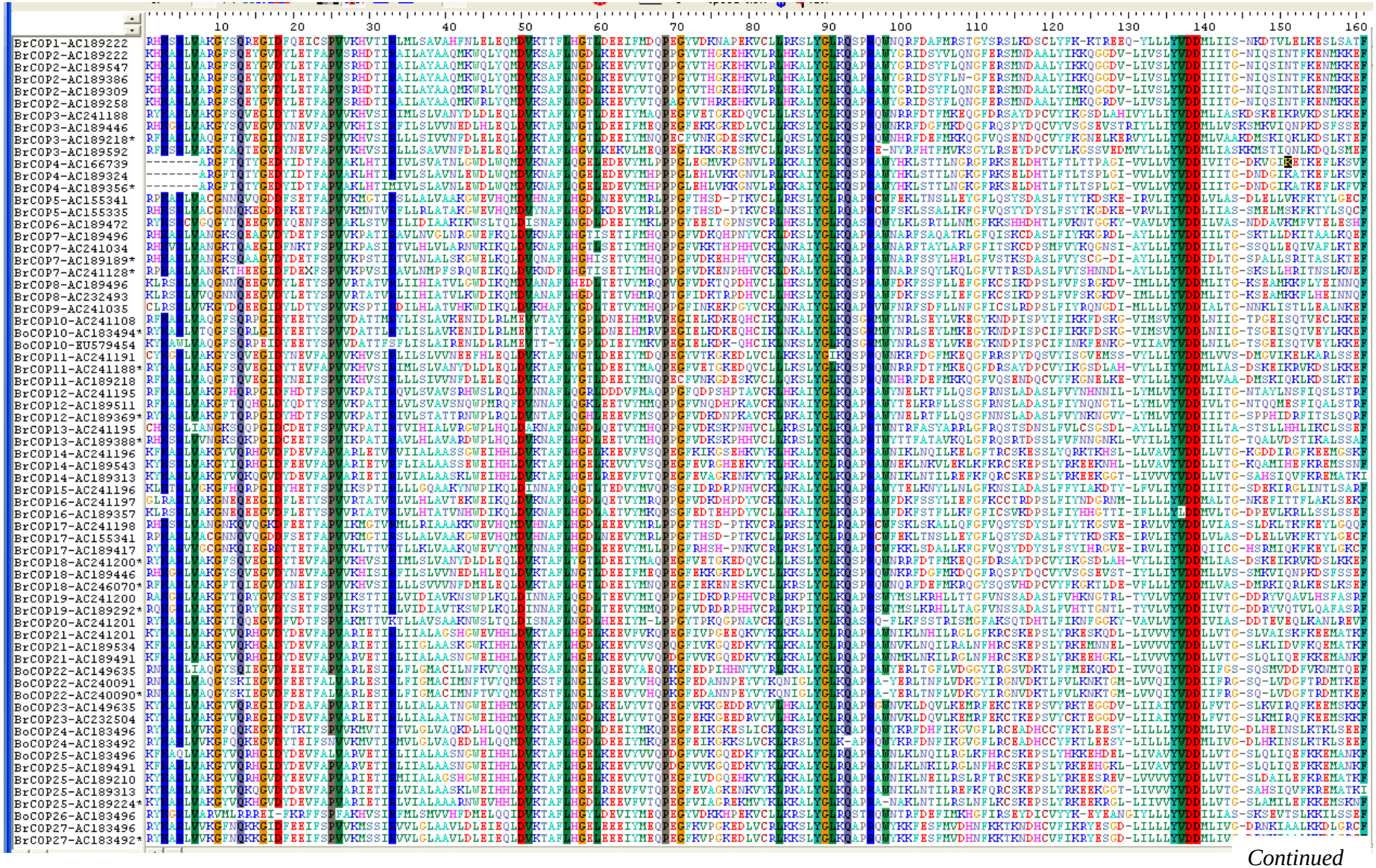


Figure 3.13: Multiple sequence alignment of RT regions of 110 Copia and Gypsy elements; 63 sequences (53 Copia and 10 Gypsy) are from *Brassica* and remaining 47 are from known retrotransposons collected from Gypsy database. The alignment of RT region was done in CLUSTALW and ~190 aa region was extracted from the original alignment and edited manually. Dashes indicate deletions; vertical coloured lines indicate homology in sequences and hence show conserved regions. The names at left identify a individual element followed by accession numbers in *Brassica* LTR retrotransposons while it gives genus and species name for known elements. Br: *Brassica rapa*. Bo: *Brassica oleracea*. Bn: *Brassica napus.* COP: Copia. GYP: Gypsy


10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
BrCOP1-AC189222
BrCOP2-AC189222
BrCOP2-AC189547
BrCOP2-AC189386
BrCOP2-AC189309
BrCOP2-AC189258
BrCOP3-AC241188
BrCOP3-AC189446
BrCOP3-AC189218*
BrCOP3-AC189592
BrCOP4-AC166739
BrCOP4-AC189324
BrCOP4-AC189356*
BrCOP5-AC155341
BrCOP5-AC155335
BrCOP6-AC189472
BrCOP7-AC189496
BrCOP7-AC241034
BrCOP7-AC189189*
BrCOP7-AC241128*
BrCOP8-AC189496
BrCOP8-AC232493
BrCOP9-AC241035
BrCOP10-AC241108
BoCOP10-AC183494*
BoCOP10-EU579454
BrCOP11-AC241191
BrCOP11-AC241188*
BrCOP11-AC189218
BrCOP12-AC241195
BrCOP12-AC189511
BrCOP12-AC189369*
BrCOP13-AC241195
BrCOP13-AC189388*
BrCOP14-AC241196
BrCOP14-AC189543
BrCOP14-AC189313
BrCOP15-AC241196
BrCOP16-AC241197
BrCOP16-AC189357
BrCOP17-AC241198
BrCOP17-AC155341
BrCOP17-AC189417
BrCOP18-AC241200*
BrCOP18-AC189446
BrCOP18-AC246070*
BrCOP19-AC241200
BrCOP19-AC189292*
BrCOP20-AC241201
BrCOP21-AC241201
BrCOP21-AC189534
BrCOP21-AC189491
BoCOP22-AC149635
BoCOP22-AC240091
BoCOP22-AC240090*
BrCOP23-AC149635
BrCOP23-AC232504
BoCOP24-AC183496
BrCOP24-AC183492
BrCOP25-AC183496
BrCOP25-AC189491
BrCOP25-AC189210
BrCOP25-AC189313
BrCOP25-AC189224*
BoCOP26-AC183496
BrCOP27-AC183496
BrCOP27-AC183492*

Continued

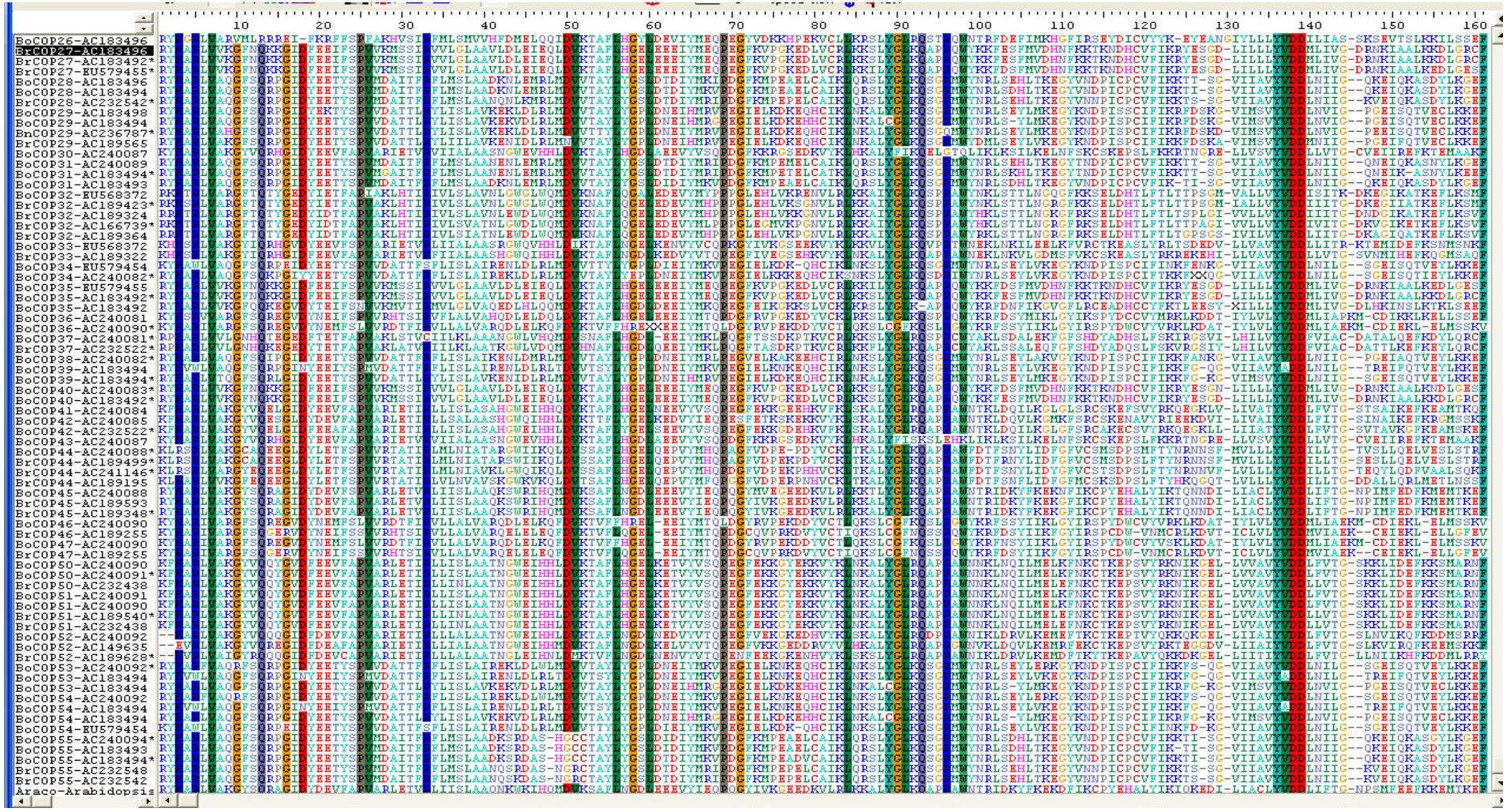


Figure 3.14: Amino acid alignments of conserved region of 138 Copia elements from *Brassica*. The RT regions were retrieved from databases and aligned using CLUSTALW and then ~160 aa region was extracted from the original alignment and edited manually. Dashes indicate deletions; vertical coloured lines indicating homology show conserved regions. The names at left identify individual elements and their database accession numbers.

CHAPTER 4
CHARACTERIZATION OF LINEs AND SINEs: UBIQUITOUS COMPONENTS OF *BRASSICA* CROP GENOMES

Summary

The non-LTR retrotransposons (retroposons) are abundant in plant genomes including *Brassicaceae*. Eight novel families of LINEs (four autonomous and four non-autonomous) and ten of SINEs were identified and characterized from *Brassica* genomes. The autonomous LINEs display two or three open reading frames, ORF1 and ORF2, where the ORF1 domain is a *gag* protein domain, while ORF2 encodes endonuclease (EN) and a reverse transcriptase (RT). Three out of four families encode an additional RNase H (RH) domain, which is common in 'R' and 'I' type of LINEs from *Drosophila*. The PCR analysis of LINEs and SINEs indicate their diversity and widespread occurrence in *Brassica* genomes. Database searches revealed the existence of LINE and SINE families in closely related genera including *Arabidopsis* indicating their origin from common ancestors predating their separation. Comparing the reverse transcriptase of *Brassica* LINEs with those of known LINEs from other plants, *Brassicaceae* LINEs clustered in separate clades. Four clades were observed in *Brassica* LINEs suggesting the separation of four families. Similarly the SINE elements from *Brassica* clustered into 10 families.

4.1 Introduction

The non-long terminal repeat (Non-LTR) retrotransposons (or retroposons) are subdivided into long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) based on their sizes, internal coding/non-coding regions and structural features. The autonomous LINEs are characterized by TSDs of varied lengths, typically possess 1 or 2 (sometimes 3) ORFs and a poly(A) tail at 3′ terminal end. They possess an endonuclease (EN/APE) and a reverse transcriptase (RT) domain, while a few LINEs exhibit additional domains such as Zinc finger (ZF) and RNase H (RH). The RNase H domain is present in 'TAD', 'R1', 'LOA' and 'I' families of LINEs (Malik *et al.,* 1999; Schmidt, 1999; Jurka *et al.,* 2007). Few LINEs are characterized in plants, although the number of reported plant LINEs is growing with genome sequencing. PCR analysis revealed the presence of LINEs in three *Beta* (*vulgaris, lomatogona, nana*) species, *Allium cepa, Oryza sativa, Secale cereale, Nicotiana tabacum* and *Antirrihnum majus* (Kubis *et*

al., 1998). The first well characterized plant LINE *Cin4* was detected in the A1 gene of *Zea mays* as an insertion in the 3´ untranslated region (Noma *et al.*, 1999). An active LINE element called *Karma* was characterized from *Oryza sativa* (Komatsu *et al.*, 2003), while another LINE *Llb* was described from *Ipomoea batatas* genome (Yamashita and Tahara, 2006). Well characterized LINEs from plants include *BLIN* from *Hordeum vulgare, del2* from *Lilium speciosum, LINE-CS* from *Cannabis sativa* and *ATLN* from *Arabidopsis thaliana* (Noma *et al.*, 2001; Vershinin *et al.*, 2002). A LINE family named *BNR* was described from the genome of *Beta vulgaris* having 3 well characterized elements (*BNR1-BNR3*). The elements range in size from 6.4-9.3 kb, flanked by 7-22 bp TSDs and exhibiting two non-overlapping ORFs (Heitkam and Schmidt, 2009).

The second group of retroposons designated SINEs are 100-500 bp large elements flanked by variable TSDs, a poly adenosine tail at 3´ terminus, an internal polymerase III promoter and non-tRNA region of variable sizes (Kapitonov and Jurka, 2003; Deragon and Zhang, 2006; Kramerov and Vassetzky, 2011). The tRNA region of the SINEs displays two well conserved sequence motifs called box A and box B, which served as internal promoter for the transcription of SINEs by RNA polymerase III. The SINEs are non-autonomous elements but are mobile and utilize the enzymatic machinery of LINEs for their transposition. *TS* family of SINEs was detected as highly repetitive family among *Solanaceae* crops like *Capsicum annum, Solanum* (*Lycopersicon*) *esculentum* and *Solanum tuberosum* (Pozueta-Romero *et al.*, 1998). SINE elements named *S1* are well characterized in *Brassica* (Goubely *et al.*, 1999), which are ~170 bp in size and widely distributed among members of *Brassicaceae*. Another *Brassica oleracea* specific SINE family (*BoS*) is distributed in *Brassica* with ~4290 estimated copies belonging to different families (Deragon and Zhang, 2006). The *Au* SINEs are very diverse elements detected in *Gramineae* (*Aegilops umbellulata, Triticum aestivam, Zea mays*), *Solanaceae* (*Nicotina tabacum, Solanum esculentum*), *Fabaceae* (*Medicago truncutula, Lotus japonicas, Glycine max*) and others (Fawcett *et al.*, 2006). A survey of SINEs in the rice genome led to the identification of 13487 copies of SINEs, of which *F524* is the most active SINE in rice with highest (119) intact copies. *SINE3_OS* have above 7000 copies but only 10 intact elements were identified, the remaining are all truncated copies (Khan *et al.*, 2011).

The present study aimed to identify the range of LINE and SINE elements in sequenced *Brassica* BACs and characterize their diversity across *Brassica* germplasms.

4.2 Results

4.2.1 Identification and general features of *Brassica* LINEs

The comparison of similar regions from *Brassica rapa* and *Brassica oleracea* BAC sequences led to the identification of six LINE elements by dot plot analysis. These sequences were further used as query in GenBank database to collect the similar sequences from *Brassica* and related genera *Arabidopsis* and 30 full lengths autonomous LINEs including the query sequences (Table 4.1) were collected. The structural features and phylogenetic analysis of these 30 autonomous LINEs split them into four different families. No strong hits against any known LINE family in TE databases were found, so these novel families were named as *Rehan*, *Faizan*, *Furqan*, and *Nouman*. Out of 30 elements, 10 are members of *Rehan* (*BrLINE1-1* to *BrLINE1-10*), 6 are members of *Faizan* (*BrLINE2-1* to *BrLINE2-6*), 4 are representing *Furqan* (*BoLINE3-1* and *BrLINE3-4*) and 10 representing *Nouman* (*BrLINE1-1* to *BrLINE1-10*) family of LINEs. The sizes, host target site duplications, poly(A) signal and the open reading frames of the elements were studied in detail. *Rehan* is the highest copy number family followed by *Nouman* with members dispersed in several *Brassica* and *Arabidopsis* genomes. In contrast, *Furqan* is considered to be the family with lowest members, where only 2 complete elements were collected from *Brassica* Nucleotide Collection database. The other families are intermediate. The autonomous LINEs in *Brassica* range is sizes from 3361 to 8038 bp. *BrLINE1-1* (8038) is the largest LINE, followed by *BrLINE3-1,* which is a 7313 bp long in size. Almost all the elements identified were flanked by host target site duplications (TSDs) of 5-19 bp, the average sizes being 13-15 bp. In all the cases the poly(A) tail is present with a 7-23 bp stretch except *BrLINE4-5* element from *Nouman* family, where 40 bp polyadenylation signals were detected in the tail region with 17 bp TSDs (Figure 4.1; Table 4.1).

4.2.2.1 Characterization and structural features of *Rehan* family of LINEs

The largest family designated *Rehan* is represented by 15 full length elements, out of which 10 well characterized elements are listed in table 5.1. *BrLINE1-2,* a 7232 bp element was the first element of this family identified from *Brassica rapa* accession ‘AC189222.2’ while identifying LTR-retrotransposons. The element was inserted in two

LTRs without any other identifiable portion of LTR retrotransposons. The detailed investigation led to the detection of a LINE element integrated in two solo LTRs. The element has two open reading frames, where ORF1 is 258 aa while ORF2 is 940 aa large. ORF2 encode a *pol* gene with 5′-EN-RT-RH-3′ domain organization. Computer based homology searches by using 7232 bp *BrLINE1-2* sequence against *Brassica* Nucleotide Collection database gave 10 full length homologues covering >70% of the query, while several other partial homologues were identified. The largest element from the family is *BrLINE1-1*, which is 8038 bp large in size including 8 bp TSDs at both ends and displaying a 7 bp poly(A) tail at its C-terminal end. Two small insertions ~400-500 bp integrated in its central region increase its size and make it defective. The element encodes a typical plant LINE structure 5′-EN-RT-RH-3′ (Figure 4.1).

BrLINE1-3 is a 6816 bp in size, flanked by TSDs of 6 bp and 16 bp poly(A) stretch at 3′ end. It is composed of 132 aa ORF1 encoding the zinc finger (ZF) domain and 1215 aa ORF2 encoding the *pol* protein domains in 5′-EN-RT-RH-3′ order. *BrLINE1-5* is 5777 bp in size including 5 bp TSDs at both ends and 21 bp poly(A) tail. The ORF1 (348 aa) is in the opposite orientation i.e. downstream to ORF2 (1369 aa). *BrLINE1-6* and *BrLINE1-7* are 5398 and 5352 bp in sizes including TSDs of 15 and 8 bp respectively. The *BrLINE1-6* display a 13 bp poly(A) tail while *BrLINE1-7* exhibit the largest (22 bp) poly(A) tail investigated in members of *Rehan* family. The element is highly A/T rich (60.3%) with many small poly(T) stretches in its internal region. *BrLINE1-8* (5033 bp) was identified from *Brassica rapa* BAC 'AC189587.2' and has rearranged sequence. It encodes two ORFs in the same frameshift but contains 2 stop codons. *BrLINE1-9* is the smallest member of *Rehan* family with a size of 3867 bp including 9 bp flanking TSDs and 19 bp poly(A) C-terminal end (Table 4.1). Like *BrLINE1-2*, another element named *BoLINE1-10* was identified as integrated in two LTRs at its both terminal ends on *Brassica rapa* BAC 'AC183494.1'. It is integrated exactly downstream to 5′-CA-3′ termini of 358 bp 5′ LTR and ends 232 bp upstream to the start of 358 bp 3′ LTR (Figure 4.1). *BoLINE1-10* is a LINE of 5845 bp flanked by a 15 bp target site duplication and 11 bp poly(A) end.

4.2.2.2 Structural features of *Faizan* family of LINEs

Although the majority of the elements have complete protein domains (5′-EN-RT-RH-3′), but *BrLINE2-6* (3361 bp) has a deleted RNase H (RH) region. The largest member of the

family is *BrLINE2-1* identified from *Brassica rapa* BAC 'AC189630.2'. The element is 6382 bp long including flanking TSDs of 15 bp and a poly(A) tail of 10 bp. It has two ORFs; ORF1 is 559 aa while ORF2 is 1338 aa long encoding 5′-EN-RT-RH-3′ protein domains. *BrLINE2-2* is 6299 bp large LINE in *Brassica rapa* (AC189430.2). It is flanked by 13 bp TSDs and a 15 bp 3′ poly(A) tail (Figure 4.1). A 5263 bp homologue was identified from *Brassica rapa* 'AC189651.2' named *BrLINE2-3*. It has two consecutive ORFs: ORF1 is 256 aa and ORF2 is 1320 aa in size and its 3′ untranslated region (UTR) has polyadenylation signal of 23 bp, the highest in this family. The element *BoLINE2-4* is 5077 bp in size including 11 bp TSDs and 9(A) tail at C-terminus. The element has a long ORF of 1243, which encode the protein domains of a typical LINE element (5′-EN-RT-RH-3′). *BrLINE2-5* and *BrLINE2-6* are 3867 and 3361 bp in sizes including TSDs of 18 and 8 bp respectively. Their 3′ UTR have polyadenylation signals of 8 and 15 bp (Table 4.1). *BrLINE2-5* has shown the typical LINE *pol* gene encoding region (5′-EN-RT-RH-3′), while a RH is deleted from the *BrLINE2-6* (Figure 4.1).

4.2.2.3 Identification and characterization of *Furqan* family of LINEs

The dot plot analysis of *Brassica* BACs to identify LTR retrotransposons led to the identification of a LINE insertion in *Brassica oleracea* BAC 'AC240078.1'. The element named *BoLINE3-1* is 7313 bp in size, flanked by TSDs of 13 bp. The structural organization of the *BoLINE3-1* revealed that it encodes two non-overlapping ORFs, ORF1 is 526 aa and ORF2 is 873 aa in size, although few other small ORFs can be observed in the element in other frameshifts. The domain organization showed the presence of EN and RT, while RH is absent in this family. *BrLINE3-2* is a 5414 bp element, flanked by TSDs of 7 bp and a poly(A) tail of 28 bp. Two ORFs are detected with a size of 247 and 942 aa in the same frameshift (Figure 4.1). The element is a defective as endonuclease is deleted during the rearrangement of the sequence during the evolutionary phase. *BrLINE3-3* and *BrLINE3-4* are 5925 and 4013 bp in sizes, flanked by 13 bp TSDs and having a poly adenosine tail of 17 and 12 bp respectively at their C-terminal end (Table 4.1).

4.2.2.4 Structural features of *Nouman* family of LINEs

Nouman is a high copy number family of LINEs after *Rehan*. The first element (*BoLINE4-2*) from the family was identified from *Brassica oleracea* BAC 'AC240089.1' by dot plot

comparison of *Brassica rapa* (AC155341.2) and *Brassica oleracea* (AC240089.1) accessions. By using this as a query sequence, several homologues were collected from *Brassica* genomes and investigated for their hallmarks (TSDs, poly(A), EN and RT). Ten full length elements with >75% query coverage and identity were enlisted and described in detail, although many other homologues with less identity were also present. The largest element *BrLINE4-1* is 6725 bp in size, generates 19 bp TSDs at both ends and 19 bp polyadenylation signals at 3′ UTR. The element encodes two ORFs in two frameshifts, ORF1 is 654 aa and ORF2 is 1317 bp in size encoding the 5′-EN-RT-RH-3′ protein domains. *BoLINE4-2* is 6560 bp large having TSDs of 17 bp and 11 bp poly adenosine tail at C-terminal end (Figure 4.3). *BrLINE4-3* was described from *Brassica rapa* BAC 'AP011511.1' having a size of 6553 bp including 19 bp TSDs and 11 bp polyadenylation signal of 11 bp. *BoLINE4-4*, a 6482 bp LINE generates 8 bp duplications upon integration to the host sites and terminated at 3′ terminus by polyadenylation signals of 14 bp. A similar sized (6424 bp) element designated *BrLINE4-5* was identified from a *Brassica rapa* (AC232437.1), exhibit 17 bp TSDs and the longest poly(A) signals (40 bp) at their 3′ untranslated region. The element encodes two ORFs, ORF1 with a size of 595 and ORF2 with a size of 1363 aa encoding typical LINE *pol* gene (5′-EN-RT-RH-3′). *BrLINE4-6* is a 5124 bp LINE, flanked by 6 bp TSDs and 16 bp poly(A) tail at 3′ terminal end. *BrLINE4-7* and *BrLINE4-8* are 4740 and 4321 bp LINEs, flanked by 5 bp TSDs, 17 and 11 bp poly adenosine tails at 3′ terminus. The smallest LINEs from *Nouman* family are *BrLINE4-9* and *BoLINE4-10*, which are 3846 and 3416 bp respectively. Both elements have a polyadenosine tail of 11 bp but the domain organization of *BoLINE4-10* (Figure 4.1) revealed that the EN is deleted from the element during the rearrangement of the element in evolutionary stages.

4.2.3 Open reading frames and domain organization of *Brassica* LINEs

Two open reading frames (ORF1 and ORF2) were identified in most *Brassica* LINEs investigated. The ORF1 is a *gag* protein domain, while ORF2 encodes the *pol* gene encoding the EN, RT and RH protein domains. In few cases, only single ORF encoding these domains were observed, while in few elements two ORFs are arranged in two different frameshifts. Few LINEs have shown the presence of stop codons in *pol* polyprotein indicative of a defective element. Five different protein domain organizations were observed as 5′-EN-RT-RH-3′, 5′-(X)-EN-RT-RH-3′, 5′-EN-RT-3′, 5′-RT-RH-3′

and 5′-RH-3′; where X is any additional ORF like ZF, unknown protein (DUF) and Pre-mRNA-splicing factor (PRP). About 80% elements showed a typical 'R' or 'I' (5′-EN-RT-RH-3′) type LINE domain organization, while the others have either an extra domain in them or deleted EN or RH domain. A 7.3 kb *BoLINE3-1* has shown a typical plant L1 LINE structure, where EN is followed by a reverse transcriptase. In its closely similar relative *BrLINE3-3* and *BrLINE3-4*, EN and RT domains are observed lacking a RH domain. All the LINE families have a RH domain in them with the exception of members from *Furqan* family, where only EN and RT domains are present (Table 4.1: Figure 4.1).

4.2.4 Identification and characterization of non-autonomous LINEs in *Brassica*

The dot plot comparison of *Brassica rapa* BAC 'AC189298.1' with its homoeologous *Brassica oleracea* BAC 'EU642504.1' exposed some insertion sites, which after detail investigations were found to be the non-autonomous LINEs. Three elements were inserted in *Brassica oleracea* (EU642504.1), while one LINE was found inserted in *Brassica rapa* accession (AC189298.1). The non-autonomous LINE elements were named *Bo-N-LINEX*, where *Bo* stands for *Brassica oleracea*, *N* indicate non-autonomous and *X* after LINE represent the number. *Bo-N-LINE1* was identified from *Brassica oleracea* (EU642504.1) at position 108339-108982 bp. The element is 690 bp in size, generating TSDs of 7 bp and polyadenosine signals of 19 bp. *Bo-N-LINE2* is a 1016 bp large element with 9 bp TSD and a tail of 10 nucleotides, out of which 7 are Adenine (A) and three are Guanine (G) nucleotides. The element is a low copy number with no significant hits to any known TEs from Repbase and Plant Repeat databases of TEs. Another non-autonomous LINE-like element designated as *Bo-N-LINE3* was identified from *Brassica oleracea* (EU642504.1). The element was 1080 bp in size generating 13 bp TSDs and having 11 bp polyadenosine tail at its 3 terminal end. The BLAST results against Repbase TE database retrieved significant hits (~60%) to *Arabidopsis 'ATLINEs'*. This suggests that this non-autonomous LINE is the defective element of *ATLINE*-like elements. Another LINE-like element (*Br-N-LINE4*) was detected residing in *Brassica rapa* (AC189298.1) accession, while absent in its homoeologous *Brassica oleracea* BAC (EU642504.1). The element is 914 bp in size including the 13 bp TSDs at both ends and terminated by a 10 polyadenosine nucleotides at 3 terminal end (Figure 4.2).

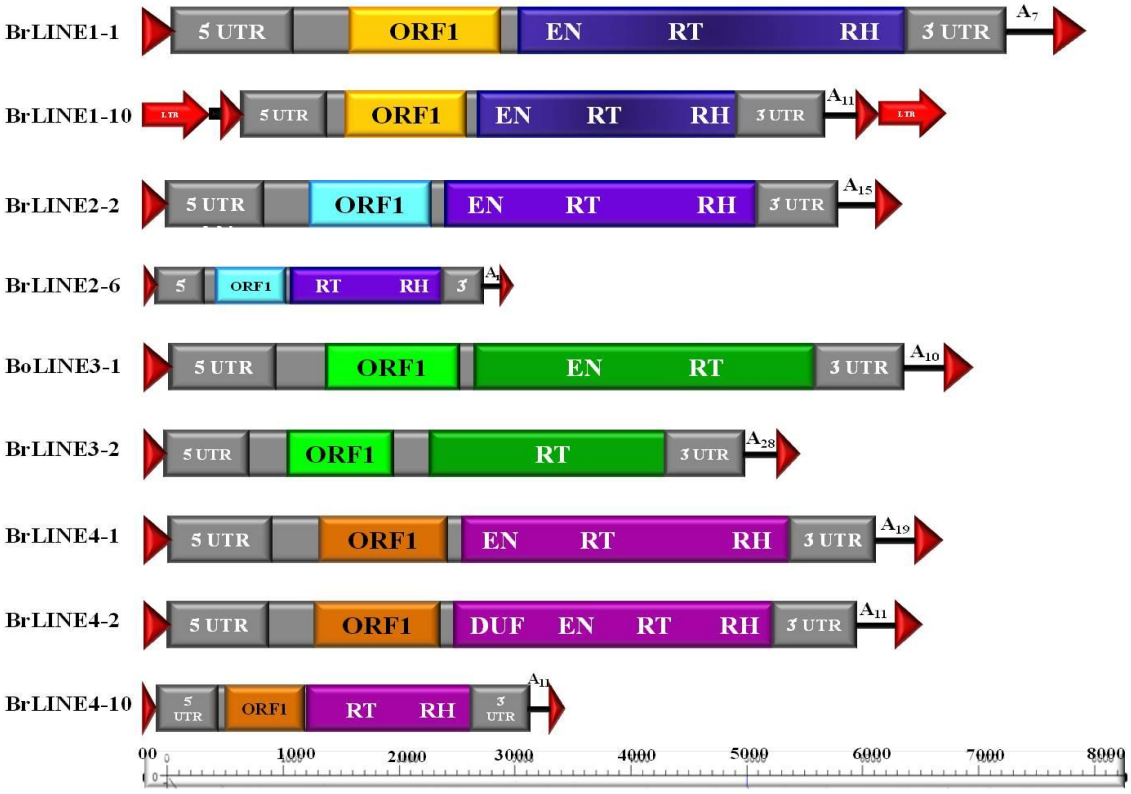


Figure 4.1: Structure and organization of LINE retroposons in *Brassica*. Red arrows indicate target site duplications, while a black line near the 3′ terminus shows the poly(A) tail with its length. Two open reading frames are indicated in different colours: ORF1 encoding a nucleocapsid protein (*gag*), and ORF2 encoding an endonuclease (EN) and the reverse transcriptase (RT). An additional RNase H (RH) domain encoded by ORF2 is present in most LINEs near the 3′ end. The location and protein domain organizations of LINEs is indicated. ORFs are flanked by untranslated regions (UTR), present at both termini. The scale below (bp) shows lengths.

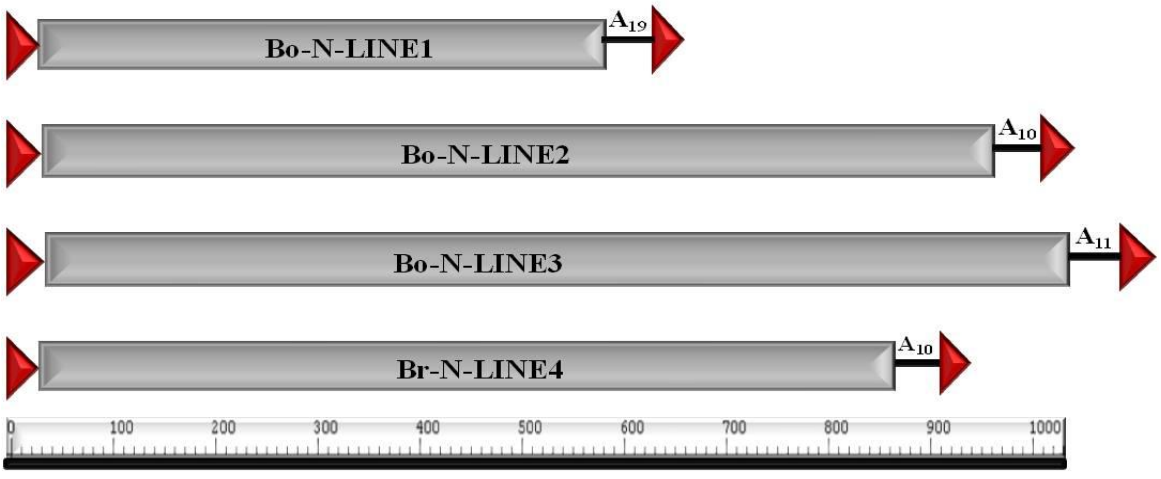


Figure 4.2: Schematic representation of non-autonomous LINEs in *Brassica*. Red arrowheads indicate target site duplications, while a black line near the 3′ terminus shows the poly(A) tail and its length. The names in the centre of the element are representing the LINE family. No protein domains are present indicating these are non-autonomous LINEs. The scale below shows lengths (bp).

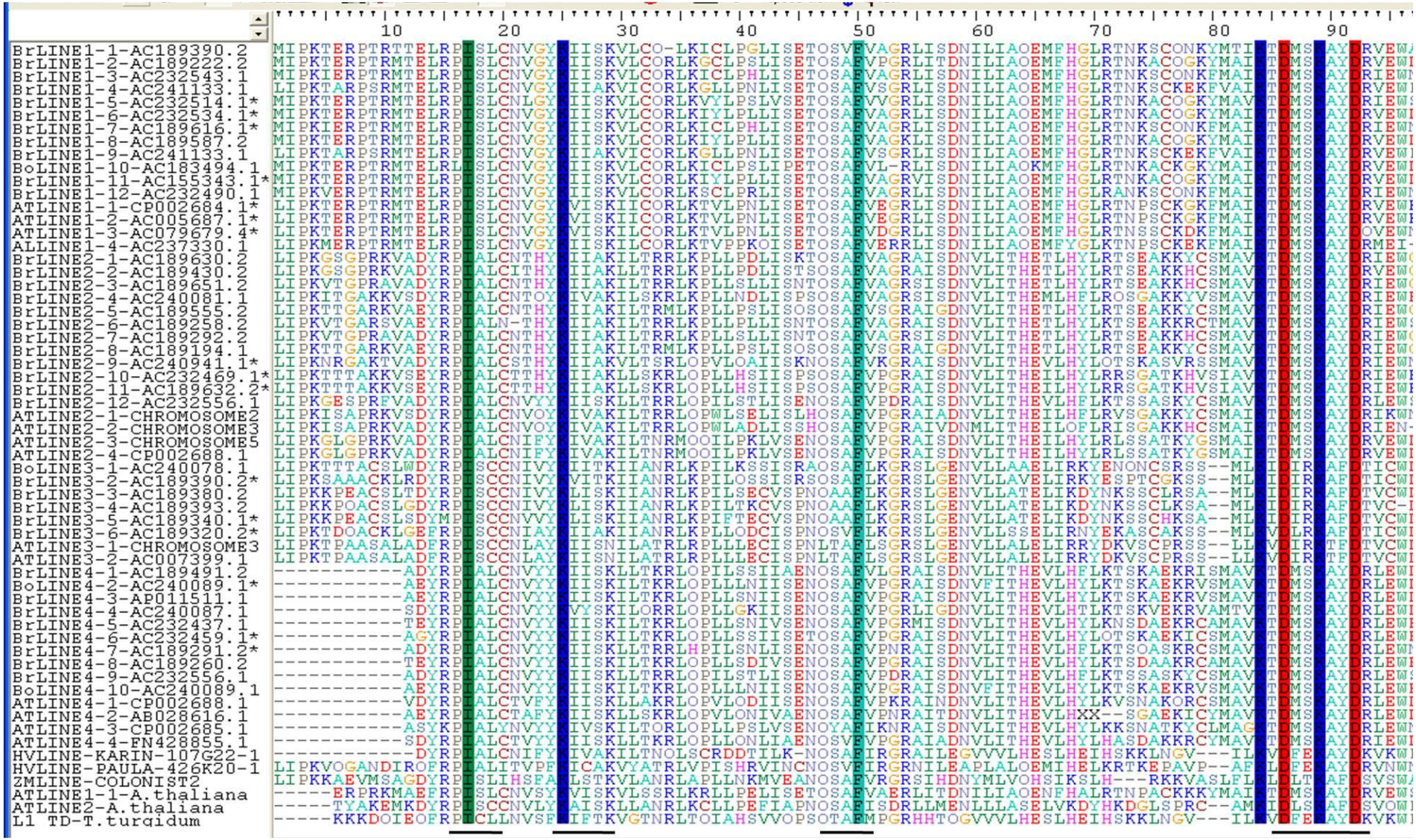
Table 4.1: List of *Brassica* autonomous and non-autonomous LINEs with BAC accessions, sizes, TSDs, poly(A) tail, ORFs and protein domains. Nucleotide sequences of representative elements are available in Appendices (attached CD).

No.	Element name	Family	BAC Accession	Species	Size	TSD	Poly(A)	ORFs	Orienta-tion	Domain Structure (5′-3′)
1	*BrLINE1-1*	*Rehan*	AC189390.2	*B. rapa*	8038	TCAGTCTC	07	6	3′-5′	EN-RT-RH
2	*BrLINE1-2*	*Rehan*	AC189222.2	*B. rapa*	7232	TACCATTATAATTA	12	2	5′-3′	EN-RT-RH
3	*BrLINE1-3*	*Rehan*	AC232543.1	*B. rapa*	6816	ACTAAC	16	2	3′-5′	ZF-EN-RT-RH
4	*BrLINE1-4*	*Rehan*	AC241133.1	*B. rapa*	6387	TTGTAAAGC	19	3	3′-5′	EN-RT-RH
5	*BrLINE1-5*	*Rehan*	AC232514.1	*B. rapa*	5777	GGAAG	21	2	5′-3′	EN-RT-RH
6	*BrLINE1-6*	*Rehan*	AC232534.1	*B. rapa*	5398	TACTAAACAACAACT	13	3	5′-3′	EN-RT-RH
7	*BrLINE1-7*	*Rehan*	AC189616.1	*B. rapa*	5352	GTAATAAT	22	2	5′-3′	EN-RT-RH
8	*BrLINE1-8*	*Rehan*	AC189587.2	*B. rapa*	5033	TTATATGGTTAATGA	21	2	3′-5′	EN-RT-RH
9	*BrLINE1-9*	*Rehan*	AC241133.1	*B. rapa*	3822	TCCATAGAA	19	3	5′-3′	EN-RT-RH
10	*BoLINE1-10*	*Rehan*	AC183494.1	*B. oleracea*	5845	TATCTCATACGATCT	11	6	3′-5′	EN-RT-RH
11	*BrLINE2-1*	*Faizan*	AC189630.2	*B. rapa*	6382	TTGTATTCATGTAAC	10	2	3′-5′	EN-RT-RH
12	*BrLINE2-2*	*Faizan*	AC189430.2	*B. rapa*	6299	TACTTGTTAAAAG	15	2	5′-3′	EN-RT-RH
13	*BrLINE2-3*	*Faizan*	AC189651.2	*B. rapa*	5263	GAGTTTGA	23	2	5′-3′	EN-RT-RH
14	*BoLINE2-4*	*Faizan*	AC240081.1	*B. oleracea*	5077	GAAGCTATTTC	9	1	5′-3′	EN-RT-RH
15	*BrLINE2-5*	*Faizan*	AC189555.2	*B. rapa*	3867	GAAATAATATTCCACATC	08	1	5′-3′	EN-RT-RH
16	*BrLINE2-6*	*Faizan*	AC189258.2	*B. rapa*	3361	GTGTTGTG	15	2	5′-3′	RT-RH
17	*BoLINE3-1*	*Furqan*	AC240078.1	*B. oleracea*	7313	CTATGATTGGGGA	10	2	5′-3′	EN-RT
18	*BrLINE3-2*	*Furqan*	AC189390.2	*B. rapa*	5414	GTTAGTC	28	2	3′-5′	RT
19	*BrLINE3-3*	*Furqan*	AC189380.2	*B. rapa*	5925	ACTCGGATAATAT	17	2	5′-3′	EN-RT
20	*BrLINE3-4*	*Furqan*	AC189393.2	*B. rapa*	4013	GTATCTTGTTTAG	12	2	5′-3′	EN-RT
21	*BrLINE4-1*	*Nouman*	AC189491.2	*B. rapa*	6725	GAACAACCTTGTTTGTGAC	19	2	5′-3′	EN-RT-RH
22	*BoLINE4-2*	*Nouman*	AC240089.1	*B. oleracea*	6560	AAATTGTTTCGACCTTG	11	2	5′-3′	DUF-EN-RT-RH
23	*BrLINE4-3*	*Nouman*	AP011511.1	*B. rapa*	6553	TGGAGTCCTACTTAAAAGT	11	2	5′-3′	PRP-EN-RT-RH
24	*BoLINE4-4*	*Nouman*	AC240087.1	*B. oleracea*	6482	CCATTGTC	14	4	3′-5′	EN-RT-RH
25	*BrLINE4-5*	*Nouman*	AC232437.1	*B. rapa*	6424	GAAAAACAATGGCAGGT	40	2	5′-3′	EN-RT-RH
26	*BrLINE4-6*	*Nouman*	AC232459.1	*B. rapa*	5124	GTCATC	16	2	3′-5′	EN-RT-RH
27	*BrLINE4-7*	*Nouman*	AC189291.2	*B. rapa*	4740	ATGAG	17	1	3′-5′	EN-RT-RH
28	*BrLINE4-8*	*Nouman*	AC189260.2	*B. rapa*	4321	ACGTC	11	2	3′-5′	EN-RT-RH
29	*BrLINE4-9*	*Nouman*	AC232556.1	*B. rapa*	3846	ATATTTT	11	2	5′-3′	EN-RT-RH
30	*BoLINE4-10*	*Nouman*	AC240089.1	*B. oleracea*	3416	TATCA	11	4	5′-3′	RT-RH
31	*Bo-N-LINE1*	*BoNAL1*	EU642504.1	*B. oleracea*	690	GCTTATA	19	ND	5′-3′	Non-autonomous
32	*Bo-N-LINE2*	*BoNAL2*	EU642504.1	*B. oleracea*	1016	GTGTATGAT	10	ND	3′-5′	Non-autonomous
33	*Bo-N-LINE3*	*BoNAL3*	EU642504.1	*B. oleracea*	1080	CGATCAGCTGTTT	11	ND	3′-5′	Non-autonomous
34	*Br-N-LINE4*	*BrNAL4*	AC189298.1	*B.* rapa	914	CAAAAAGTAGTTT	10	ND	3′-5′	Non-autonomous

4.2.5 Sequence analysis and phylogenetic relationship of RT from *Brassica* and other plant LINEs

A total of 60 LINE RT domains were analysed, out of which 40 were collected from *Brassica,* 14 from *Arabidopsis* and 6 from Repbase database of TEs. The most conserved region of the RT (~200-210 aa) around the DD conserved motif is considered for comparative analysis. The similarity among various sequences is much higher (>75-95%) within the members of the same family as compared to other families. The sequence analysis of RT region from *Brassica* and other plant LINEs showed high homology and some conserved regions. The most conserved motif is D-DD, where D is 45 aa upstream to DD (D45DD). This motif is present in all RT sequences from *Brassica, Arabidopsis*, *Hordeum vulgare* (*KARIN*, *PAULA*), *Triticum turgidum* (*L1_TD*) and *Zea mays* (*COLONIST2*). A seven amino acid signature (HLLFADD) including the DD motif is conserved in all RT sequences. Another conserved motif (GLRQGD) 45 aa upstream to DD is present in all *Brassica*, *Arabidopsis*, *Triticum*, *Hordeum* and *Zea mays* RT sequences aligned. A conserved signature (KTDMSKAY/FD) is also observed in all LINE RT sequences, whereas several other regions with 1-6 aa conserved regions are shared by various RT (Figure 4.3).

Based on the alignment, the phylogeny of *Brassica* LINEs among themselves and other plant LINEs were investigated by constructing Neigbour-Joining tree with 1000 bootstrap repetitions in Geneious program. The LINE elements clustered into 5 family specific clusters, where all LINEs from Triticale clustered in a clade, while all other families make family-specific clade. The RT regions from *Brassica* and *Arabidopsis* LINEs clusterd together in same families, which indicate their origin from a common ancestor. The largest family Rehan contains 17 elements, of which 12 were *Brassica* and 5 were from *Arabidopsis*. In contrast, the smallest family Furqan contains 9 elements, 6 from *Brassica* and remaining 3 from *Arabidopsis*. Similarly the Faizan and Nouman families also include the *Brassica* and *Arabidopsis* LINE elements together. The sharing of same family from *Brassica* and *Arabidopsis* LINEs suggest that the LINE elements are old elements and were present in the ancestral genome before the separatiuon of two genera (Figure 4.4).



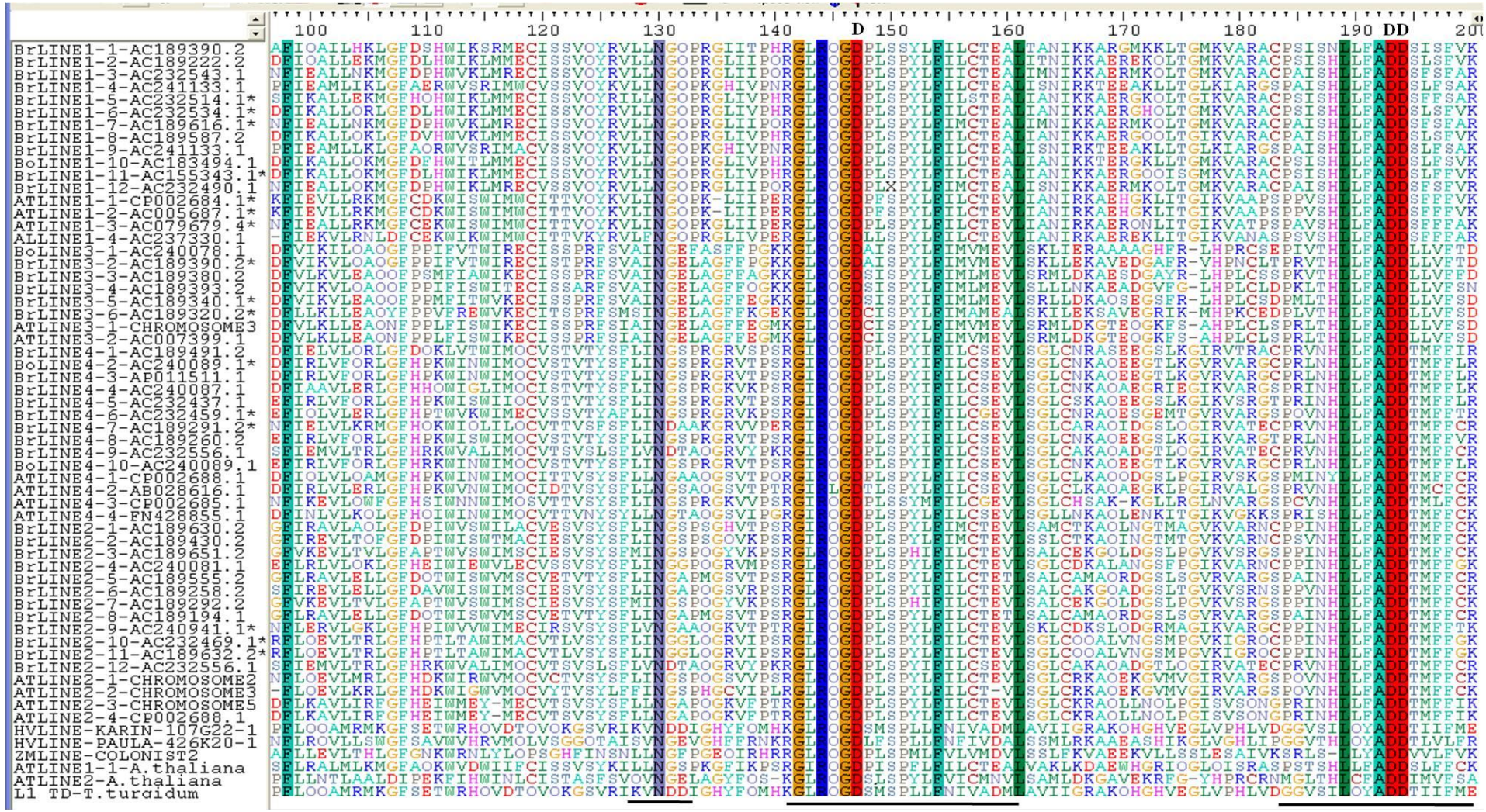


Figure 4.3: Alignment of reverse transcriptase (RT) regions of *Brassica* and other LINEs including *Arabidopsis.* Approximately 200 aa from RT was aligned in CLUSTALW. The alignment shows several conserved motifs underlined by grey lines and highlighted by vertical coloured bars of conserved amino acids. The DD motif is most conserved and present in almost all the LINEs. *Brassica* and *Arabidopsis* LINEs show high homology in RT regions.

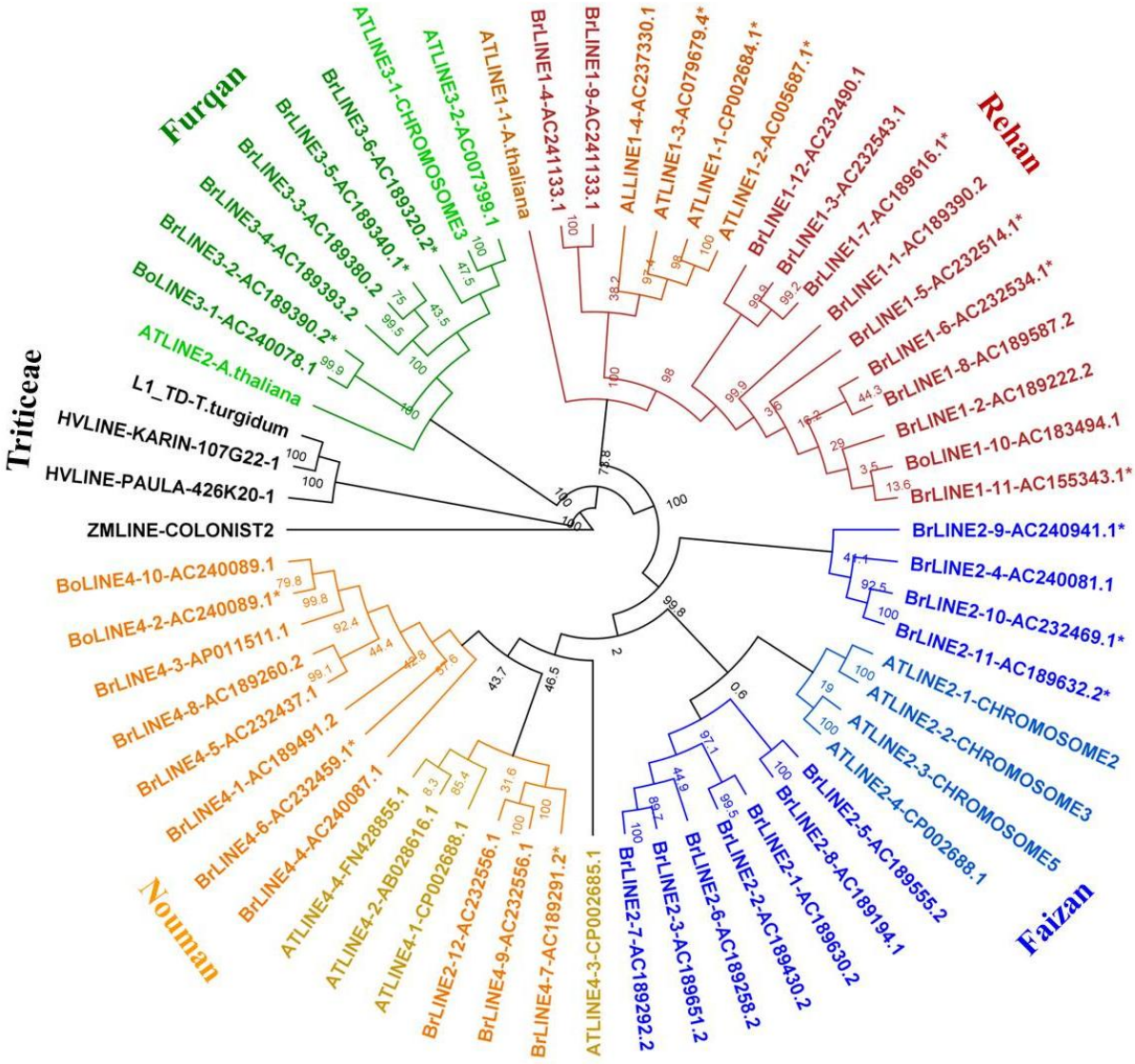


Figure 4.4: Phylogenetic tree of reverse transcriptase of *Brassica*, *Arabidopsis* and known plant LINEs. The tree was constructed on ~ 200 amino acid conserved region around DD motif (192 upstream and 6 aa downstream) by the Neighbour-Joining method with 1000 bootstrap replicates using the Geneious Pro program. *Zea mays* LINE *COLONIST2* was used to root the tree. The bootstrap support is shown near the nodes. The known LINEs represented by black colours were obtained from Repbase database of eukaryotic transposable elements. The sequences clustered into five clades representing four LINE families from *Brassicaceae* and one clade representing LINEs from other plants from *Triticeae*.

4.2.6 PCR amplification of RT from *Brassica Rehan* LINEs family

Degenerate oligonucleotide primers pair BoLINE1F 5′-GTTGACCTGAAACCATCTCA-3′ and BoLINE1R 3′-CAACTAGGATGACGGAACTG-5′ were designed by inspection of conserved amino acid sequences of the reverse transcriptase. PCR amplification of a 645 bp RT region was performed from 40 *Brassica* cultivars. The results showed

amplification of the RT region from 30 diploid and polyploid *Brassica* crops. *Brassica rapa* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona) amplified the product, while no amplification was observed from 'Vertus' and 'Suttons' accessions. No amplification from *Brassica nigra* accessions suggested a separate evolutionary line in B vs A-C-genomes. Among C-genomes, *Brassica oleracea* cultivars (De Rosny, Kai Lan, Early Snowball, Cuor Di Bue Grosso, GK97361) amplified the RT regions, while no amplification was seen in *Brassica oleracea italica* 'Precoce Di Calabria'. All *Brassica juncea* '(NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna) amplified the products except *Brassica juncea* (*TSAI SIM*). All the six *Brassica napus* cultivars (New, Mar, Last and Best, Fortune, Drakker, Tapidor) and five *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) yielded the expected product indicating the distribution of *Rehan* family members (Figure 4.5a). The four resynthesized hexaploid *Brassica* cultivars also generated bands suggesting the diversity, abundance and distributions of LINEs among *Brassica* genomes.

Table 4.2: List of primers with their names, sequences and sizes of the expected products to amplify the LINEs and SINEs elements in *Brassica*.

Sr.No.	Family	TE Size	Product Size	Primer Name	Primer Sequence
1	*Rehan*	7132	645	BoLINE1F BoLINE1R	GTTGACCTGAAACCATCTCA CAACTAGGATGACGGAACTG
2	*Faizan*	6382	694	BoLINE2F BoLINE2R	GTTCGATTGATTCCCAAAGG CGACTTCAGCAGGTTGATCC
3	*Furqan*	7313	724	BoLINE3F BoLINE3R	TGTAGCCTTTGGGACTACCG CACGCTTGAAACCTGAGATG
4	*Nouman*	6560	726	BoLINE4F BoLINE4R	CCATCGCTCTCTGCAATGTC CGGTACCTCCCTCTTTCTGG
5	*BoNAL1*	690	906	BoNLINE1F BoNLINE1R	CAAAATTAACCCAAATGAGG TGGCATCAAACTTGAACGAA
6	*BoNAL2*	914	1144	BoNLINE2F BoNLINE2R	GGATTTAAGGAAATAGTGGT TGTATACGGATAGATGAAAC
7	*BoNAL3*	1016	1265	BoNLINE3F BoNLINE3R	GAGGTTGCTTCGTATCTTAC CGTCTTATGATCATTGTCCG
8	*BrNAL4*	1080	1286	BrNLINE4F BrNLINE4R	CTGTATTGAGAAATCCTCTA ACGAGTTGTTCTACCATTTG
9	*BoSINE2*	219	365	BoSINE2F BoSINE2R	GAACAAGAAAAATGCAGGG CGTACCATCACATCTCTTTC
10	*BoSINE3*	272	585	BoSINE3F BoSINE3R	TTCGTTCAAGTTTGATGCCA AAAGATCCTCACTGGAATCA
11	*BoSINE9*	524	735	BoSINE9F BoSINE9R	AGCTATTACCATGTCGTTCC ACATAACATTGATACTCCGC
12	*BrSINE10*	376	615	BrSINE10F BrSINE10R	CAAACACTACAAGTGAATAC GCAAGGTGGAGAAGATAAG

4.2.7 *Faizan* LINE family is proliferating in A and C-genome *Brassica*

The diversity and distribution of *Faizan* LINE family among various *Brassica* crops were tested by using primer pair BoLINE2F 5′-GTTCGATTGATTCCCAAAGG-3′ and BoLINE2R 3′-CGACTTCAGCAGGTTGATCC-5′ amplifying a 694 bp RT region. The product was amplified from all *Brassica rapa* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons) and *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, Gk97361) cultivars suggesting its proliferation in A and C-genomes. No amplification from *Brassica nigra* suggests its absence from B-genome. From nine *Brassica juncea* cultivars tested, all except two (NATCO, Varuna) amplified the RT regions of Faizan family suggesting its distribution among different cultivars. Similarly, all *Brassica napus* and *Brassica carinata* and 4 hexaploids *Brassica* (AABBCC) amplified the 694 bp RT product indicating the abundance and distribution of the LINE in almost all A and C-genome diploids and polyploids (Figure 4.5b).

4.2.8 *Furqan* LINE family is proliferating in C-genomes

The diversity and distribution pattern of *Furqan* family of LINEs among *Brassica* genomes is performed by PCR analysis. The degenerative primer pair BoLINE3F 5′-TGTAGCCTTTGGGACTACCG-3′ and BoLINE3R 3′-CGACTTCAGCAGGTTGATCC-5′ were design from conserved region of LINE RT to amplify a 724 bp product. Out of 40 cultivars tested, 19 C-genome specific diploids and polyploids produced the bands. The amplification pattern yielded no expected products from any of A or B-genome *Brassica*. Strong bands of ~724 bp products were amplified from all six cultivars of *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361). Again no amplification from *Brassica juncea* (AABB) strengthens the hypothesis of absence of this LINE from A and B-genomes. One Pakistani origin *Brassica juncea* (NARC-II) amplified the band, whose authenticity was unclear and we assume that it is a mixed hybrid having the introgression of C-genome chromosomes by cross hybridization of species common in Pakistan and Indian regions. The allotetraploids (AACC, BBCC) and hexaploids (AABBCC) yielded the expected products from all the genomes (Figure 4.5c). This suggests the proliferation of *Furqan* LINE family in C-genome species and its hybrids.

4.2.9 Diversity and distribution of *Nouman* LINE family among *Brassica* cultivars

The diversity and distribution of *Nouman* family was confirmed by PCR analysis of 726 bp RT region from 40 *Brassica* cultivars. The primer pair BoLINE4F 5′-CCATCGCTCTCTGCAATGTC-3′ and BoLINE4R 3′-CGGTACCTCCCTCTTTCTGG-5′ was designed from conserved regions of the transposase. Products were amplified from 34 (1 weak), out of 40 cultivars tested. The PCR analysis revealed that *Brassica rapa* cultivars (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons) amplified the bands, while very weak band was amplified in cultivar 'Pak Choy'. The amplification from all six *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361) accessions indicate its distribution among various *oleracea* crops. Out of nine *Brassica juncea* tested, 6 amplified the RT, while 3 (NARC-I, Giant Red Mustard, Varuna) showed no amplification. The allotetraploid crops (AABB, BBCC) and hexaploid (AABBCC) amplified the RT bands revealing the high diversity and distribution of LINE in nearly all cultivars tested (Figure 4.5d).

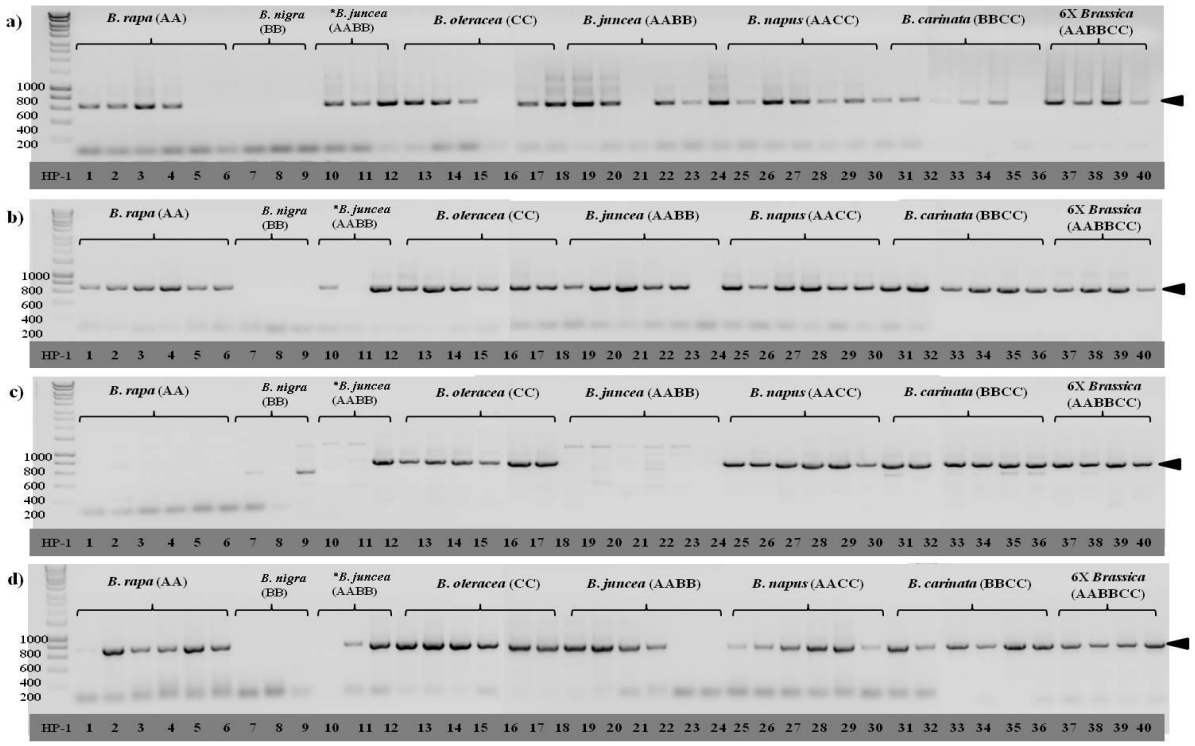


Figure 4.5: PCR amplification of *Brassica* LINEs. a) *BoLINE1*; b) *BoLINE2*; c) *BoLINE3*; d) *BoLINE4.* The upper bands indicated by arrowheads represent the amplification of LINEs RT region from various *Brassica.* (This and subsequent PCR figures show reversed images of ethidium-bromide stained agarose gels following size separation by electrophoresis of PCR products; ladder band sizes shown in bp; numbers below identify DNA accessions listed in Table 2.1.)

4.2.10 Transoson insertional polymorphisms (TIPs) of non-autonomous LINEs in *Brassica* genomes

To observe whether these insertions are sequence specific or across various *Brassica* genomes, four pairs of TIP markers were designed to target insertion including flanking ends (Table 4.2) from the homologous flanking sequences upstream and downstream to these insertions. The amplification patterns of these insertions have showed polymorphisms. By using the oligonucleotide primer pair (BoNLINEF + BoNLINER) flanking the non-autonomous *Bo-N-LINE1* insertion, 22 bands were amplified from various *Brassica* accessions including *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361), *Brassica napus* (New, Mar, Last And Best, Fortune, Drakker, Tapidor), *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and four hexaploids *Brassica* (AABBCC). The amplification pattern showed the insertion amplification from all *Brassica oleracea*, *Brassica napus*, *Brassica carinata* and four hexaploids *Brassica* genomes indicating the C-genome specificity of insertion. In all *Brassica oleracea* accessions, only upper band with insertion was observed, while in *Brassica napus*, *Brassica carinata* and hexaploids *Brassica*, both upper (insertion) and lower bands (pre-insertion sites) were amplified. This suggests that the *Brassica rapa* (AA), *Brassica nigra* (BB) and *Brassica juncea* (AABB) have amplified only lower bands amplifying pre-insertion sites only (Figure 4.6a).

The insertion polymorphisms of *Bo-N-LINE2* revealed its proliferation in C-genome and its polyploids. A 1265 bp band was amplified from 18 *Brassica* accessions including six *Brassica oleracea* accessions (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361), two *Brassica napus* (New, Tapidor), six *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and 4 hexaploids *Brassica* accessions. All the six *Brassica oleracea* (CC) and six *Brassica carinata* (AACC) amplified the upper bands only while two *Brassica napus* and four hexaploids *Brassica* (AABBCC) amplified the upper and lower bands representing two different genomic alleles; one with insertion and other without insertion. A ~150 bp flanking sequence (lower band) was amplified from *Brassica rapa* (AA) and *Brassica juncea* (AABB) but not from any *Brassica nigra* cultivar (BB) (Figure 4.6b).

PCR amplification of 1080 bp *Bo-N-LINE3* was performed from *Brassica* genomes. A total of 23 bands amplifying the insertions were obtained from C-genome specific *Brassica* cultivars from *Brassica oleracea*, *Brassica napus*, *Brassica carinata* and 4 synthetic hexaploids *Brassica* accessions. In contrast, lower bands (210 bp) amplifying the pre-insertion sites were obtained from all *Brassica rapa*, *Brassica nigra* and *Brassica napus* genomes (Figure 4.6c). The primers (BRLINE4F + BRLINE4R; Table 4.2) flanking the *Br-N-LINE4* insertion successfully amplified 38 bands (~920-1144 bp) from *Brassica rapa*, *Brassica nigra*, *Brassica napus*, *Brassica carinata* and 3 hexaploid *Brassica* cultivars (*B. napus* x *B. nigra*). In majority of the genomes, an 1144 bp band was amplified, while is few cases ~920 bp band was amplified. This lower ~920 bp band is BB-genome specific, as observed in its polyploids, such as *Brassica napus* (AABB) and *Brassica carinata* (BBCC) (Figure 4.6d).

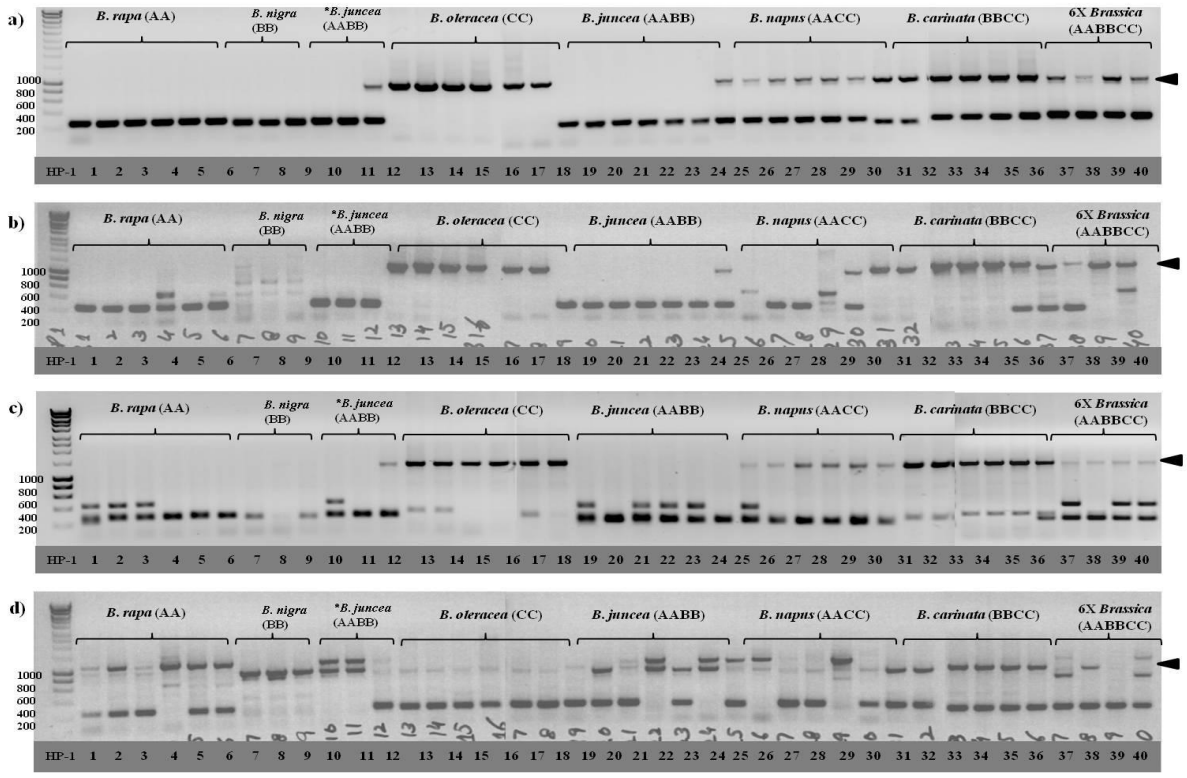


Figure 4.6: Insertional polymorphisms of *Brassica* non-autonomous LINEs showing presence or absence in various *Brassica* accessions. PCR products were obtained with primers hybridizing to the flanking regions of each of the four members a) *Bo-N-LINE1*; b) *Bo-N-LINE2*; c) *Bo-N-LINE3*; d) *Bo-N-LINE4*. The upper bands indicated by arrowheads represent amplification of LINEs, while lower bands lack the amplicons (pre-insertion sites).

4.2.11 Copy number estimation of autonomous LINEs

Approximately 412 and 1026 LINEs from four families were estimated from *Brassica rapa* and *Brassica oleracea* whole genomes respectively. The estimated members from *Rehan* family are 134 and 260 in *Brassica rapa* and *Brassica oleracea* respectively. The *Faizan* family is represented by 104 and 240 copies in A and C-genomes. *Furqan* family of LINEs is considered as low copy number family with 50 and 136 estimated copies in *Brassica rapa* and *Brassica oleracea* respectively. The largest family *Nouman* is estimated to have 124 and 390 copies in *Brassica rapa* and *Brassica oleracea* genomes. It was concluded that the *Brassica oleracea* genomes harbour more than 1 fold LINEs in its genome as compared to *Brassica rapa*. We also speculate that the number of LINEs in *Brassica rapa* and *Brassica oleracea* is more than this estimate, as several truncated and partial LINEs were found distributed in *Brassica* genomes.

4.3 Identification of novel families of SINEs in *Brassica* genomes

4.3.1 SINE identification by comparative sequence analysis

Novel SINE insertions were identified from *Brassica* genomes by homoeologous BAC sequence comparison and their homologues were collected from the GenBank database. By comparing *Brassica rapa* and *Brassica oleracea* accessions (AC189298.1 x EU642504.1), a SINE insertion was detected in *Brassica rapa* and 4 SINEs in *Brassica oleracea*. Similarly, the comparison of *Brassica rapa* (AC155341.2) x *Brassica oleracea* (AC240089.1) and *Brassica rapa* (CU984545.1) x *Brassica oleracea* (EU579455.1) led to the identification of 4 and 1 SINE insertions respectively (see Conclusion; Figure 10.1-10.3). The newly identified SINEs (reference) were used as query in BLASTN searches to identify other relatives residing in *Brassica* species. The sequences were considered as members of the same family, if they generate host TSDs, poly(A) tail at 3´ terminus and >75% coverage in entire lengths. The phylogenetic analysis of all collected SINEs revealed that they clustered into 10 different groups or families. The families were named as *BoSINE1-BrSINE10* (Table 4.3).

4.3.2 Estimation of full length SINE copy numbers from whole *Brassica* genomes

SINEs are diverse retroposons present in the *Brassicaceae*. As only 9% and 1% sequence data is available for *Brassica rapa* and *Brassica oleracea* respectively (before February, 2012, as complete BAC sequences where there are few gaps compared to the genomic sequence of *Brassica rapa* available at other websites) in *Brassica* Nucleotide Collection database at NCBI, 143 intact copies were collected, that displayed >70% identity to the query over their entire length. The total numbers of SINEs in *Brassica rapa* and *Brassica oleracea* are estimated as 1440 and 2210 respectively. The copy number of each SINE family was also estimated and low, middle and high copy number families were identified. *BrSINE10* is the largest and highly diverse family of SINEs with 505 and 450 copies in *Brassica rapa* and *Brassica oleracea* respectively. It is the only family where estimated number of copies is higher in A-genome compared to C-genome. *BoSINE8* is considered to be the second abundant family displaying 356 and 510 intact copies from A and C-genomes. *BoSINE9* is the smallest family with 26 and 74 estimated copies in *Brassica rapa* and *Brassica oleracea* whole genomes followed by *BoSINE7* (Table 4.4).

4.3.3 Structural features of *Brassica* SINEs

Like other SINEs described in various plants, the *Brassica* SINEs are small in sizes with typical SINE features displaying TSDs, head regions, internal regions (body) of variable sizes and a poly(A) tail at the 3′ terminus. The structural features of all *Brassica* SINE families are more or less similar. The smallest SINE investigated is a member of *BoSINE2* family, which is 206 bp in size, while larger elements belong to BoSINE9. *BoSINE1* has 10 members from 213-225 bp, flanked by TSDs of 7-14 bp and terminated with a 3′ poly(A) tail of 19-21 bp. The first SINE (*BoSINE1-1*) from this family was identified as an insertion residing in *Brassica oleracea* (EU642504.1) sequence. The size of the *BoSINE1-1* is 216 bp, terminated by 5′-CAAAAAAAAAAAAAAAAAA-3′ C-terminal end and flanked by 14 by TSDs (Figure 4.7). The *BoSINE2* family presents a low copy number family with members having sizes from 206-219 bp, flanked by TSDs of 13-18 bp and polyadenylation signals of 10-27 bp at their 3′ terminal end. A 206 bp smallest SINE (*BoSINE2-3*) belongs to this family. The first element (*BoSINE2-1*) from this family was identified in *Brassica oleracea* (EU642504.1), where a 219 bp insertion was found flanked by 18 bp TSDs and a tail terminating with $CTT(A)_8$ (Table 4.3 & 4.4).

The family *BoSINE3* presents the members ranging in sizes from 256-277 bp including TSDs of 10-17 bp and terminating by a poly(A)$_{\mathbf{9\text{-}11}}$ tail at their carboxylic terminal ends. The well characterized member is a 272 bp *BoSINE3-1* having a 13 bp TSDs and 9 bp poly(A) tail. The members of family *BoSINE4* generally range in sizes from 361-397 bp with the exception of *BoSINE4-1*. The elements are flanked by TSDs of 07-15 bp (except *BoSINE4-1*) and terminated with 8-34 bp poly(A) tail. *BoSINE4-1* is the first element described from this family as an insertion of 442 bp residing in *Brassica oleracea* (EU579455.1) accession. It generates largest TSD of 42 bp, but another inner 15 bp TSDs are also present in its genome. With 42 bp TSDs the size of the element is 442 bp while with short TSDs (15 nt), the size of the element is 361 bp. It was concluded that the longest TSDs might be the result of an error during the 5′ and 3′ host DNA nicking and integration of the element to a new site. Fourty two bp TSDs were identified by viewing the flanking regions of insertion in dot plot analysis, otherwise the other homologues of *BoSINE4-1* generates 8-17 bp TSDs (Table 4.3 & 4.4).

The sequences from *BoSINE5* are similar in sizes (225-229 bp), flanked by short TSDs (3-4 bp) and a poly(A) tail of 8-11 bp. The first element was characterized from *Brassica oleracea* (AC240089.1) as a 225 bp insertion including 4 bp TSDs and 5′-CAAAAAAAA-3′ C-terminal tail. *BoSINE6* family of *Brassica* SINEs represents 321-335 bp large members generating TSDs (05-11) and having poly(A) tail. *BoSINE6-1* is the first and well characterized member of the family with a size of 335 bp including 11 bp TSDs at both ends and an 18 bp polyadenylation tail at its C-terminal end. A low copy number family *BoSINE7* is characterized by having representatives ranging in sizes from 392-401 bp, including TSDs (3-8 bp) and a poly(A)$_{\mathbf{11\text{-}13}}$ (Table 4.3 & 4.4). The first identified member from the family is *BoSINE7-1* from *Brassica oleracea* (AC240089.1) residing as an insertion. The element is 401 bp in size including 8 bp TSDs at both ends and polyadenylation signals of 8 nucleotides. The second largest family of *Brassica* SINEs *BoSINE8* represents diverse members dispersed in *Brassica rapa* and *Brassica oleracea* genomic sequences. The elements range in sizes from 480-506 bp including the host target site duplications (5-13 bp) and poly(A)$_{11\text{-}28}$ tail adjacent to C-terminal end. Generally the terminal tail have 11-19 bp poly(A) stretch but few elements generate a longer stretch (21-27 bp). *BoSINE8-1* represents the first identified member of the family from *Brassica oleracea* (AC240089.1) accession with 13 bp TSDs and a 5′-CAAAAAAAAAAA-3′ C-terminal tail (Figure 4.7; Table 4.3 & 4.4).

Table 4.3: Full length SINEs identified by comparative dot plot analysis of *Brassica* BAC sequences.

No.	Reference Elements	Family	BAC Accession	Species	Size	TSD	Poly (A) Tail	GC%
1	*BoSINE1-1*	*BoSINE1*	EU642504.1	*B.oleracea*	216	14	CAAAAAAAAAAAAAAAAAA	52.0
2	*BoSINE2-1*	*BoSINE2*	EU642504.1	*B.oleracea*	219	18	CTTAAAAAAAA	48.4
3	*BoSINE3-1*	*BoSINE3*	EU642504.1	*B.oleracea*	272	13	CAAAAAAAAA	47.1
4	*BoSINE4-1*	*BoSINE4*	EU579455.1	*B.oleracea*	443	44	CAAAAAAAA	37.0
5	*BoSINE5-1*	*BoSINE5*	AC240089.1	*B.oleracea*	225	04	CAAAAAAAA	37.3
6	*BoSINE6-1*	*BoSINE6*	AC240089.1	*B.oleracea*	335	11	CAAAAAAAAAAAAAAAAAA	37.3
7	*BoSINE7-1*	*BoSINE7*	AC240089.1	*B.oleracea*	401	08	CAAAAAAAAAAA	41.6
8	*BoSINE8-1*	*BoSINE8*	AC240089.1	*B.oleracea*	484	13	CAAAAAAAAAAA	37.6
9	*BoSINE9-1*	*BoSINE9*	EU642504.1	*B.oleracea*	524	11	CAAAAAAAAAA	48.1
10	*BrSINE10-1*	*BrSINE10*	AC189298.1	*B.rapa*	376	13	CGTTAAAAAAAAA	41.0

4.3.4 Structural features of low and high copy number SINE families

A 524 bp insertion (*BoSINE9-1*) flanked by 11 bp TSDs and a poly(A) tail yielded no significant hits but the NCBI EST database yielded two sequences with >85% identity in their entire lengths. The retrieved sequences were *Brassica napus* cDNA, mRNA sequences and designated as *BnSINE9-2* and *BnSINE9-3*. The elements were 558 bp in sizes generating 6 bp TSDs and a largest poly(A) tail comprising 50 adenine and a single guanine nucleotide. The copy number estimation in *Brassica rapa* (26) and *Brassica oleracea* (74) suggests that this is the lowest copy number family of SINEs studied in present work. The largest family is *BrSINE10* with 505 and 450 members in A and C-genomes respectively. *BrSINE10-1* is a 376 bp large SINE including 13 bp TSDs and a 5′-CGTTAAAAAAAAA-3′ tail. The BLASTN searches against GenBank database retrieved about 50 full length elements, of which 47 were from *Brassica rapa* and 3 were from *Brassica oleracea* BAC sequences. The high copy numbers in *Brassica rapa* is due to the availability of high percentage of available sequenced data (51.3 Mbp) as compared to *Brassica oleracea* (4.7 Mbp). The representative of *BrSINE10* family range is sizes from 368-378 bp including TSDs (5-15 bp) and a poly(A) tail of 9-15 bp. A 5 bp conserved motif (TCAGC) was observed in majority of the elements adjacent to poly(A) tail of the elements (Table 4.4).

4.3.5 Transposon insertional polymorphisms (TIPs) of SINEs in *Brassica* genomes

The distribution and abundance of SINEs in various *Brassica* species was investigated by TIP based PCR markers. A total of 40 *Brassica* accessions/cultivars were tested for the presence or absence of SINEs at a particular site/locus. The collinear sequences from A and C-genome *Brassica* are highly similar with few gap points indicating an insertion in one but lacking in other. Based on this, higher bands were amplified with insertions and lower bands amplifying the pre-insertion sites (flanking regions). Four set of primers were used to amplify four SINE families among various *Brassica* species. *BoSINE2-1* (219 bp) was found inserted in *Brassica oleracea* (EU642504.1) accession from 51788-52006 bp. A primer pair BoSINE2F 5′-GAACAAGAAAAATGCAGGG-3′ and BoSINE2R 3′-CGTACCATCACATCTCTTTC-5′ was designed to amplify it. It gave large (insertion) and small (pre-insertion site) bands in various *Brassica* accessions. *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso and GK97361) and *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67 and NARC-PK) accessions amplified *BoSINE2-1* and lower bands (without insertion) indicating they are heterozygous with a site in the C-genome, while B-genomes have the flanking sequence but lack insertions. Weak bands from all the six *Brassica napus* (New, Mar, Last and Best, Fortune, Drakker, Tapidor) were not further characterized because of time constraints. All the other *Brassica* LINEs amplified the lower product with pre-insertional sites (Figure 4.8a). A 272 *BoSINE3-1* was tested for its presence in various *Brassica* accessions. The results showed its amplification from all six *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso and GK97361) and *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) accessions. Weak bands were detected in 3 *Brassica* hexaploids (*B. napus* x *B. nigra*) (Figure 4.8b). The *Brassica rapa*, *Brassica nigra*, *Brassica juncea* and *Brassica napus* accessions amplified only pre-insertional sites.

A 524 bp *BoSINE9-1* was identified residing in *Brassica oleracea* (EU642504.1) accession from position 37170-37693 bp. The BLASTN searches against *Brassica* EST database returned only 3 sequences; one from the same accession (EU642504.1) and two others from *Brassica napus* cDNA sequences. The primers were designed from the flanking regions and were tested against 40 *Brassica* genomes. Interestingly, only one *Brassica oleracea* accession GK97361 produced the expected product size (735 bp). The

amplicon was sequenced and aligned with *BoSINE9-1* achieving >98% identity. All other *Brassica* accessions except *Brassica nigra* amplify the pre-insertion sites (~210 bp) (Figure 4.8c). The primers designed from flanking regions of *BrSINE10-1* produced upper and lower bands to indicate the presence and absence of *BrSINE10* family in 40 *Brassica* accessions belonging to 6 different species. The PCR results confirmed the abundance of *BrSINE10* is various *Brassica* genomes. The amplification of the *BrSINE10-1* was seen in *Brassica rapa* (AA) species and the their allotetraploids having A-genome in them. Many of the genomes amplified additional bands of varied sizes other than the expected product. The bands after sequencing showed similarity to the *BrSINE10* family. It was concluded that multiple copies of the element are dispersed in *Brassica rapa* genomes (Figure 4.8d).

Table 4.4: Average lengths, TSDs, Pre-tail motifs and estimated copy numbers of each SINEs family in *Brassica*. The name of the family is given on the basis of the first element identified in *Brassica*. ECN: Estimated copy numbers.

No.	Family Name	Size of elements	TSDs	Pre-tail Motifs	C (A)n	Strong Hits	ECN in whole *B. rapa*	ECN in whole *B. oleracea*
1	*BoSINE1*	213-225	07-14	TTATC	$C(A)_{19-21}$	10	92	130
2	*BoSINE2*	206-219	13-18	TTATC	$C(A)_{10-17}$	5	55	260
3	*BoSINE3*	256-277	10-17	TTTTC	$C(A)_{9-11}$	10	100	135
4	*BoSINE4*	361-443	07-15/44	TTAGC	$C(A)_{8-17}$	12	115	145
5	*BoSINE5*	225-229	03-04	TTTTC	$C(A)_{08-11}$	5	62	128
6	*BoSINE6*	321-335	05-11	TTAGC	$C(A)_{16-18}$	6	25	245
7	*BoSINE7*	392-401	03-08	TTACC	$C(A)_{11-13}$	4	52	138
8	*BoSINE8*	480-506	05-13	TTGTC	$C(A)_{11-27}$	38	356	510
9	*BoSINE9*	524-558	05-11	TGATC	$C(A)_{10-12/51}$	3	26	74
10	*BrSINE10*	368-378	05-15	TCAGC	$C(A)_{9-15}$	50	505	450
					Total	**143**	**1440**	**2210**

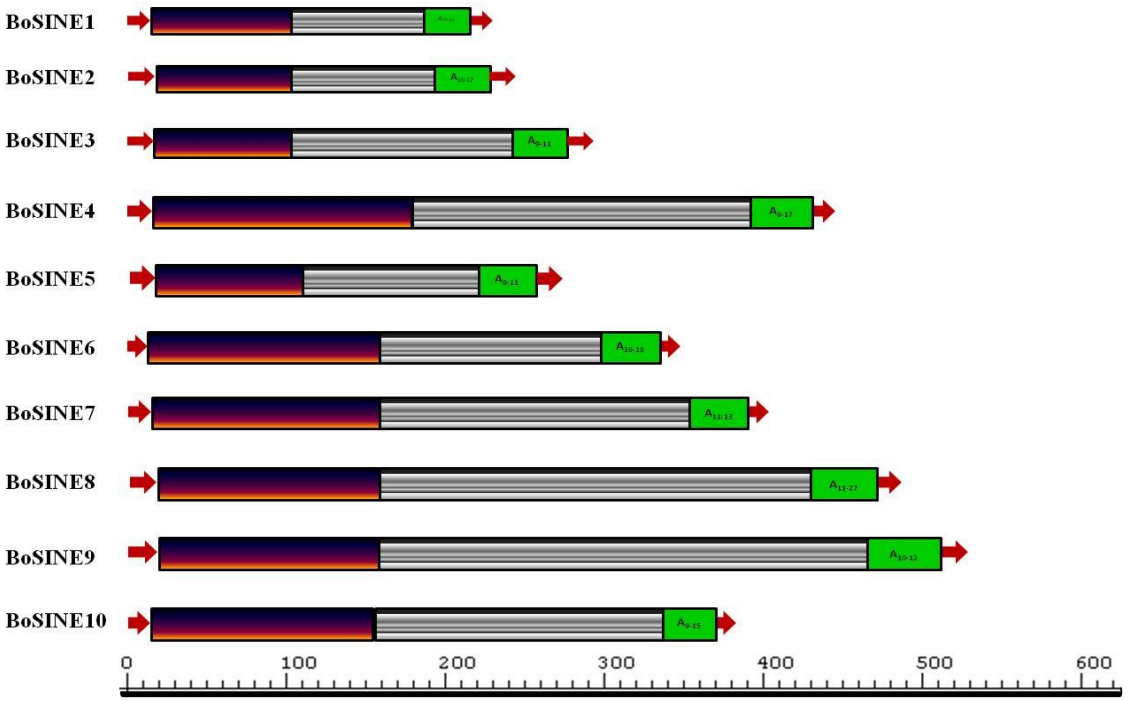


Figure 4.7: Schematic representation of *Brassica* SINE families. SINEs are composed of a tRNA-derived region (coloured in 5´ terminal end), an unrelated DNA sequence (light grey) and a LINE-related region or tail (green box). The variable sized TSDs are represented by red arrows at both terminal ends. Scale in bp.

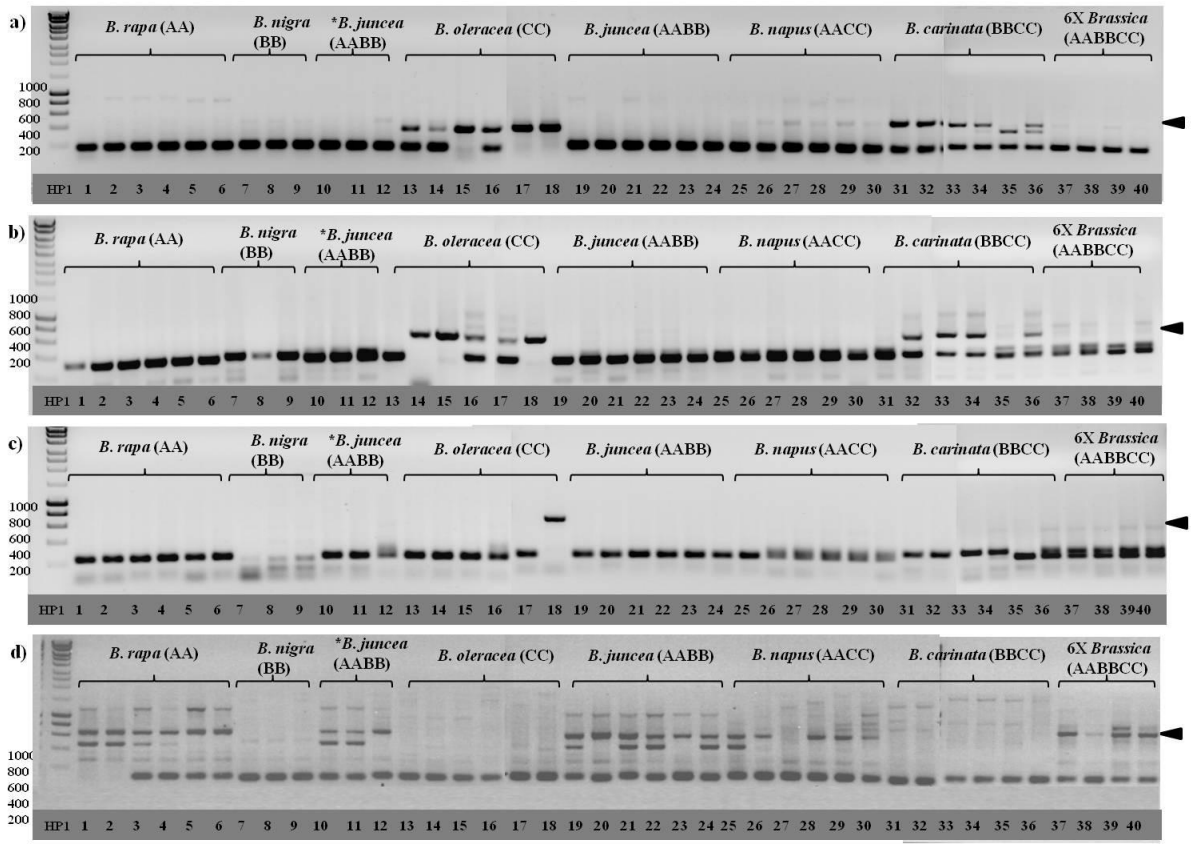


Figure 4.8: Transposon insertional polymorphism (TIP) of various SINE families: a) *BoSINE2*; b) *BoSINE3*; c) *BoSINE9*; d) *BrSINE10*. Higher bands indicate amplification of specific SINEs, while lower bands represent the pre-insertion sites (sites without SINE insertions).

4.4 Discussion

The results here show the value of comparison of BAC sequences for identification of the full range of LINE and SINE-like transposable elements on the basis of their activity and homology. Previous methods have required assumptions about motifs and structures; the results show that these methods were efficient in that most elements would have been identified, although a few novel groups not represented in repeat element databases were found here. These will be valuable in annotation and assisting in assembly of whole genome shotgun (WGS) sequencing data in the future. Current WGS approaches have difficulty in assembling LINE and SINE rich regions of the genomes, even using paired-end strategies, where long and duplicated elements, particularly when heterozygous, prevent conting ends from being overlapped unequivocally. In *Brassica* sequencing projects, due to partial assembly, it is difficult to study the full copies of TEs or novel elements present in the genome; in the assembly of date palm, it is notable that the raw reads include. Here, four novel families of each autonomous and non-autonomous LINEs were detected distributed among *Brassica* crops. Similarly, 10 SINE families were identified and characterized from *Brassica* genomes. The structural features and distribution of the elements were studied in detail by computational and molecular analysis. The analysis confirmed that retroposons are a diverse group of TEs scattered among *Brassica* and closely related genera *Arabidopsis*.

4.4.1 Reverse transcriptase is the most conserved region among plants LINEs

Higher similarities are observed when comparing amino acid sequences of the RT from various plant LINEs. The overall similarity observed in two variable sequences is ~47% while the similarity between two close members is >96%. High homology is observed in the members of *Rehan* family (83-96%) at amino acid level. The homology between *Brassica and Arabidopsis* RT sequences are >80%, while the homology of *Brassica* LINEs with grass family is upto 56%. Several conserved motifs were also observed in RT region from *Brassica*, *Arabidopsis* and other plants (Figure 4.3). In studies supported by the work here, RT regions of LINEs from various organisms were aligned, which revealed 11 conserved blocks of identity. Of these 11 blocks, 7 blocks are highly conserved among all LINEs collected from various species (Malik *et al.*, 1999).

4.4.2 RNase H containing LINEs are dominant in *Brassica*

In the present study, the majority of the *Brassica* LINEs identified display an RH domain at their C-terminal end. Of the four autonomous LINE families, the three families have structures encoding RH in their *pol* gene except *Furqan* family, where all the elements encode only EN and a RT in their *pol* gene. The detail structural features of these LINEs indicate that the RH domain is present in extreme C-terminal end upstream to the poly(A) tail. The RH domain is composed of ~300-400 bp in various members of LINE families. The protein domain analysis of plant LINEs revealed the encoding of EN and RT domains, which is a typical feature of L1 of human LINEs, identified from various plants (Heitkam and Schmidt, 2009). Very few plant LINEs have shown such features encoding an RNase H (RH) domain. In a detail analysis of all LINEs described, it was reported that only four clades designated as Tad, R1, LOA and I contain RNase H domains (Malik *et al.*, 1999). *Brassica* LINEs encoding an RNase H domain were considered close to Tad, R1, LOA and I clades due to similar structural and domain organization.

4.4.3 SINEs display a conserved motif before poly(A) tail

The SINEs studied from *Brassica* and *Arabidopsis* genomes display a conserved motif upstream to their poly(A) tail at 3′ terminus. The motif is generally AT rich and is highly conserved within the family members and across various families. It mostly starts with T and ends with a C nucleotide (Table 4.4). The structural analysis of other known *Brassicaceae* SINEs, including the S1, AtSN1/RAtheE3 from *Brassica*, AtSN2/RAtheE1 and RAthE2 from *Arabidopsis thaliana* showed a conserved motif upstream to poly(A) tail at 3′ terminus (Deragon *et al.*, 1994; Lenoir *et al.*, 2001; Myouga *et al.*, 2001). This suggests that all the SINEs from *Brassicaceae* share more or less similar motifs at their pre-tail ends. The highly characterized *BoS* family of SINEs from *Brassicaceae* exhibit a conserved motif (TTATC) immediately upstream to 3′ terminal end. Two bp purines residues are located upstream to this motif (Zhang and Wessler, 2005).

4.4.4 Target Site Preference of *Brassica* SINEs

The SINEs have shown a preference for their insertion to a heavily populated AT rich region. To determine the SINEs insertional preference within the genomes, SINE

insertions with extra 20 bp flanking nucleotides on both ends were collected and aligned to observe their insertion preference. Analysis of insertion SITEs of *Brassica* SINEs revealed that all the members belonging to 10 families of SINEs have shown an insertion preference in AT rich regions. The 5′ end of the SINEs displayed the strong preference of AT rich regions as compared to the 3′ end. *BoSINE1* family showed an insertion preference of AT rich regions at both ends like other families. *BoSINE3*, *BoSINE4*, *BoSINE7* and *BrSINE10* families are inserted into highly AT rich regions while *BoSINE8* showed purines rich flanking regions (Figure 4.9). The *BoS* elements from *Brassica* had shown insertional preference in AT rich regions (Zhang and Wessler, 2005).

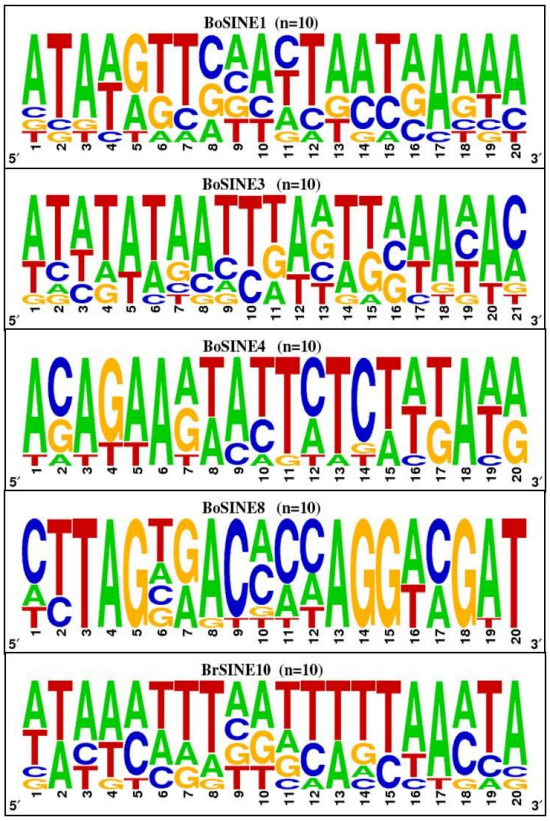


Figure 4.9: Frequency plot indicating the insertional preference of five SINE families into AT rich regions. The weblogo indicates the SINE preference for AT rich regions created using 20 bp of aligned sequence flanking the SINE insertions for each family.

4.4.5 *Brassica* SINEs are ancient retroposons in *Brassicaceae*

The SINE families investigated in this study are considered to be the old families that were present before the separation of *Arabidopsis-Brassica* species. This can be confirmed by the high similarity of *Brassica* SINEs with the *Arabidopsis genomes*

(~75%). The BLASTN searches against the GenBank database retrieved many sequences from *Arabidopsis thaliana* and *Arabidopsis lyrata*. Some of the hits from *Arabidopsis* showed very high sequence similarity with *Brassica* SINEs suggesting their common origin from the same ancestor. *Brassica* SINEs were considered as old as the divergence of *Arabidopsis-Brassica* (16-19 Mya) occured. It is believed that *Arabidopsis-Brassica oleracea* diverged 16-19 million years ago (Myo) from a common ancestor (Deragon and Zhang, 2006). *BrSINE10* is considered to be the youngest family due to high homology within its members. *BoSINE1*, *BoSINE3*, *BoSINE4*, and *BoSINE8* are considered as middle aged families while *BoSINE9* is considered to be recently introduced due to fewer copies and high homology (84-88%) between sequences. The *Brassica BoS* family is also an old family, whose members are dispersed among various *Brassica* species suggesting their divergence before the separation of these species. It is thought the oldest members have diverged ~20 Mya, whereas the youngest members have originated ~2-3 million years ago (Zhang and Wessler, 2005).

4.4.6 SINEs as molecular markers in phylogenetic studies

SINEs can be used as molecular markers to investigate the evolutionary relations of species or to trace phylogeny. The SINEs can be used in two different ways in phylogenetic studies. The first approach is the identification of a specific SINE family in species by PCR analysis or dot hybridization. All species which display the presence of a specific SINE family are treated as close to each other than to other species, which lack them. The second approach is the site specific insertional polymorphism, where the PCR primers are designed from the common flanking regions around SINEs. The species having SINE insertions generate higher products, while those who lack generated the shorter product (Deragon and Zhang, 2006). Species sharing the SINE insertions are considered to be close as compared to others, who lack them (Kramerov and Vassetzky, 2011). Similar methodology was use to observed the presence/absence of SINEs at various loci in the present studies. The species amplifying a specific SINE family are more close to the others lacking them. Thus *Brassica oleracea* (CC) and *Brassica carinata* (BBCC) are closer to each other as compared to *Brassica juncea* (AABB), which failed to amplify *BoSINE3*. The results revealed that the SINEs can be used as molecular markers in evolutionary studies and to trace the phylogeny among various species.

4.5 Conclusions

Non-LTR retrotransposons or retroposons are present in enormous numbers in all eukaryotic genomes. They have played a major role in genome diversification and evolution. Despite the progress in our understanding of retroposon biology, many aspects remain unclear. We set the trends in the field with the identification and characterization of LINEs and SINEs among *Brassicaceae*, with an emphasis on evolutionary relationships of retroposons from *Brassicaceae* and other plants. The insertional polymorphisms of retroposons were explored indicating their absence in some species and abundance in other ones and also used them in a wide range of phylogenetic studies. The analysis will help in annotation and characterization of several related retroposons from plant families and will be used as molecular markers to study the diversity among closely related cultivars and varieties.

CHAPTER 5
CACTA AND HARBINGER DNA TRANSPOSONS: CHARACTERIZATION AND IMPACT ON *BRASSICA* GENOMES

Summary

CACTA and Harbinger are diverse superfamilies of DNA transposons. A combination of dot plot analysis and BLASTN searches led to the identification of 35 autonomous and 7 non-autonomous CACTA, and five autonomous and several non-autonomous Harbinger elements in *Brassica*. The PCR analysis amplified the CACTA and Harbinger transposases from 40 and 38 *Brassica* genomes respectively suggesting their abundant distribution among various *Brassica* crops. A detailed characterization and evolutionary analysis of the identified elements allowed some to be placed in genome-specific groups. The protein domains of transposons from *Brassica* and other plants revealed similar organizations with minor differences. Both transposases (TNPD, TNPA) are present in most CACTA elements, while a few CACTA harbour an additional ATHILA ORF1-like domain in opposite orientation. The autonomous Harbinger has a transposase and 1 or 2 additional SANT and NAM-like putative DNA-binding protein motifs. The TIRs of both CACTA and Harbinger are highly conserved and can be used to differentiate the superfamilies. The high copy numbers of CACTA and Harbinger in *Brassica* led to the conclusion that 3 bp generating transposons (CACTA and Harbinger) contribute significantly to genome size and evolution of *Brassica* genomes.

5.1 Introduction

CACTA, also called En/Spm, elements constitute a diverse group of DNA transposons identified from various plants and include *Caspar* from *Triticum* (Sergeeva *et al.*, 2010), *Tam1* and *TamRS1* from snapdragon (*Antirrhinum majus*) (Nacken *et al.*, 1991; Roccaro *et al.*, 2007), En/Spm from maize (Gierl, 1996), soybean (Zabala and Vodkin, 2008; Xu *et al.*, 2010), CAC1 from *Arabidopsis thaliana* (Miura *et al.*, 2001), Ps1 from *Petunia hybrida* (Snowden and Napoli, 1998), Pis1 from *Pisum sativum* (Shirsat, 1988), Tnr3 and Tnr1 from *Oryza sativa* (Motohashi *et al.*, 1996; Han *et al.*, 2000), *Caspar* from *Triticeae* (Wicker *et al.*, 2003) and the non-autonomous elements as the maize *dSpm* (Gierl, 1996). The CACTA superfamily of DNA transposons received its named due to the conserved

'CACTA' DNA sequence signature in termini of their TIRs. CACTA elements are flanked by 3 bp TSDs, 10-28 bp TIRs (with CACTA in their termini) and DDD/E type transposase. En/Spm elements are the autonomous elements, while there non-autonomous partners I/dSpm lack transposase enzyme (Wicker *et al.*, 2003; Tian, 2006). The CACTA elements are used as molecular markers in many crops as in maize, where the markers were developed from TIRs of *Issac*-CACTA transposons, which distinguished the maize imbred lines (Lee *et al.*, 2005).

The PIF-Harbinger is a superfamily of DNA transposons characterized by generating 3 bp TSDs, flanked by 14-25 or up to 50 bp TIRs and a DDD/E transposase (Kapitonov and Jurka, 2004). The diverse PIF-Harbinger elements are easily distinguishable into two subgroups, named *PIF* and *Pong* (Zhang *et al.*, 2004). Harbinger is highly diverse superfamily of DNA transposons with members distributed among protists, insects, worms, vertebrates and plants. Their autonomous elements encode two protein domains; superfamily specific transposase and DNA-binding domain. The DNA binding domain is characterized by having different conserved motifs as SANT/myb/trihelix (~70 aa), while the other region of DNA binding domain showed no significant similarities studied in different species (Kapitonov and Jurka, 2004). Generally, the Harbinger are flanked by TAA/TTA target site duplications, but some families generate other TSDs, as CAG target sites observed in Zebra fish Harbinger2-3_DR. The phylogenetic studies based on Harbinger-transposases suggest their horizontal transfer. As, the transposases from the *Arabidopsis* and maize Harbinger and *PIF* elements are more similar to diatom Harbinger *1-2_TP* transposase as compared to their closely related rice *Pong* and *Arabidopsis ATIS112A.* The *PIF* and Harbinger were considered as two separate superfamilies prior to 2001, which merged to a single superfamily due to high similarities between the elements (Jurka and Kapitonov, 2001; Kapitonov and Jurka, 2004).

In this chapter, the aim was to identify CACTA and Harbinger class II DNA transposons in *Brassica* genomic sequences, and to analyse their structures (including the internal regions encoding the protein domains like transposase and its associated domains), the evolutionary diversity, mobility and consequences for *Brassica* genome organization. The structural and evolutionary relationships of *Brassica* CACTA and Harbinger were compared with CACTA and Harbinger identified from other crop species.

5.2 Results

5.2.1 CACTA identification and characterization by Dot plot and BLASTN analysis

The identification of transposable elements by dot plot comparison of *Brassica* homoeologous BAC sequences led to the identification of various insertions flanked by 3 bp TSDs. The detailed structural and molecular analysis revealed the identification of CACTA transposons. A CACTA transposon was identified by comparing *Brassica rapa* subsp. pekinensis clone KBrB028I01 (AC189298.1) against its homoeologue *Brassica oleracea* var. alboglabra clone BoB028L01 (EU642504.1). The element was autonomous CACTA and is named as *BoCACTA1*. Two other I/dSpm or non-autonomous CACTA elements were identified by comparing *Brassica rapa* (AC155341.2) against its homoeologous *Brassica oleracea* (AC240089.1). The BLASTN analysis of autonomous *BoCACTA1* retrieved several homologues from *Brassica rapa* and *Brassica oleracea* with homology of 50-100%. Our identified elements *BoCACTA1, BoCACTA2* and *BoCACTA3* were found to be the *Bot1-1, Bot1-2,* and *Bot1-3* elements identified by Karin Alix and her co-authors in *Brassica oleracea* (Alix *et al.*, 2008). According to their findings, *Bot1* elements are *Brassica oleracea* specific and have played a major role in the divergence of *Brassica* genomes. Many other homologues of *Bot1* were isolated and characterized from various *Brassica* BAC sequences on the basis of conserved CACTA TIRs. All identified CACTA from *Brassica rapa* and *Brassica oleracea* were collectively included in *Bot1* family due to their similarity with *Bot1-1-Bot1-3* elements. A total of 35 autonomous CACTA elements were identified, out of which 19 were from *Brassica oleracea*, 14 from *Brassica rapa* and two from *Brassica napus* BAC clone sequences. Seven non-autonomous CACTA elements were isolated and characterized from different *Brassica* BAC clone sequences, which were further blast to find their autonomous copies to relate their progenitors. The autonomous elements range in sizes from 3 kb to 11 kb. Non-autonomous CACTA are smaller in sizes ranging from 1.2 kb to 3.2 kb (Table 5.1).

5.2.2 Structural features of *Brassica oleracea* CACTA elements

The *BoCACTA* and related homologues display all the characteristics of CACTA transposons including the 3 bp TSDs, TIRs of 15-17 bp (mostly15 bp), CACTA signature in their termini of TIRs and possessing two transposase named TNPD and TNPA. *BoCACTA1* (*Bot1-1*) was identified by comparing *Brassica rapa* accession AC189298.1

against its homeologue *Brassica oleracea* accession 'EU642504.1', where it is residing from position 20580-29972 bp in 3′-5′ orientation. *BoCACTA1* is 9399 bp large in size including 3 bp TSDs. It has perfect 15 bp TIRs (5′-CACTACAAGAAAACA-3′) and a CACTA signature and its reverse compliment at the termini of TIRs. The element has both transposase TNPD and TNPA at N-terminal and C-terminal ends respectively. A transposase associated domain (TAD) is present towards the N-terminal end of TNPD transposase. Similarly two domains named DUF4218 and DUF4216 are present towards the C-terminal end of TNPD (Figure 5.1). The function of these domains is unknown but their presence in all identified CACTA indicates that they are accessory domains, which aid transposase in the transposition and integration of CACTA elements. The nearest homologues of *BoCACTA1* are *BoCACTA2* and *BoCACTA3*, which are synonym of *Brassica oleracea Bot1-2* and *Bot1-3* CACTA elements of Alix (Alix *et al.*, 2008). *BoCACTA2* was identified from *Brassica oleracea* accession 'EU642505.1' from position 44789-55702 bp within BAC sequence. It is 10914 bp in size with 3 bp TSDs at both ends and 15 bp 5′-CACTACAAGAAAACA-3′ TIRs. The genome of *BoCACTA2* displays the presence of both transposase TNPD and TNPA. The transposase associated domain is present in N-terminal of TNPD while DUF4218 and DUF4216 are located at C-terminal end of TNPD and N-termini of TNPA (Figure 5.1; Table 5.1).

The largest *Brassica* CACTA identified in present study is *BoCACTA3*, identified from *Brassica oleracea* accession 'EU642506.1' from 19777-30844 bp (Figure 5.1). The element is 11068 bp including 3 bp TSDs generated at both terminal ends. The element has 15 bp perfect 5′-CACTACAAGAAAACA-3′ TIRs and several sub-terminal repeats in terminal 80 bp. According to Alix *et al*., (2008), the *Bot1-3* has 64 bp TIRs but in present study, first 15 bp were considered as TIRs and the other discontinuous repeats as sub-terminal repeats. This is based on 15 bp conserved TIRs in almost all *Brassica* CACTA investigated in *Brassica*. The genome of *BoCACTA3* displays only the presence of transposase TNPD. The TAD domain is present in N-terminal end of TNPD, while DUF4218 and DUF4216 are located at C-terminal end of TNPD. There is another extra domain named DUF7241 present in sub-terminal region of 3′ terminal end (Figure 5.1). Interestingly both elements *BoCACTA2* and *BoCACTA3* capture an ATHILA ORF-1 domain, which is integral component of Ty3/gypsy LTR retrotransposons identified in *Arabidopsis thaliana*. The detailed analysis showed that the ATHILA ORF-1 is present in opposite orientations in comparison to the CACTA protein domains (Figure 5.1). Two

other copies of CACTA designated as *BoCACTA4* and *BoCACTA5* were detected in *Brassica oleracea* accession 'EU642505.1' from 21474-29678 and 78098-85744 bp respectively. *BoCACTA4* and *BoCACTA5* are 8205 and 7647 bp in sizes with 3 bp TSDs and TIRs of 15 bp. Two other CACTA named *BoCACTA18* and *BoCACTA30* are 10682 and 10728 bp respectively display the similar structure with capturing ATHILA ORF-1 domain in opposite orientation of their coding region (Table 5.1 & 5.2).

During a dot plot analysis for the identification of retrotransposons, a 7265 bp CACTA element was identified from *Brassica oleracea* accession 'EU579455.1'. The element is named as *BoCACTA19*, which posses 3 bp TSDs and 15 bp TIRs similar to other *Brassica oleracea* CACTA elements. The blast analysis of this sequence provided many other copies from *Brassica oleracea*. The element display TNPD transposase with its associated domain (TAD) and ATHILA ORF-1 domain in opposite orientations. *BoCACTA21* and *BoCACTA22* are 8210 and 7170 bp large elements with 3 bp TSDs and 15 bp (5′-CACTACAAGAAAACA-3′) TIRs. They encode both transposase proteins TNPD and TNPA with their associated domains without capturing ATHILA ORF-1 domain (5′-TAD-TNPD-DUF4218-DUF4216-TNPA-3′). *Brassica oleracea* accession (AC183496.1) harbour four complete copies of CACTA (*BoCACTA30-BoCACTA33*). *BoCACTA30* is the largest (10728 bp) with ATHILA ORF-1 domain in its genome while *BoCACTA31* and *BoCACTA32* are 7157 and 6075 bp large is size including 3 bp TSDs and 15 bp TIRs with internal region encoding both transposase proteins and associated domains (Table 5.1). *BoCACTA33* is a 5916 bp CACTA with typical *Brassica* CACTA TSDs and TIRS but encoding only a TNPD with its associated domain TAD.

5.2.3 Molecular characterization of *Brassica rapa* CACTA elements

The homologues of *BoCACTA1* (*Bot1*) were also detected in *Brassica rapa* genomes. The first A-genomic CACTA (*BrCACTA9*) was identified from *Brassica rapa* accession 'AC172883.2' as an insertion from 114211-122180 bp with 3 bp TSDs and 15 bp TIRs (5′-CACTACAAGAAAACA-3′). The blast searches retrieved many other copies with homology in their internal regions. The elements showing >60% homology in their entire lengths and >70% homology in coding regions were collected. The 15 bp TIRs and their reverse complements were used to define 5′ and 3′ end respectively. Fourteen intact autonomous CACTAs were identified from *Brassica rapa* genomes (*BrCACTA6, BrCACTA7, BrCACTA9-BrCACTA17, BrCACTA26, BrCACTA34* and *BrCACTA35*).

They have shown similar TIRs as observed in *Brassica oleracea* CACTA elements. The largest among the *Brassica rapa* CACTA is *BrCACTA6* residing in *Brassica rapa* accession 'AC189480.2'. The element is 9393 bp in size generating 3 bp TSDs and 15 bp TIRs (5′-CACTACAAGAAAACA-3′). The genome of *BrCACTA6* displays the typical features of plant CACTA elements (TAD-TNPD-DUF4218-DUF4216-TNPA) (Table 5.1 & 5.3). *BrCACTA7* is 8288 bp large including 3 bp TSDs at both ends and 15 bp TIRs similar to *BrCACTA6*. *BrCACTA11* and *BrCACTA16* are 7829 and 5442 bp with perfect protein domain organization (TAD-TNPD-DUF4218-DUF4216-TNPA) in the same frame. A 4952 bp element designated *BrCACTA16* was identified from *Brassica rapa* accession 'AC189360.2'. The element includes 3 bp TSDs and 15 bp TIRs and an internal region coding transposase proteins and associated domains. The smallest autonomous *Brassica* CACTA is 3029 bp with 3 bp TSDs, perfect 15 bp TIRs and internal region encoding transposase and its associated domain (TAD-TNPD). The average sizes of *Brassica rapa* CACTA elements range from 7-8 kb (Table 5.1 & 5.2).

5.2.4 Identification of CACTA in *Brassica* allotetraploids

Among *Brassica* allotetraploids (AABB, AACC, and BBCC) two complete CACTA and several transposase like sequences were retrieved from *Brassica napus* BAC clones available in GenBank. The first complete *Brassica napus* CACTA named *BnCACTA8* was identified from *Brassica napus* (AJ245479.1), which is 8164 bp including 3 bp TSDs and 15 bp TIRs (5′-CACTACAAGAAAACA-3′). The element is a perfect autonomous element with internal regions encoding transposase and all associated domains necessary for transposition and integration (5′-TAD-TNPD-DUF4218-DUF4216-TNPA-3′). Another complete CACTA named *BnCACTA27* was identified from *Brassica napus* accession 'AC236784.1' from 93542-100733 bp. The element is 7192 bp including typical 3 bp TSDs and 15 bp TIRs but its internal region encodes only TNPD with its associated domains (5′-TAD-TNPD-DUF4218-DUF4216-3′). Due to the lack or scarcity of available sequence (<5 Mbp) in GenBank database for the allotetraploid *Brassica juncea* (AABB), *Brassica napus* (AACC) and *Brassica carinata* (BBCC) genomes, we cannot build a picture of the CACTA elements from their genomes. However, the diversity and distribution of CACTA in *Brassica* allotetraploids was confirmed by PCR analysis, with *Brassica nigra*, *Brassica juncea* and *Brassica carinata* which amplified the transposase indicating the diversity of *Bot1* elements.

5.2.5 Protein domain organization in plant CACTA elements

The autonomous CACTA transposons mostly display a single transcriptional unit, which generates four to six protein domains (5′-TAD-TNPD-DUF4218-DUF4216-TNPA-3′) (Figure 5.1). TNPD and TNPA are the transposase genes required for transposition and integration of CACTA transposons. The transposase associated domain (TAD) is present towards N-terminal end of TNPD. The exact function of TAD is not known but it is the accessory component of TNPD transposase aiding it during transposition. Two domains named DUF4218 and DUF4216 are present towards the C-terminus of TNPD. The exact function of both domains is not known but their presence in almost all plant CACTA suggests their important role in mobilization of these elements. TNPA is localized next to DUF4218 and DUF4216 domains.

The domain pattern and organization of autonomous plant CACTA were investigated. Two major patterns of protein domain organizations were found. The first pattern is exhibited by majority of plant CACTA, where both transposases are present with other domains as 5′-TAD-TNPD-DUF4218-DUF4216-TNPA-3′. The second pattern of protein domain organizations is 5′-TAD^{+}-TNPD^{+}-DUF4218^{+}-DUF4216^{+}-[ATHILA-ORF1^{-}]-TNPA^{+}-3′, where signs + and – indicate plus and minus orientations. The first pattern of domain organization was studied in *Brassica rapa, Brassica oleracea, Arabidopsis thaliana, Petunia hybrid, Solanum tuberosum, Medicago truncatula, Zea mays, Oryza sativa* and few other plants (Table 5.2), while the second pattern of protein domain organization is only observed in *Brassica oleracea* where an additional ATHILA-ORF1 domain is present in opposite orientation. These two major patterns have a few other sub-patterns, where one or more domains are missing: in *BoCACTA5*, DUF4216 is missing, while in *BrCACTA10, BrCACTA17, BoCACTA23, BoCACTA24, BnCACTA27, EnSpm-13, EnSpm-10_ZM, EnSpm_OS, EnSpm1_TM, EnSpm-1_TA, EnSpm-1_HV, EnSpm-15_SB*, one transposase (TNPA) is missing (5′-TAD-TNPD-DUF4218-DUF4216-3′). Few CACTA from the grass family members (*Zea mays, Oryza sativa, Triticum monococcum, Triticum aestivum, Hordeum vulgare, Sorghum bicolor*) lack transposase (TNPD) is their structures, but still are active and mobile. The simplest type of the protein domain organization is observed in *Daucus carota* element *TDC1* where only two transposase proteins are residing with TAD domain. Similarly, *Arabidopsis thaliana ATENSPM1* element only displays a TNPA domain in its molecular structure (Table 5.2).

Table 5.1: *Brassica* CACTA elements with BAC sequences, sizes, number of TSDs, TIRs and orientation. Nucleotide sequences of representative CACTA elements are available in Appendices (attached CD).

Element Name	BAC Accession	Species	Size	Position	TSD	TIR	Orientation
BoCACTA1	EU642504.1	*B. oleracea*	9399	20580-29972	3	15	3′-5′
BoCACTA2	EU642505.1	*B. oleracea*	10914	44789-55702	3	15	5′-3′
BoCACTA3	EU642506.1	*B. oleracea*	11068	19777-30844	3	15	5′-3′
BoCACTA4	EU642505.1	*B. oleracea*	8205	21474-29678	3	15	5′-3′
BoCACTA5	EU642505.1	*B. oleracea*	7647	78098-85744	3	15	3′-5′
BrCACTA6	AC189480.2	*B. rapa*	9393	87937-97329	3	15	3′-5′
BrCACTA7	AC232490.1	*B. rapa*	8288	61958-70245	3	15	5′-3′
BnCACTA8	AJ245479.1	*B. napus*	8164	44881-53044	3	15	3′-5′
BrCACTA9	AC172883.2	*B. rapa*	7970	114211-122180	3	15	3′-5′
BrCACTA10	AC189446.2	*B. rapa*	7861	5462-13322	3	15	5′-3′
BrCACTA11	AC189321.2	*B. rapa*	7829	92374-100202	3	15	5′-3′
BrCACTA12	AC189341.2	*B. rapa*	7802	99395-107196	3	15	3′-5′
BrCACTA13	AC189496.2	*B. rapa*	7779	56849-64627	3	15	5′-3′
BrCACTA14	AC189314.1	*B. rapa*	7669	21683-29351	3	15	5′-3′
BrCACTA15	AC189655.2	*B. rapa*	6996	39384-46379	3	15	3′-5′
BrCACTA16	AC189360.2	*B. rapa*	5442	59073-64514	3	15	5′-3′
BrCACTA17	AC229605.1	*B. rapa*	4952	83111-88062	3	15	5′-3′
BoCACTA18	AC183492.1	*B. oleracea*	10682	81000-91686	3	15	3′-5′
BoCACTA19	EU579455.1	*B. oleracea*	7265	82206-89482	6	15	3′-5′
BoCACTA20	AC183495.1	*B. oleracea*	9661	104704-114364	3	15	3′-5′
BoCACTA21	AC183495.1	*B. oleracea*	8210	159474-167683	3	15	3′-5′
BoCACTA22	AC183495.1	*B. oleracea*	7170	237844-245013	3	15	3′-5′
BoCACTA23	AC183493.1	*B. oleracea*	8072	228710-236781	3	15	3′-5′
BoCACTA24	AC183492.1	*B. oleracea*	8362	61770-70131	3	16	5′-3′
BoCACTA25	AC183492.1	*B. oleracea*	3735	183789-187523	3	15	5′-3′
BrCACTA26	AC172883.2	*B. rapa*	7970	114211-122180	3	15	3′-5′
BnCACTA27	AC236784.1	*B. napus*	7192	93542-100733	3	15	3′-5′
BoCACTA28	AC240086.1	*B. oleracea*	8741	29332-38072	3	15	3′-5′
BoCACTA29	AC240092.1	*B. oleracea*	9900	32432-42331	3	15	3′-5′
BoCACTA30	AC183496.1	*B. oleracea*	10728	171084-181811	3	15	3′-5′
BoCACTA31	AC183496.1	*B. oleracea*	7157	350861-358017	3	15	3′-5′
BoCACTA32	AC183496.1	*B. oleracea*	6075	302434-308508	3	15	3′-5′
BoCACTA33	AC183496.1	*B. oleracea*	5916	138717-144632	3	15	3′-5′
BrCACTA34	AC189565.2	*B. rapa*	5123	57417-62539	3	15	5′-3′
BrCACTA35	AC232476.1	*B. rapa*	3029	93851-96879	3	15	5′-3′
Bo-N-CACTA1	AC240092.1	*B. oleracea*	3265	48182-51446	3	15	3′-5′
Br-N-CACTA2	AC155342.2	*B. rapa*	2559	58153-60711	3	15	3′-5′
Bo-N-CACTA3	AC240087.1	*B. oleracea*	2662	89991- 92652	3	15	3′-5′
Bo-N-CACTA4	AC240080.1	*B. oleracea*	2773	66450-69222	3	15	3′-5′
Br-N-CACTA5	AC155341.2	*B. rapa*	1419	29398-30816	3	13	3′-5′
Br-N-CACTA6	AC241034.1	*B. rapa*	1288	6551-7838	3	13	3′-5′
Br-N-CACTA7	AC189489.2	*B. rapa*	1288	108610-109897	3	15	3′-5′

Table 5.2: Protein domain organizations and TIRs of *Brassica* and other plant CACTA elements. TAD: Transposase associated domain. DUF: Domain of unknown function.

Element Name	Plant Species	Size	TIR sequence (5′-3′)	Domains (5′-3′)
BoCACTA1	*Brassica oleracea*	9399	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA2	*Brassica oleracea*	10914	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-DUF4271/ATHILA*
BoCACTA3	*Brassica oleracea*	11068	CACTACAAGAAAACA	TNPD-DUF4218-DUF4216/TAD*-ATHILA*
BoCACTA4	*Brassica oleracea*	8205	CACTACAAGAAAACA	TAD-TNPD- DUF4218-DUF4216-TNPA
BoCACTA5	*Brassica oleracea*	7647	CACTACAAGAAAACA	TAD-TNPD-DUF4218-TNPA
BrCACTA6	*Brassica rapa*	9393	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA7	*Brassica rapa*	8288	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BnCACTA8	*Brassica napus*	8164	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA9	*Brassica rapa*	7970	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA10	*Brassica rapa*	7861	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216
BrCACTA11	*Brassica rapa*	7829	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA12	*Brassica rapa*	7802	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA13	*Brassica rapa*	7779	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA14	*Brassica rapa*	7669	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216
BrCACTA15	*Brassica rapa*	6996	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA16	*Brassica rapa*	5442	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA17	*Brassica rapa*	4952	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216
BoCACTA18	*Brassica oleracea*	10682	CACTACAAGAAAACA	TAD-TNPD-DUF4218/ATHILA*
BoCACTA19	*Brassica oleracea*	7265	CACTACAAGAAAACA	TAD-TNPD-/ATHILA*
BoCACTA20	*Brassica oleracea*	9661	CACTACAAGAAAACA	TAD-TNPD-/ATHILA*
BoCACTA21	*Brassica oleracea*	8210	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA22	*Brassica oleracea*	7170	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA23	*Brassica oleracea*	8072	CACTACAAAAAAACA	TAD-TNPD-DUF4218-DUF4216
BoCACTA24	*Brassica oleracea*	8362	CACTACAAGAAAcACA	TAD-TNPD-DUF4218-DUF4216
BoCACTA25	*Brassica oleracea*	3735	CACTACAAGAAAACA	TAD-TNPD
BrCACTA26	*Brassica rapa*	7970	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BnCACTA27	*Brassica napus*	7192	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216
BoCACTA28	*Brassica oleracea*	8741	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA29	*Brassica oleracea*	9900	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA30	*Brassica oleracea*	10728	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-DUF4271/ATHILA*
BoCACTA31	*Brassica oleracea*	7157	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA32	*Brassica oleracea*	6075	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BoCACTA33	*Brassica oleracea*	5916	CACTACAAGAAAACA	TAD-TNPD
BrCACTA34	*Brassica rapa*	5123	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
BrCACTA35	*Brassica rapa*	3029	CACTACAAGAAAACA	TAD-TNPD
Bo-N-CACTA1	*Brassica oleracea*	3265	CACTACAAGAAAACA	NON-AUTONOMOUS
Br-N-CACTA2	*Brassica rapa*	2559	CACTACAAGAAAACA	NON-AUTONOMOUS
Bo-N-CACTA3	*Brassica oleracea*	2662	CACTACAAGAAAACA	NON-AUTONOMOUS
Bo-N-CACTA4	*Brassica oleracea*	2773	CACTACAAGAAAACA	NON-AUTONOMOUS
Br-N-CACTA5	*Brassica rapa*	1419	CACTACAAGAAAACA	NON-AUTONOMOUS
Br-N-CACTA6	*Brassica rapa*	1288	CACTACAAGAAAAagCA	NON-AUTONOMOUS
Br-N-CACTA7	*Brassica rapa*	1288	CACTACAAGAAAACA	NON-AUTONOMOUS
BRENSPM1	*Brassica rapa*	7811	CACTACAAGAAAACA	TAD-TNPD-DUF4218-DUF4216-TNPA
Chester-1	*Arabidopsis thaliana*	8216	CACTACAAGAAATAT	TAD-TNPD-DUF4218-DUF4216-TNPA
Cac1	*Arabidopsis thaliana*	8479	CACTACAA	TAD-TNPD-DUF4218-DUF4216-TNPA
ATENSPM1	*Arabidopsis thaliana*	4548	CACTACAAGAAAACAGTCGTTTTGCGAGG	TNPA
TDC1	*Daucus carota*	5251	CACTACAAGAAAACGCG	TAD-TNPD-TNPA
PSL	*Petunia hybrida*	9932	CACTACAAAAAA	TAD-TNPD-DUF4218-DUF4216-TNPA
EnSpm-2_STu	*Solanum tuberosum*	17800	CACTACAAAAAAACCC	TAD-TNPD-DUF4218-DUF4216-TNPA
EnSpm-13	*Vitis vinifera*	12363	CACTACTACAAAA	TAD-TNPD-DUF4218-DUF4216
EnSpm_MT	*Medicago truncatula*	8153	CACTACAAGAAAAAT	TAD-TNPD-DUF4218-DUF4216-TNPA
EN1	*Zea mays*	8287	CACTACAAGAAAA	TAD-TNPD-DUF4218-DUF4216-TNPA
Dopia4_ZM	*Zea mays*	8463	CACTACAGGAAACGCCTAAATTTTCGTGGGCC	TAD-TNPD-DUF4218-DUF4216-TNPA
EnSpm-10_ZM	*Zea mays*	8351	CACTACCGGAATCCGGGCTTTGCCGAGTG	TAD-TNPD-DUF4218-DUF4216
EnSpm_OS	*Oryza sativa*	11265	CACTACTGGAGATGGGAAGGCTCCCGTGTGCAT	TAD-TNPD-DUF4218-DUF4216
Rim2-569	*Oryza sativa*	20352	CACTGGTGGAGAAACC	TAD-TNPD-DUF4218-DUF4216-TNPA
EnSpm1_TM	*Triticum monococcum*	9841	CACTACTGGAATCAGCTAGTTTGCC	TAD-TNPD-DUF4218-DUF4216
EnSpm-1_TA	*Triticum aestivum*	14742	CACTACTAGGGAAAAGCCT	TAD-TNPD-DUF4218-DUF4216
EnSpm-1_HV	*Hordeum vulgare*	11744	CACTACTGGAATCA	TAD-TNPD-DUF4218-DUF4216
EnSpm-15_SB	*Sorghum bicolor*	14652	CACTAGTAG	TAD-TNPD-DUF4218-DUF4216

5.2.6 *Brassica* CACTA captures an ATHILA ORF-1 domain

Brassica CACTA transposons have captured an ATHILA ORF-1 domain in their coding regions. ATHILA ORF-1 domain is the integral part of *Arabidopsis thaliana* Ty3/gypsy LTR retrotransposons. The *Brassica oleracea* CACTA showed homology in ~1200-1280 bp (~400-428 aa) region of ATHILA ORF-1 domain from *Arabidopsis thaliana* Gypsy retrotransposon. The ATHILA ORF-1 domain is present in *BoCACTA2, BoCACTA3, BoCACTA18, BoCACTA19, BoCACTA20* and *BoCACTA30*. All these CACTA elements are larger in sizes from other homologues indicating the presence of ATHILA ORF-1 domain insertions in their sequences. In general, ~3.1 kb insertion was detected in *Brassica* CACTA, with ~1.2 kb region homologous to ATHILA ORF-1 domain. The insertional preference of this insertion is AT rich regions and exact terminal ends of the insertions were not identifiable due to the presence of a lot of direct repeat sequences in its terminal regions (Figure 5.1; Table 5.2).

5.2.7 Characterization of non-autonomous *Brassica* CACTA

A 2559 bp CACTA was identified from *Brassica rapa* accession 'AC155342.2' from 58153-60711 bp within BAC sequence. The element generated 3 bp TSDs and 15 bp (5′-CACTGGTGGAGAAACC-3′) TIRs. The 300 bp terminal regions were used to locate its En/Spm or autonomous CACTA and *BrCACTA6* and related homologues were found as its descendents. The element is named *Br-N-CACTA2*, where *Br* specifies *Brassica rapa*, *N* indicate non-autonomous, CACTA represents transposons superfamily and 2 indicate the number of identified element. Another 3265 bp non-autonomous CACTA (*Bo-N-CACTA1*) was identified from *Brassica oleracea* 'AC240092.1' residing from 48182-51446 bp with 3 bp TSDs and having 15 bp TIRs (5′-CACTGGTGGAGAAACC-3′) similar to other *Brassica* CACTA transposons (Figure 5.2). *Bo-N-CACTA3* and *Bo-N-CACTA4* are 2662 and 2773 bp in sizes with 3 bp TSDs and 15 bp TIRs identified from *Brassica oleracea* accessions 'AC240087.1' and 'AC240080.1' respectively. The comparison of *Brassica rapa* accessions 'AC155341.2' x 'AC189489.2' led to the identification of two non-autonomous CACTA named *Br-N-CACTA5* and *Br-N-CACTA6*, which are 1419 bp and 1288 bp respectively. *Br-N-CACTA7* is similar to *Br-N-CACTA6* in size and homology but present in another *Brassica rapa accession* 'AC241034.1' (Figure 5.2; Table 5.1).

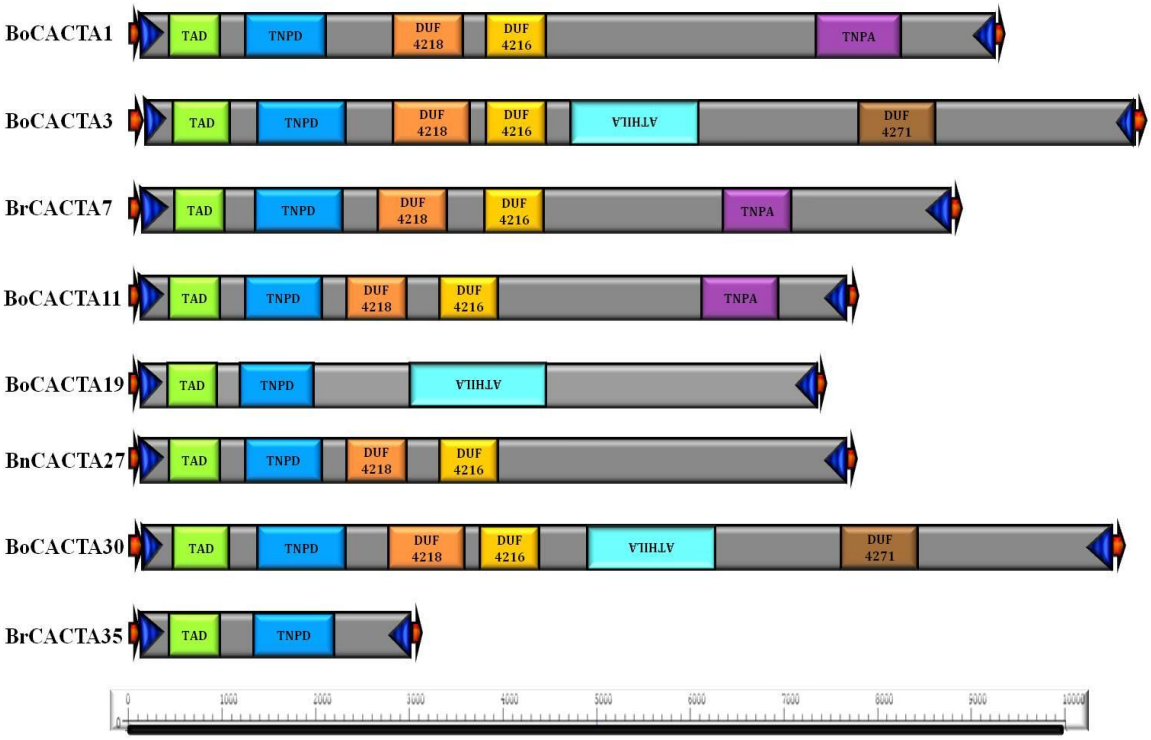


Figure 5.1: Schematic representation of CACTA elements studied in *Brassica*. Red arrows at termini represent TSDs, while blue triangles indicate TIRs. Transposases TNPD and TNPA are shown as blue and purple boxes. The transposase associated domain (TAD) is shown in green while three domains of unknown functions DUF4218, DUF4216 and DUF4271 are shown in different colours. ATHILA-ORF1 domain is shown in light blue colour in the opposite orientation. The names and sizes of domains were obtained by blasting the sequences against the known proteins in the conserved domain database of NCBI. The scale below the elements shows sizes in bp.

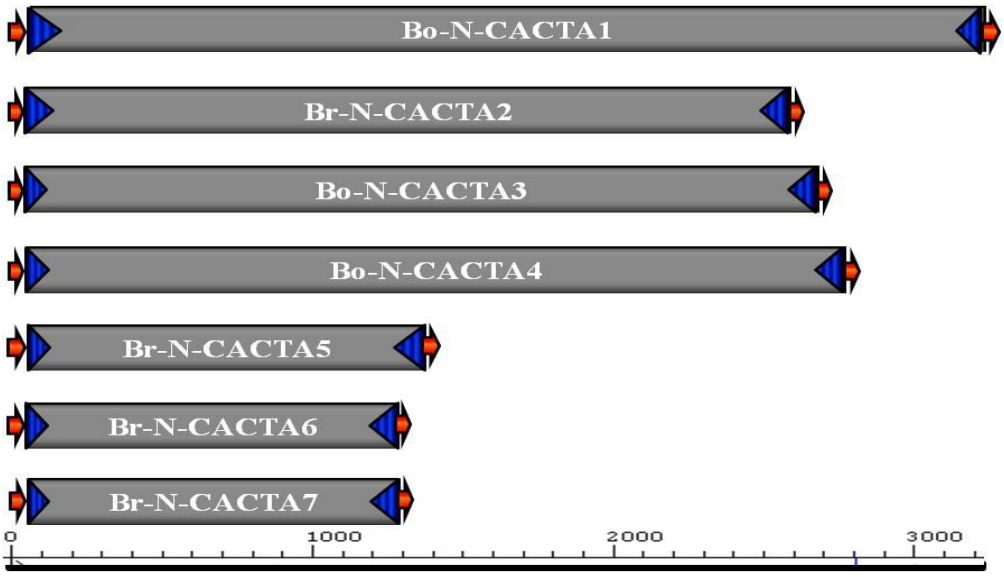


Figure 5.2: Schematic representation of non-autonomous CACTA elements studied in *Brassica*. The names given to the elements are printed in their internal regions. TSDs are shown by red arrows, while blue triangles indicate TIRs. The elements have not shown any protein coding domains like transposase or any other protein. The TIRs of the elements were similar to their ancestral autonomous CACTA elements. The scale below the elements shows sizes in bp.

5.2.8 PCR amplification of ATHILA ORF-1 insertion in *Brassica*

To investigate, whether ATHILA ORF-1 is only captured by *Brassica oleracea* CACTA or *Brassica rapa* CACTA elements also harbour this, the primers BoATHILAF (5′-ACATTGAAGGGCTGTTCCAG-3′) and BoATHILAR (3′-AGCTTGTACTGGCTGGAGTC-5′) were designed from the ATHILA ORF-1 domain. Out of 40 *Brassica* diploids and polyploids lines, a 1 kb ATHILA ORF-1 was amplified from 28 accessions indicating its presence in most of the *Brassica* genomes. A weak band of ~1 kb size was amplified from *Brassica rapa* (Pak Choy, San Yue Man, Vertus, Suttons) accessions. All the three *Brassica nigra* cultivars failed to amplify ATHILA ORF-1 domain. This revealed that *Brassica nigra* CACTA lack this domain while *Brassica rapa* and *Brassica oleracea* CACTA possess it in their genes. All the six *Brassica oleracea* cultivars amplified the 1 kb band of ATHILA ORF-1 revealing that this domain is mostly captured by *Brassica oleracea* CACTAs. Out of nine *Brassica juncea* genomes (NARC-II, Kai Choy, W3) accessions amplified the product. Strong bands were amplified from *Brassica napus* cultivars (New, Mar, Last and Best, Fortune, Drakker, Tapidor). Similarly all *Brassica carinata* accessions amplified the 1 kb band. The amplification of ATHILA ORF-1 in *Brassica napus* and *Brassica carinata* indicates that C-genome is the contributor of this domain in allotetraploids *Brassica*. All the four *Brassica* hexaploids amplified a strong band of 1 kb suggesting its diversity in these genomes (Figure 5.3b).

5.2.9 PCR amplification of *Brassica* CACTA transposase_21 (TNPD)

To amplify CACTA transposase, degenerate primers pair BoCACTA1F and (5′-CCTCAGGTGGACCATCAAAC-3′) and BoCACTA1R (3′-GACGAAAAGGTTGCAGAGGT-5′) were designed from the conserved DDD/E region of transposase (TNPD). PCR was carried out to amplify the 580 bp (~190 aa) of DDD/E domain region. A total of 40 *Brassica* genomes were used to amplify the CACTA transposase. *Brassica* CACTA transposase was successfully amplified from all the 40 diploid and polyploids *Brassica* lines suggesting its high diversity among *Brassica* species. A 580 bp strong band was amplified from all A-genome *Brassica* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons), B-genome *Brassica nigra* (HRIGRU011011, HRIGRU010978, HRIGRU010919), C-genome *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361), allotetraploid *Brassica juncea*,

Brassica napus and *Brassica carinata* accessions. The synthetic hexaploid *Brassicas* also amplified the CACTA transposase. The amplification of CACTA in all *Brassica* species indicate its diversity, presence and ancient nature and suggests their amplification before the separation of *Brassica nigra* and *Brassica rapa/Brassica oleracea* clades ~17 Mya (Figure 5.3a).

Table 5.3: List of *Brassica* CACTA and Harbinger primers with, size of the elements, size of the expected products, names and sequence of primers.

Sr.No.	TE Superfamily	TE Size	Product Size	Primers	Primer Sequence
1	CACTA	9399	580	BoCACTA1F	CCTCAGGTGGACCATCAAAC
				BoCACTA1R	GACGAAAAGGTTGCAGAGGT
2	CACTA	1200	1000	BoATHILAF	ACATTGAAGGGCTGTTCCAG
				BoATHILAR	AGCTTGTACTGGCTGGAGTC
3	Harbinger	3843	566	BoHARB1F	CGATGAGTACTTAAGAAGAC
				BoHARB1R	GGCAAGATTATGAGAGCATG
4	Harbinger	1521	2672	BrHARB5AF	CGCCATTGTTTCATGTGTGT
				BrHARB5AR	GCATTCAGATGATGTTGTGC
5	Harbinger	1392	2672	BrHARB5BF	GCACAACATCATCTGAATGC
				BrHARB5BR	GTACTACTGTCTACGTATGG
6	Harbinger	1199	1357	BoNHARB1F	ACTAGCCATTTCCATCTTCT
				BoNHARB1R	GTATTCACTTGTAGTGTTTG
7	Harbinger	819	1100	BrNHARB2F	ACATGCATAGATTGCGCTTG
				BrNHARB2R	TTTTCACATTCGGCATGAGT

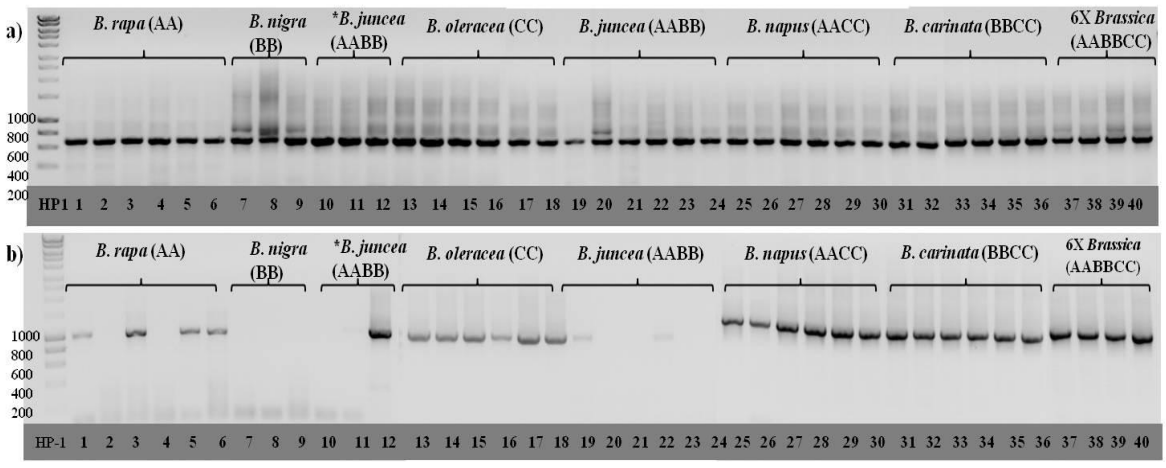


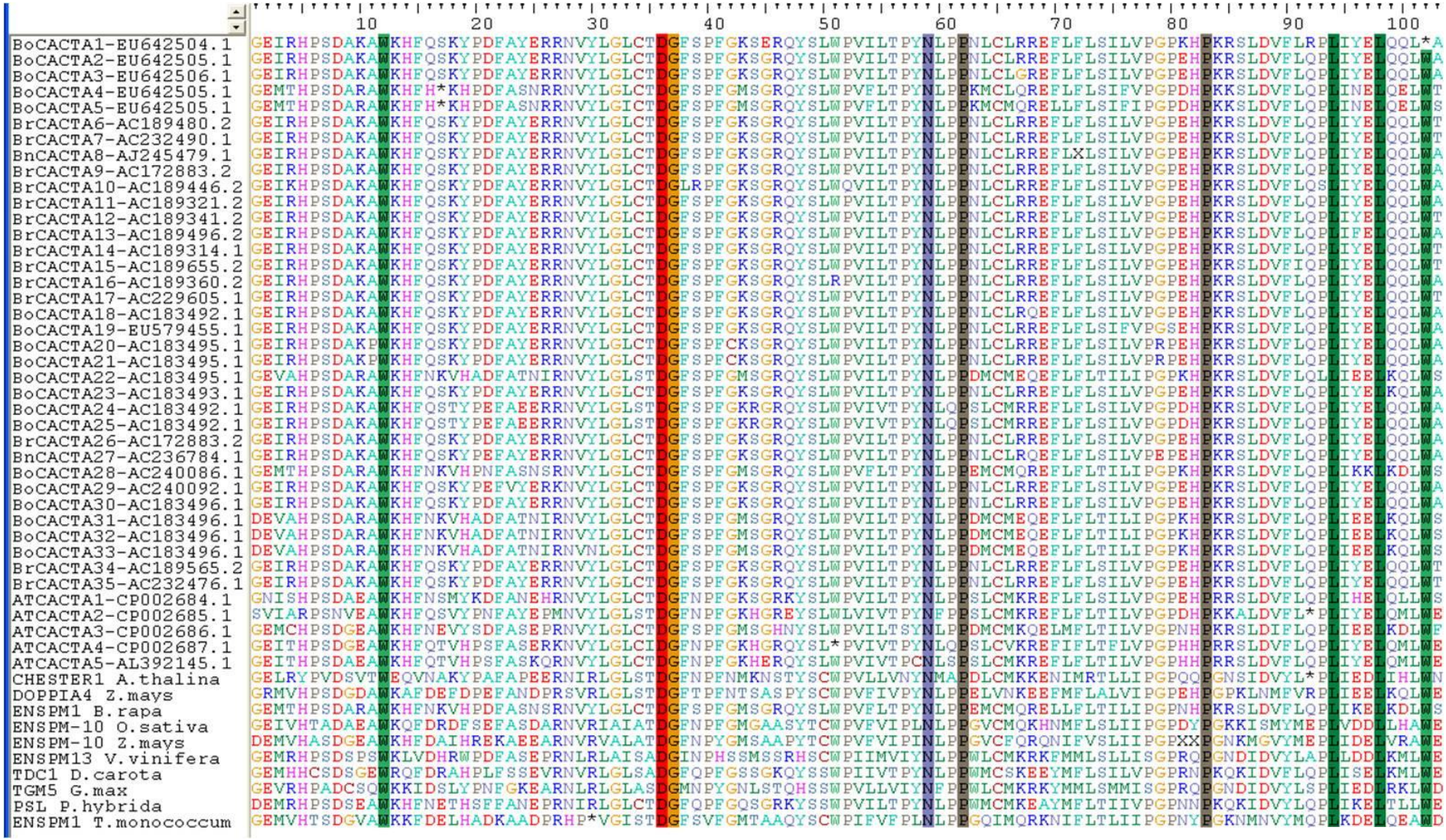
Figure 5.3: PCR amplification of a) CACTA transposase from 40 *Brassica* lines. The 580 bp bands amplify the CACTA transposase from all 40 genomes. b) BoATHILA ORF-1 domain: the 1000 bp band shows the presence of this domain in *Brassica* but in contrast to the CACTA transposase, it is not present in all accessions of *Brassica rapa* and *Brassica juncea*. All PCR figures show inverted images of size-separated ethidium bromide stained PCR products following agarose gel electrophoresis; numbers below the lanes identify each cultivar listed in table 2.1 and ladders indicate sizes in bp.

5.2.10 Phylogeny of *Brassica* CACTA transposons

The alignment of transposase (TNPD) from 35 *Brassica*, 5 *Arabidopsis thaliana* and 10 known plant transposase were performed by CLUSTALW available in BioEdit program. The *Brassica* and *Arabidopsis thaliana* CACTA transposases were collected from NCBI. The 10 transposases of well known CACTAs from various plants were collected from Repbase database (Jurka *et al.*, 2005) of eukaryotic transposable elements. The alignment of 50 transposases allows the identification of motifs essential for the transposition. The transposase sequences were mostly perfect but few are interrupted by stop codons, small indels, frameshift mutations or lacking the translation initiation codons (Figure 5.4). The highly conserved catalytic triad motif $D_{93}D_{39}D$ was present in all the transposases. In addition to the DDD triad, many other specific conserved amino acid domains are present in CACTA transposases. The amino acid residues around the DDD triad and between the second and third aspartate residue ($D_{39}D$) is the most conserved region among all plant transposase.

Phylogenetic tree was generated by using 50 transposase sequences (TNPD/Transposase_21) by the Neighbour-Joining method with 1000 bootstrap repititions and the genetic distance was calculated by Jukes-Cantor model (Figure 5.5). The tree was rooted with the grass family CACTA *Dopia4_ZM* from *Zea mays*. Phylogenetic analysis using the 210 amino acid sequences from 50 plant CACTA transposases generates two major clades. One clade represents the CACTA from other monocot and dicot plants and the other clade clustered all CACTA from the *Brassicaceae* family (*Brassica* and *Arabidopsis*) except *Chester-1* of *Arabidopsis thaliana*. The first major clade represented by 8 CACTA transposases from variable plants further splits into three sub-clades. The first sub-clade (ENSPM) is represented by the grass family members as *EnSpm10_TM* from *Triticum monococcum*, *EnSpm10_OS* from *Oryza sativa* and *EnSpm10_ZM* from Zea mays. In the second sub-clades (CHESTER1), *Chester-1* from *Arabidopsis thaliana*, *EnSpm-13* from *Vitis vinifera* and *TGM5* from *Glycine max* clustered together, while *TDC1* from *Daucus carota* and *PSL* from *Petunia hybrida* clustered together by making a third sub-clade (TDC1-DC). This suggests that in spite of high homology in monocot and dicot plant CACTA transposase, there is a divergence in both groups and they follow a different evolutionary pattern.

The second clade is represented by 41 CACTA transposases from *Arabidopsis* and *Brassica* suggesting their monophyletic origin from a common ancestor before the separation of two genuses. This major clade further resolved into 3 groups, each representing species-specific sub-clades. Thus all the four *Arabidopsis thaliana* CACTA elements (*ATCACTA1, ATCACTA2, ATCACTA4, ATCACTA5*) make a sister group with *Brassica oleracea* and *Brassica rapa* whereas *ATCACTA3* appears as an out group due to slight mutations in C-terminal end of its transposase. Although the BLASTN searches retrieved many transposase sequences from *Arabidopsis thaliana*, only 1 transposase from each chromosome was collected and analyzed in this study. Thus *ATCACTA1* represents a CACTA located on chromosome 1 and *ATCACTA*5 on chromosome 5 of *Arabidopsis thaliana*. The 19 *Brassica oleracea* with one *Brassica rapa* (*ENSPM1*) and one *Brassica napus* CACTAs clustered in three dispersed groups, in which *Arabidopsis thaliana* (*ATCACTA*) and *Brassica rapa* (*BrCACTA*) specific groups are intervened between C-genome specific CACTA (*BoCACTA*). This suggests that *Brassica* CACTA *Bot1* family transposases are not only conserved in diploid *Brassicas* but actively proliferating in allotetraploid *Brassicas* (*B. juncea, B. napus, B. carinata*). The high homology in the transposase of *Brassica* and *Arabidopsis* CACTA elements suggests their common ancestry and presence before the separation of two genera ~20 Mya (Figure 5.4 & 5.5).


10
20
30
40
50
60
70
80
90
100
BoCACTA1-EU642504.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSERQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPKHPKRSLDVFLRPLIYELQQL*A
BoCACTA2-EU642505.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA3-EU642506.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLGREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA4-EU642505.1 GEMTHPSDARAWKHFH*KHPDFASNRRNVYLGLCTDGFSPFGMSGRQYSLWPVFLTPYNLPPKMCLQREFLFLSIFVPGPDHPKKSLDVFLQPLINELQELWT
BoCACTA5-EU642505.1 GEMTHPSDARAWKHFH*KHPDFASNRRNVYLGICTDGFSPFGMSGRQYSLWPVFLTPYNLPPKMCMQRELLFLSIFIPGPDHPKKSLDVFLQPLINELQELWS
BrCACTA6-AC189480.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
BrCACTA7-AC232490.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
BnCACTA8-AJ245479.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLXLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BrCACTA9-AC172883.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BrCACTA10-AC189446.2 GEIKHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGLRPFGKSGRQYSLWQVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQSLIYELQQLWA
BrCACTA11-AC189321.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BrCACTA12-AC189341.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCIDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
BrCACTA13-AC189496.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIFELQQLWA
BrCACTA14-AC189314.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
BrCACTA15-AC189655.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFIQPLIYELQQLWT
BrCACTA16-AC189360.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLRPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BrCACTA17-AC229605.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
BoCACTA18-AC183492.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRQEFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA19-EU579455.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSIFVPGSEHPKRSLDVFLQPLIYELQQLWA
BoCACTA20-AC183495.1 GEIRHPSDAKPWKHFQSKYPDFAYERRNVYLGLCTDGFSPFCKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPRPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA21-AC183495.1 GEIRHPSDAKPWKHFQSKYPDFAYERRNVYLGLCTDGFSPFCKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPRPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA22-AC183495.1 GEVAHPSDARAWKHFNKVHADFATNIRNVYLGLSTDGFSPFGMSGRQYSLWPVILTPYNLPPDMCMEQEFLFLTILIPGPKHPKRSLDVFLQLLIEELKQLWS
BoCACTA23-AC183493.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELKQLWA
BoCACTA24-AC183492.1 GEIRHPSDAKAWKHFQSTYPEFAEERRNVYLGLSTDGFSPFGKRGRQYSLWPVIVTPYNLQPSLCMRREFLFLSILVPGPDHPKRSLDVFLQPLIYELQQLWA
BoCACTA25-AC183492.1 GEIRHPSDAKAWKHFQSTYPEFAEERRNVYLGLSTDGFSPFGKRGRQYSLWPVIVTPYNLQPSLCMRREFLFLSILVPGPDHPKRSLDVFLQPLIYELQQLWA
BrCACTA26-AC172883.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BnCACTA27-AC236784.1 GEIRHPSDAKAWKHFQSKYPEFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRQEFLFLSILVPEPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA28-AC240086.1 GEMTHPSDARAWKHFNKVHPNFASNSRNVYLGLCTDGFSPFGMSGRQYSLWPVFLTPYNLPPEMCMQREFLFLTILIPGPKHPKRSLDVFLQPLIKKIKDLWS
BoCACTA29-AC240092.1 GEIRHPSDAKAWKHFQSKYPEFVYERKNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA30-AC183496.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWA
BoCACTA31-AC183496.1 DEVAHPSDARAWKHFNKVHADFATNIRNVYLGLCTDGFSPFGMSGRQYSLWPVILTPYNLPPDMCMEQEFLFLTILIPGPKHPKRSLDVFLQPLIEELKQLWS
BoCACTA32-AC183496.1 DEVAHPSDARAWKHFNKVHADFATNIRNVYLGLCTDGFSPFGMSGRQYSLWPVILTPYNLPPDMCMEQEFLFLTILIPGPKHPKRSLDVFLQPLIEELKQLWS
BoCACTA33-AC183496.1 DEVAHPSDARAWKHFNKVHADFATNIRNVNLGLCTDGFSPFGMSGRQYSLWPVILTPYNLPPDMCMEQEFLFLTILIPGPKHPKRSLDVFLQPLIEELKQLWS
BrCACTA34-AC189565.2 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
BrCACTA35-AC232476.1 GEIRHPSDAKAWKHFQSKYPDFAYERRNVYLGLCTDGFSPFGKSGRQYSLWPVILTPYNLPPNLCLRREFLFLSILVPGPEHPKRSLDVFLQPLIYELQQLWT
ATCACTA1-CP002684.1 GNISHPSDAEAWKHFNSMYKDFANEHRNVYLGLCTDGFNPFGKSGRKYSLWPVILTPYNLPPSLCMKREFLFLSILVPGPEHPRRSLDVFLQPLIHELQLLWS
ATCACTA2-CP002685.1 SVIARPSNVEAWKHFQSVYPNFAYEPMNVYLGLSTDGFNPFGKHGREYSLWLVIVTPYNFPPSLCMKREFLFLTILVPGPDHPKKALDVFL*PLIYELQMLWE
ATCACTA3-CP002686.1 GEMCHPSDGEAWKHFNEVYSDFASEPRNVYLGLCTDGFSPFGMSGHNYSLWPVILTSYNLPPDMCMKQELMFLTILVPGPNHPKRSLDIFLQPLIEELKDLWF
ATCACTA4-CP002687.1 GEITHPSDGEAWKHFQTVHPSFASERKNVYLGLCIDGFNPFGKHGRQYSL*PVIVTPYNLQPSLCVKREFIFLTFLVPGPHHPRRSLDVFLQPLIYELQMLWE
ATCACTA5-AL392145.1 GEITHPSDAEAWKHFQTVHPSFASKQRNVYLGLCTDGFNPFGKHERQYSLWPVIVTPCNLSPSLCMKREFLFLTILVPGPHHPRRSLDVFLQPLIYELQMLWE
CHESTER1 A.thalina GELRYPVDSVTWEQVNAKYPAFAPEERNIRLGLSTDGFNPFNMKNSTYSCWPVLLVNYNMAPDLCMKKENIMRTLLIPGPQQPGNSIDVYL*PLIEDIIHLWN
DOPPIA4 Z.mays GRMVHPSDGDAWKAFDEFDPEFANDPRSVRLGLSTDGFTPFNTSASPYSCWPVFIVPYNLPPELVNKEEFMFLALVIPGPEHPGPKLNMFVRPLIEEIKQLWE
ENSPM1 B.rapa GEMTHPSDARAWKHFNKVHPDFASNSRNVYLGLCTDGFSPFGMSGRQYSLWPVFLTPYNLPPEMCMQRELLFLTILIPGPNHPKRSLDVFLQPLIKEIKDLWS
ENSPM-10 O.sativa GEIVHTADAEAWKQFDRDFSEFASDARNVRIAIATDGFNPFGMGAASYTCWPVFVIPLNLPPGVCMQKHNMFLSLIIPGPDYPGKKISMYMEPLVDDLLHAWE
ENSPM-10 Z.mays DEMVHASDGEAWKHFDAIHREKAEEARNVRVALATDGFNPYGMSAAPYTCWPVFVIPINLPPGVCFQRQNIFVSLIIPGPXXPGNKMGVYMEPLIDEIVRAWE
ENSPM13 V.vinifera GEMRHPSDSPSWKLVDHRWPDFASEPRNLRLAISADGINPHSSMSSRHSCWPIIMVIYNLPPWLCMKRKFMMLSLLISGPRQPGNDIDVYLAPLLDDLKMLWE
TDC1 D.carota GEMHHCSDSGEWRQFDRAHPLFSSEVRNVRLGLSADGFQPFGSSGKQYSSWPIIVTPYNLPPWMCSKEEYMFLSILVPGPRNPKQKIDVFLQPLISEIKMLWE
TGM5 G.max GEVRHPADCSQWKKIDSLYPNFGKEARNLRLGLASDGMNPYGNLSTQHSSWPVLLVIYNFPPWLCMKRKYMMLSMMISGPRQPGNDIDVYLSPLIEDIRKLWD
PSL P.hybrida DEMRHPSDSEAWKHFNETHSFFANEPRNIRLGLCTDGFQPFGQSGRKYSSWPVILTPYNLPPWMCMKEAYMFLTIIVPGPNNPKQKIDVYLQPLIKEITLLWE
ENSPM1 T.monococcum GEMVHTSDGVAWKKFDELHADKAADPRHP*VGISTDGFSVFGMTAAQYSCWPIFVFPLNLPPEGQIMQRKNIFLTLIIPGPNYPGKNMNVYMQPLKDELQEAWD

Continued

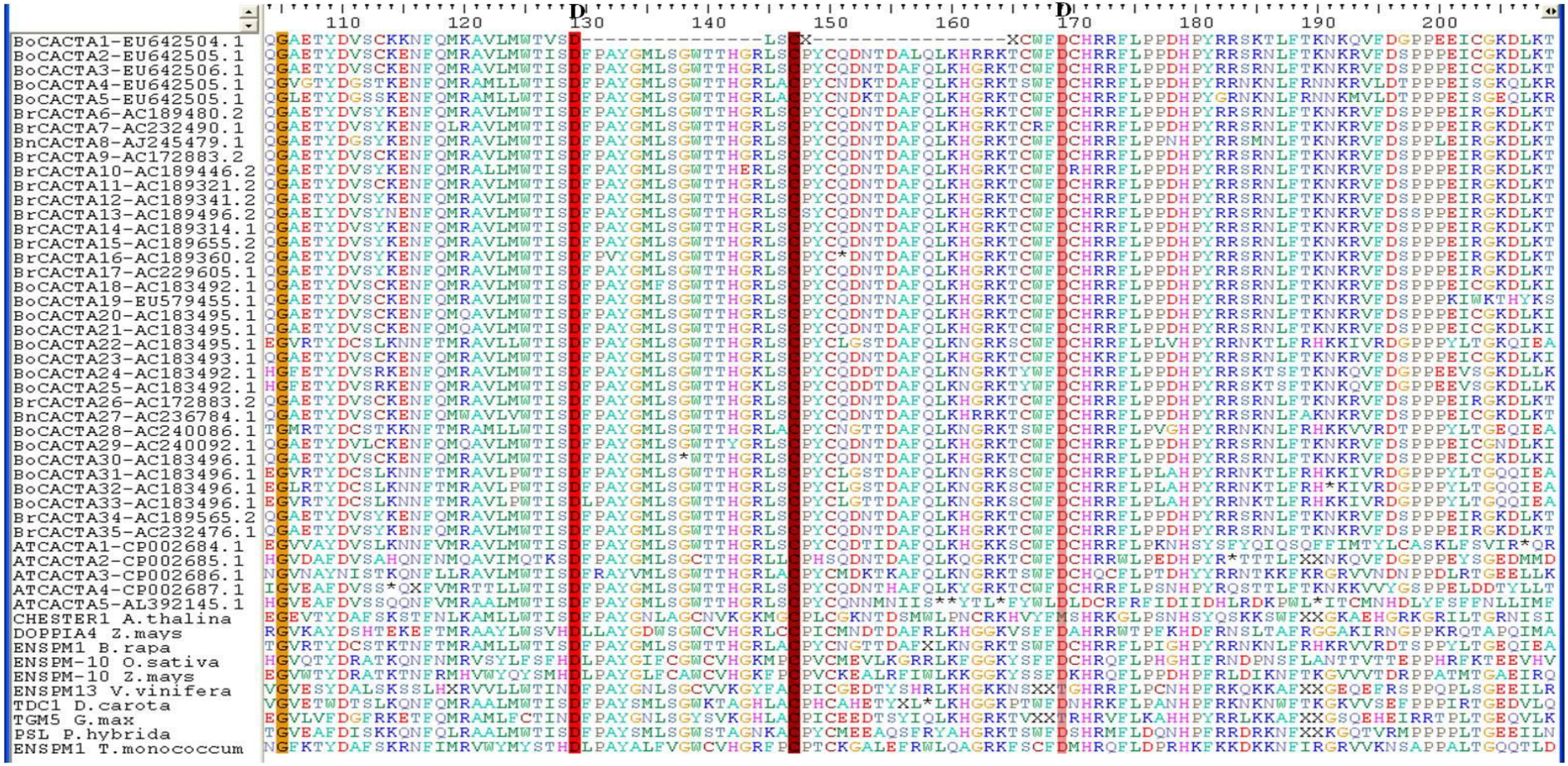


Figure 5.4: Multiple alignment of transposase encoded by *Brassica* and other plant CACTAs. The conserved DDD triad is indicated by the letters at the top. *Brassica* and *Arabidopsis* showed very high homology in their transposase, while a small difference was seen in other plant transposases. Several conserved regions are indicated by coloured vertical lines. The conserved ~210 aa region around the DDD/E triad is selected and aligned. Frameshifts were introduced if necessary with small insertions/deletions without frameshift. Asterisks in the sequences show stop codons and dashes indicate gaps.

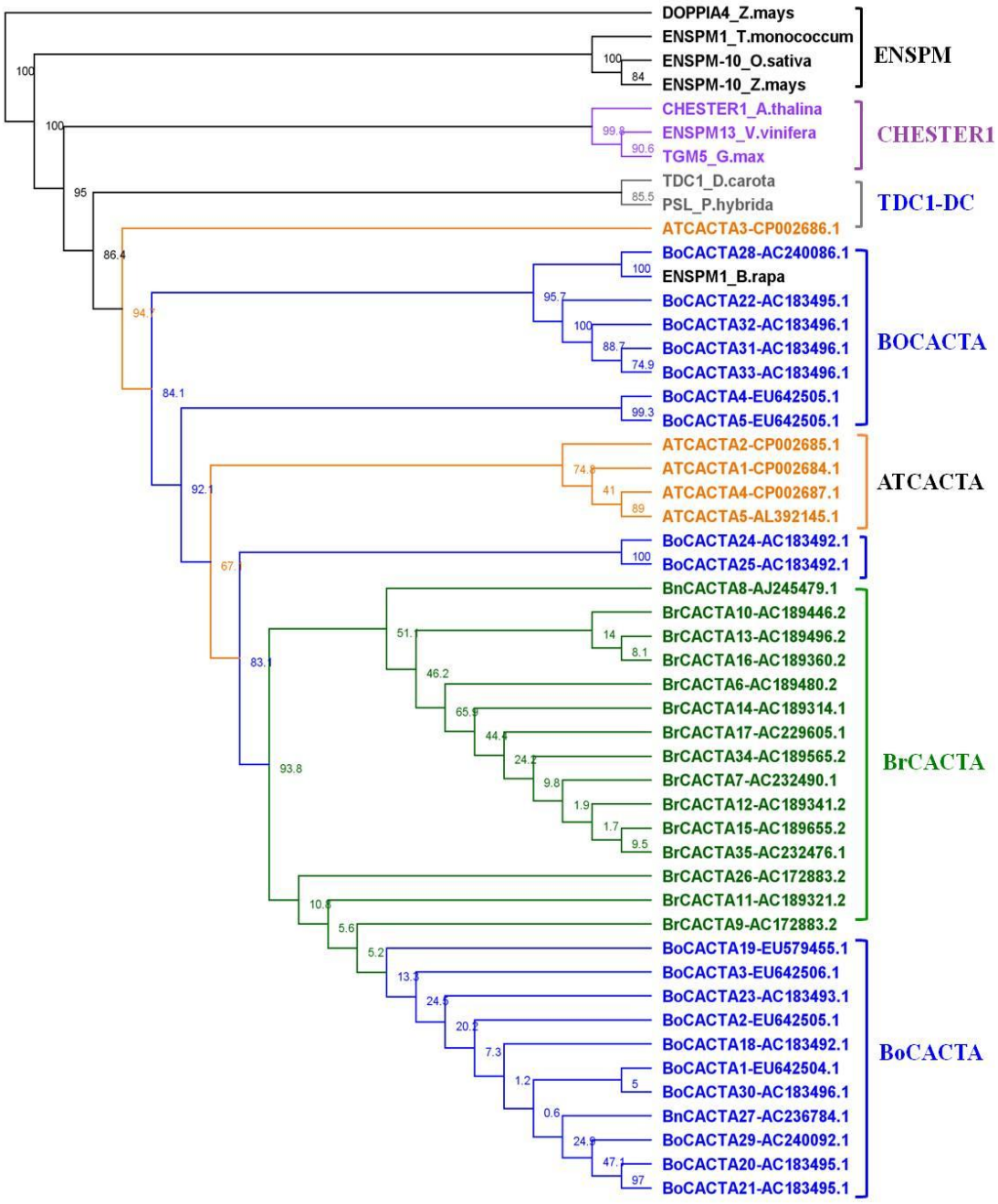


Figure 5.5: Neighbour-Joining tree showing relationship of CACTA family TNPD-transposase (Transposase-21). The phylogenetic tree of *Brassica* CACTA based on the transposase coding DNA sequence constructed by the Neighbour-Joining method with 1000 bootstrap replicates using the Geneious Pro program. The tree is rooted with *DOPIA4* transposase from *Zea mays*. The bootstrap support (%) is shown near the nodes. The known CACTA transposase sequences were obtained from Repbase database. The names of the elements are followed by the BAC numbers from which they were identified. Different sub-clades are shown in colours.

5.3 Identification of Harbinger transposons in *Brassica*

The autonomous and non-autonomous Harbinger from *Brassica* genomes were identified by dot plot comparison of homeologous BAC sequences. The dot plot comparison of *Brassica rapa* accession 'AC155344.1' against its homologue *Brassica oleracea* 'AC240081.1' led to the identification of two Harbinger *BoHARB1* and *BoHARB2* from *Brassica oleracea*. Similarly *Brassica rapa* accession 'AC155341.2' was plotted against its homologue *Brassica oleracea accession* AC240089.1, which detect a 4 kb autonomous (*BoHARB3*) and a 514 bp non-autonomous (*Bo-N-HARB3*) Harbinger in *Brassica oleracea* (AC240089.1) BAC clone. *BoHARB3* was used as query in blast searches to further detect its homologues, but only a 3.5 kb intact Harbinger from *Brassica rapa* accession 'AC189588.2' was identified. The dot plot sequence comparison of *Brassica rapa* accession (CU984545.1) against its homoeologous *Brassica oleracea* accession (EU579455) directed the identification of a 2672 bp insertion terminated by 3 bp TSDs and 17 bp TIRs. The element was named *BrHARB5* due to typical hallmarks of Harbinger transposons. Another non-autonomous Harbinger (*Br-N-HARB2*) was detected in *Brassica rapa* accession 'AC189298.1' residing within BAC from 46497-47315 bp. The autonomous Harbinger were characterized on the basis of 3 bp TSDs, 15-60 bp TIRs and internal region with a DDD/DDE transposase (TPase), whereas the non-autonomous Harbinger were characterized on the basis of TSDs, TIRs and by comparing their 5′ and 3′ terminal regions with the known Harbinger.

5.3.1 Detailed structural analysis of *Brassica* Harbinger

The first Harbinger was identified from *Brassica oleracea* (AC240081.1) from position 5984-9826 bp and was named *BoHARB1* (*Bo* indicate *Brassica oleracea*, where *HARB1* indicate 1st Harbinger from *Brassica*). The element generates a typical Harbinger-like TAA target site repeat on integration with 42 bp TIRs (Figure 5.6). The *BoHARB1* is highly AT rich (60%), with a high AT rich region (75%) in the first 350 bp immediately after the 5′ TIR. The detailed analysis of internal regions of this element revealed that it only exhibits SANT domain and lacks a transposase domain necessary for its transposition and mobilization. The SANT domain is the other protein encoded by Harbinger transposons which is considered to be a part of DNA-binding domains. On the basis of lacking the Harbinger TPase, *BoHARB1* is considered to be a defective

Harbinger. Another Harbinger-like insertion was identified from *Brassica oleracea* accession 'AC240081.1' from position 53192-56946 bp. The 3755 bp insertion exhibit the structural features of Harbinger displaying TTA TSDs and 15 bp TIRs with a mismatch of 2 bp. The element is rich in AT content (63%) with many small poly A/T sequences dispersed within the molecule. The blast hits gave no significant hits to any other element. The molecular organization of *BoHARB2* displays the encoding of two protein domains TRX and ATP11. The thioredoxin (TRX)-like protein superfamily is a highly diverse and large group of proteins containing a TRX domain with a redox active CXXC motif (FC *et al.*, 2012). The other protein (ATP11) is located at sub-terminal region of 3′ end, and is dispersed in many eukaryotic proteins.

The *BoHARB3* was identified from *Brassica oleracea* accession AC240089.1 starting from 86355 bp and ending at 90417 bp. This 4063 bp large Harbinger is terminated by AAG TSDs and 18 bp imperfect TIRs (5′-GCTTAGAGCATGATTATC-3′). The element is A/T rich (62%) with high A/T percentage (76%) in the terminal 400 nucleotides excluding TIRs at 3′ end. The molecular structure of *BoHARB3* revealed that it encodes a transposase protein in sub-terminal region of 3′ end. Besides a transposase two other proteins TRX (TRX domain containing family) and a GPCR family are encoded by it. The GPCR is a Serpentine type chemoreceptor family of proteins. This protein is located towards the C-terminal end of SANT protein domain and N-terminal end of transposase (TNP). *BrHARB4* is the only autonomous Harbinger identified from the *Brassica rapa* genome. The element is 3527 bp large in size with typical TAA TSDs and 15 bp TIRs. A ~200 bp 'CT' SSRs are present 250 bp away from the start of 5′ TIR (Figure 5.6). The element is A/T rich (60%) with several simple poly(AT) repeats. *BrHARB4* showed >75% homology in its entire length and >90% homology in transposase region indicating the members of the same family. The protein domain organization of *BrHARB4* displays a transposase domain and two other protein domains named SANT and NAM in its structure. NAM is an abbreviation of No apical meristem-associated C-terminal domain. This domain is present in several types of plant proteins including NAM-like proteins. Another Harbinger named *BrHARB5* was isolated as an insertion in *Brassica rapa* accession 'CU984545.1' from 36506-39177 bp. It is flanked by TAA TSDs and 17 bp TIRs with high A/T content (60%) in its molecule. The internal region contains a SANT and NAM associated protein superfamilies (Figure 5.6).

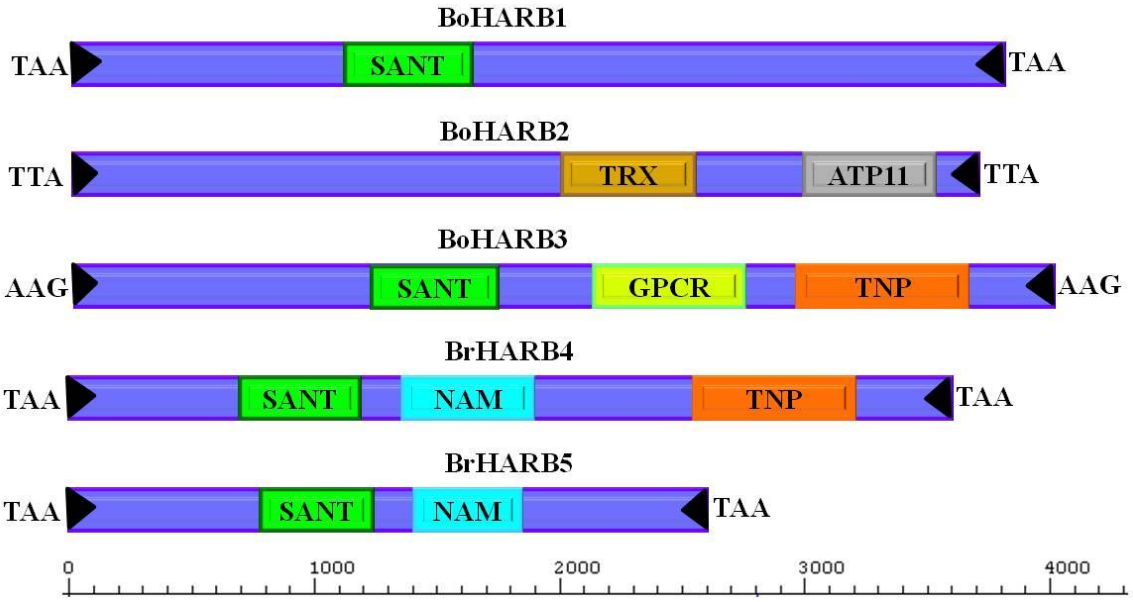


Figure 5.6: Schematic representation of *Brassica* Harbinger. The 3 bp at termini represent TSDs. Black triangles indicate TIRs. The orange box represents the transposase (TNP). SANT motifs are shown in green, NAM with blue and other domains with different colours. The protein domains were identified by screening these sequences against known proteins in the conserved domain database (CDD). The scale below shows sizes in bp. ATP11: ATP11 protein family. GPCR: Serpentine type 7TM GPCR chemoreceptor. NAM: No apical meristem-associated C-terminal domain. TRX: Thioredoxin protein superfamily.

5.3.2 Structural features of non-autonomous Harbinger in *Brassica*

In addition to the TPase-containing Harbinger, three elements were identified from *Brassica genomes* lacking any coding capacity and thus considered as non-autonomous Harbinger. They were characterized as Harbinger on the basis of 3 bp TSDs and TIRs >15 bp. The first non-autonomous Harbinger was detected from *Brassica oleracea* accession 'EU642504.1' as an insertion from 68290-69477 bp within the BAC sequence. The element is 1199 bp in size and named as *Bo-N-HARB1*. It is flanked by TTA TSDs and 24 bp perfect TIRs (5′-GAGAATCTCCAAAAGAAACTCTAT-3′). The element is highly AT rich (76%) with dispersed poly AT sequences. It captures a ~500 bp ND5 domain, which is a NADH dehydrogenase subunit. Using *Bo-N-HARB1* as a query in GenBank database, 365 sequences showed homology to the element, of which approximately half of the elements have shown >75% identity in the entire lengths. The members of this family range in sizes from 1042 bp to 1215 bp all terminated by TAA/TTA TSDs and 24-25 bp TIRs, which are highly conserved with the exception of 1-3 bp mismatches (Figure 5.7; Table 5.5). Another non-autonomous Harbinger name *Br-N-HARB2* was isolated from *Brassica rapa* accession 'AC189298.2' from 46497-47315 bp. The element is 819

bp in size terminating with TAC TSDs and 23 bp TIRs. *Bo-N-HARB3* is a 514 nucleotides large element identified from *Brassica oleracea* accession AC240089.1 inserted in position 9672-10185 bp. The element generate 26 bp imperfect TIRs (Table 5.4).

Table 5.4: Harbinger transposons studied in *Brassica* with sizes, TSDs, TIRs and positions in BAC sequences.

Name	Accession	Host	Size	TSDs	TIR (5´-3´)	Position
BoHARB1	AC240081.1	*B. oleracea*	3843	TAA	CAATAGGTCTGTTCGTTTGGTGCCC GCAGATTCCTGCGGCTG	5984-9826
BoHARB2	AC240081.1	*B. oleracea*	3755	TTA	GACCATCATTATCCC	53192-56946
BoHARB3	AC240089.1	*B. oleracea*	4063	AAG	GCTTAGAGCATGATTATC	86355-90417
BrHARB4	AC189588.2	*B. rapa*	3527	TAA	TTAATGGTTGCTTTA	34915-38440
BrHARB5	CU984545.1	*B. rapa*	2672	TAA	GAGCATCTTTATCCATG	36506-39177
Bo-N-HARB1	EU642504.1	*B. oleracea*	1199	TTA	GAGAATCTCCAAAAGAAACTCTAT	68290-69477
Br-N-HARB2	AC189298.1	*B. rapa*	819	TAC	AATATGGTGAATTGAAATAGAAT	46497-47315
Bo-N-HARB3	AC240089.1	*B. oleracea*	514	TCA	ATTGTCAATCTCTAAGACCATCGTT	9672-10185

Table 5.5: List of non-autonomous *Bo-N-HARB1* and its homologues studied in *Brassica* with sizes, TSDs and TIRs.

Name	Accession	Species	Size	TSDs	TIR (5´-3´)
Bo-N-HARB1-1	EU642504.1	*B. oleracea*	1199	TTA	*GAGAATCTCCAAAAGAAACTCTAT*
Bo-N-HARB1-2	AC183494.1	*B. oleracea*	1095	TTA	*TAGCATCTCCAAAAGACACTCTAT*
Bo-N-HARB1-3	AC183493.1	*B. oleracea*	1096	TTA	*GAGCATCTCCAAAAGACACTCTAT*
Br-N-HARB1-4	AC189475.2	*B. rapa*	1212	TAA	*GAGCATCTCCAAAAGAAACTCTAT*
Br-N-HARB1-5	AC189364.2	*B. rapa*	1212	TCA	*GAGCATTTCCAAAAGAAACTCTAT*
Br-N-HARB1-6	AC189237.1	*B. rapa*	1215	TTA	*GAGCATCTCCAAAAGAAACTCTAT*
Br-N-HARB1-7	AC189430.2	*B. rapa*	1213	TAA	*GAGCATCTCCAAAAGAAACTCTAT*
Br-N-HARB1-8	AC232512.1	*B. rapa*	1136	TTA	*CAGCATCTCCAAAAGAAACTCTAT*
Br-N-HARB1-9	AC189375.2	*B. rapa*	1102	TTA	*GAGCATCTCCAAAAAATATTCTAT*
Br-N-HARB1-10	AC189300.2	*B. rapa*	1086	TAA	*GAGCATCTCCAAAAGACACTCTAT*
Br-N-HARB1-11	AC189225.2	*B. rapa*	1063	TAA	*GAGTATCTCCAAAAGACACTCTAT*
Br-N-HARB1-12	AC232514.1	*B. rapa*	1208	TAA	*GAGCATCTCCAAAAGAAACTCTAT*
Br-N-HARB1-13	AC189183.2	*B. rapa*	1117	TAA	*GAGCATCTCCAAAAGAAACTCTAT*
Br-N-HARB1-14	AC189592.2	*B. rapa*	1042	TAA	*GAACATCTCCAAAAGAAACTTTAT*
Bn-N-HARB1-15	AC236787.1	*B. napus*	1095	TTA	*TAGCATCTCCAAAAGACACTCTAT*

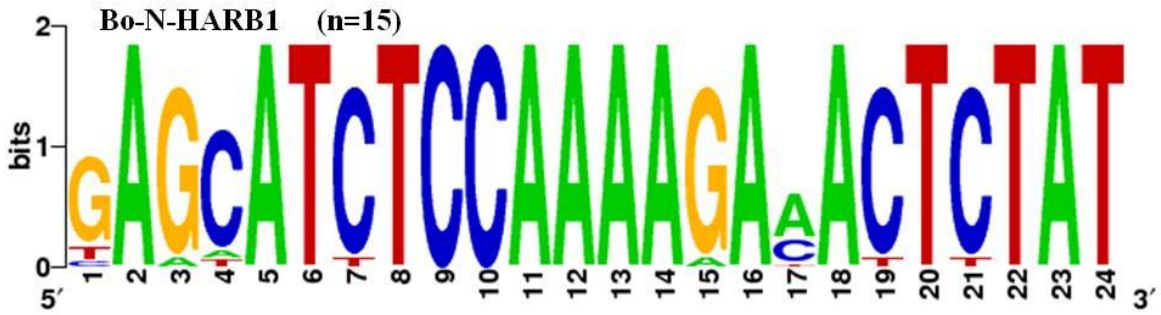


Figure 5.7: Pictrogram representing TIRs of non-autonomous Harbinger elements. Fifteen TIRs were used to generate the logo. Nucleotides 1, 3, 4 and 17 are most variable, while others particularly 8 to 14, are highly conserved among various elements.

5.3.3 Insertional polymorphism of non-autonomous Harbinger in *Brassica*

The Insertional polymorphisms of *Brassica* non-autonomous Harbinger were performed by using TIP markers designed from flanking regions of insertions. The higher and lower bands were achieved on the basis of presence or absence of insertions at specific loci. The BoNHARB1F (5′-ACTAGCCATTTCCATCTTCT-3′) and BoNHARB1R (3′-GTATTCACTTGTAGTGTTTG-5′) primers pair was used to amplify 1199 bp *Bo-N-HARB1* element with a product size of ~1357 bp including the flanking regions. The amplification of *Bo-N-HARB1* was not observed in any of A-genome, but B and C-genome *Brassica* diploids yielded the expected bands. All the three *Brassica nigra* (HRIGRU011011, HRIGRU010978, HRIGRU010919) and six *Brassica oleracea* accessions (De Rosny, Kai Lan, Early Snow Ball, Precoce Di Calabria, Cuor Di Bue Grosso, GK97361) amplified the ~1357 bp segments. Similarly, four *Brassica napus* (Mar, Last And Best, Fortune, Drakker) and six *Brassica carinata* accessions (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) amplified the *Br-N-HARB1* elements. One hexaploid *Brassica* (*B.carinata* x *B.rapa*) amplified the expected band (Figure 5.8a). Another primer pair BoNHARB2F (5′-ACATGCATAGATTGCGCTTG-3′) and BoNHARB2R (3′-TTTTCACATTCGGCATGAGT-5′) was designed to amplify a 819 bp *Bo-N-HARB2* element with a product size of 1100 bp including ~180 bp flanking regions. The primer amplified the desired bands from *Brassica rapa* (Pak Choy, Chinese Wong Bok) and *Brassica juncea* (NARC-I, NATCO, W3, Varuna) genomes. Additional bands of ~500 bp were amplified from all genomes except one *Brassica nigra* and four *Brassica juncea* lines (Figure 5.8b).

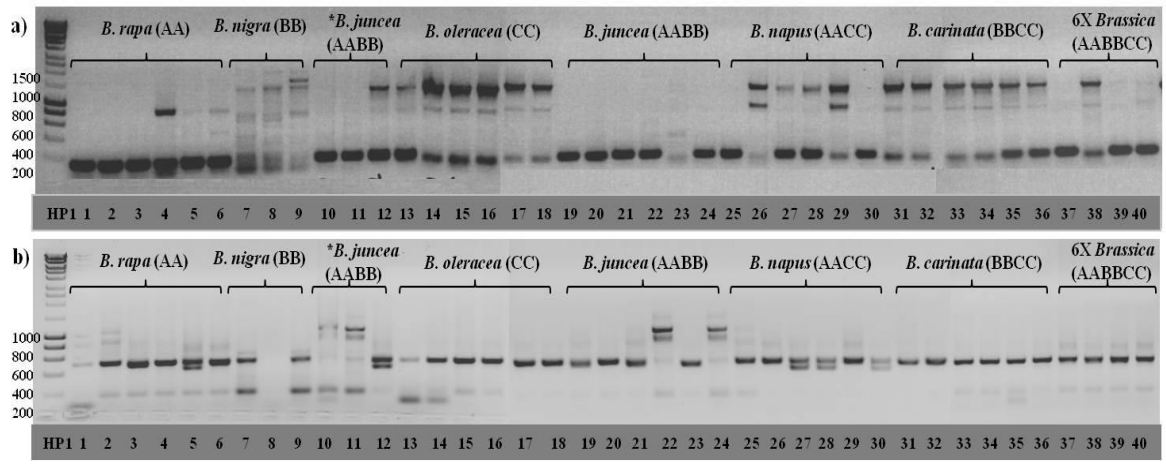


Figure 5.8: Insertional polymorphism of non-autonomous Harbingers. a) Upper bands (1357 bp) amplifying 1199 bp *Bo-N-HARB1* b) upper bands (1100 bp) amplifying 819 bp *Br-N-HARB2* from various *Brassica* lines.

5.3.4 PCR amplification of Harbinger transposase in *Brassica*

The diversity and amplification pattern of Harbinger specific transposase was performed using 40 *Brassica* cultivars. The oligonuclotide primers were designed from the transposase region around the DDD/E motif. The blast analysis showed a high diversity of Harbinger transposase in *Brassica* genomes. This was further confirmed by PCR amplification of transposase from various *Brassica* species. Out of 40 *Brassica* accessions tested for the presence of Harbinger transposase, a 566 bp TPase region was amplified from 38 diploids and polyploids *Brassica*. The only genomes failed to amplify the TPase were *Brassica rapa chinensis* (Pak Choy) and *Brassica rapa rapa* (Vertus). Very weak band was observed in *Brassica rapa pekinensis,* where PCR was repeated to gain a strong band at different annealing temperature. The lack of amplification in *Brassica rapa* accessions (Pak Choy and Vertus) might be due the difference in annealing temperatures or there might be a possibility that the primer specific region of TPase is either variable, defective or absent in these genomes. The expected product size of 566 bp transposase was amplified from the three *Brassica nigra* accessions (HRIGRU011011, HRIGRU010978, HRIGRU010919) suggesting its presence in B-genome *Brassica*. In addition to the amplification of expected band, additional bands of ~380 bp were also amplified (Figure 5.9a). Besides amplifying 566 bp products, additional bands of ~550 bp were amplified from all six *Brassica oleracea* accessions. The 350 bp band was also amplified by all *Brassica oleracea* lines. All nine *Brassica juncea* accessions (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna),

Brassica carinata (AACC) accessions (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and synthetic hexaploids also amplified the Harbinger TPase (Figure 5.9a). The amplification of Harbinger TPase shows that Harbinger transposons are ancient superfamily of DNA transposons and were present in *Brassica nigra*, *Brassica rapa* and *Brassica oleracea* before their divergence from a common ancestor.

5.3.5 Insertional polymorphism of *BrHARB5* in *Brassica* genomes

BrHARB5 was identified from *Brassica rapa* accession 'CU984545'. The sequence was used as a query in blast searches against the GenBank database to collect the other copies of the element. No significant hits were received from any of *Brassica* genomes. The question arise: is *BrHARB5* unique to *Brassica rapa* or are other homologues dispersed in various *Brassica* species? To answer these questions, the markers were developed: one from 5′ end and insertion and the other from the 3′ end and insertion amplifying the 1516 bp first half (including 192 bp flanking region) and 1521 bp last half (including 153 bp flanking region) respectively (Figure 5.9d). Both first and last parts of *BrHARB5* were amplified in all the six *Brassica rapa* accessions (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons). The bands were very strong and amplified the first half and last half indicating the presence of complete *BrHARB5* in all *Brassica rapa* genomes. In contrast, no amplification was observed in any of *Brassica nigra* (BB genome) and *Brassica oleracea* (CC genome) accessions. This confirmed the A-genome specificity of *BrHARB5*. The amplification pattern of *BrHARB5* in allotetraploids and hexaploids further strengthens the A-genome specific nature of this element, where only *Brassica juncea* (AABB), *Brassica napus* (AACC) and hexaploid *Brassica* (AABBCC) amplified the products, while *Brassica carinata* (BBCC) failed to amplify this. Out of 9 *Brassica juncea* used, only *Brassica juncea* (Kai Choy and Tsai Sim) amplified the 1516 and 1521 bp first and last part of *BrHARB5*. All the five *Brassica napus* accessions except 'Last and Best' amplified both first and last part of *BrHARB5*. Similarly 2 hexaploids (AABBCC: *B. napus* x *B. nigra*) amplified the complete *BrHARB5* (Figure 5.9b-d).

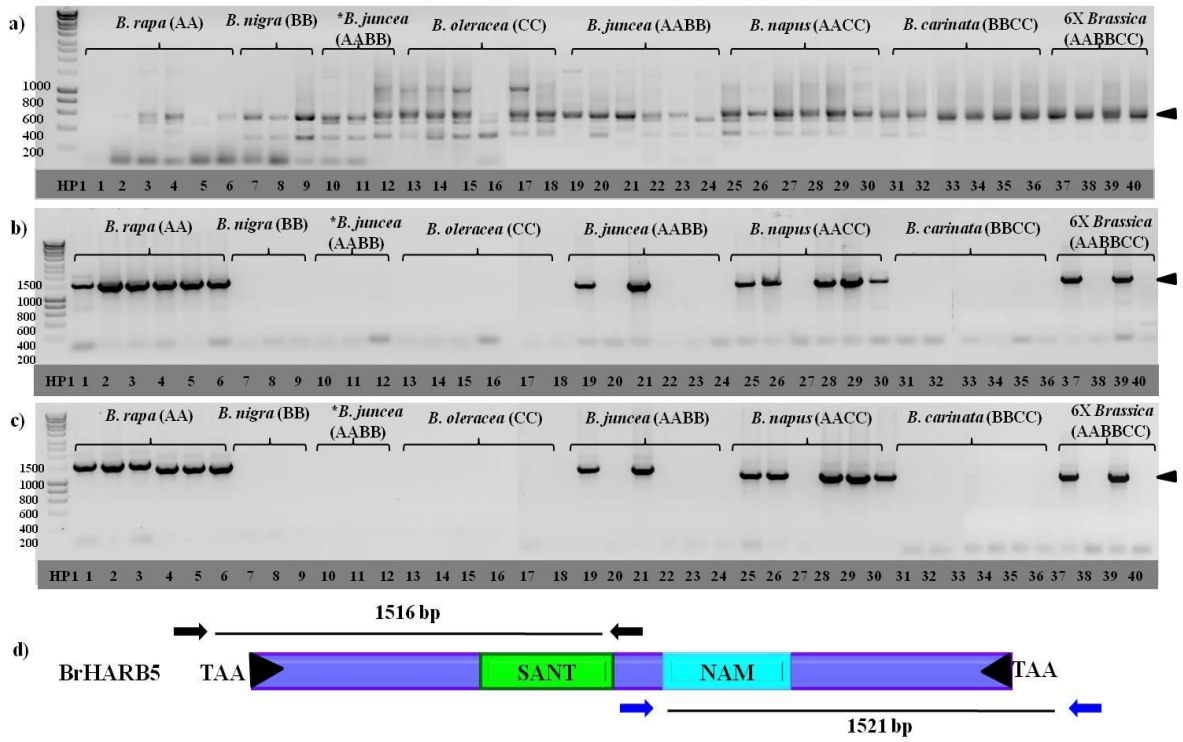


Figure 5.9: PCR amplification of a) 566 bp *Brassica Harbingers* transposase. The transposase is present in most of *Brassica* genomes except accessions 1 and 5. b) First part of *BrHARB5* (1566 bp) c) Last part of *BrHARB5* amplified from A-genome and its allotetraploids (AABB, AACC) and hexaploids (AABBCC). d) Showing the position of markers (primers) with product sizes from *BrHARB5*.

5.3.6 The phylogenetic relationship of *Brassica* and other plant Harbinger

The transposase domains from 22 Harbingers around DDE domain (~200 aa) region were aligned in CLUSTALW. The alignment revealed that the transposase from various plants showed high homology and several conserved regions with maximum homology within *Brassicaceae* members. A highly conserved $D_{88}D_{38}E$ triad can be observed in all the transposases except one (*HARB1_OS*), where 'E' amino acid is missing due to incomplete transposase analyzed (Figure 5.10). The *Zea mays* Harbinger (HARB2_ZM) was used to root the tree. The evolutionary tree based on this alignment categorized *Brassica* and other plant Harbingers into four clades. Four elements, including the previously described elements *HARB-1_OS, HARB-1_TA, MTISI12A_MT* and *HARB-1_MD* grouped together in a clade named *HARB1_OS*. The grass family members clustered together separating from dicot plant Harbingers. The second clade is represented by 3 members: *Sorghum bicolor SolHARB-10_SB, Solanum tuberosum* element *HARB-3_ST* and *Vitis vinifera Harbinger-1*. All the five identified *Arabidopsis thaliana* Harbinger with 1 known *Arabidopsis* Harbinger (*ATIS112A*) cluster together in third clade named *ATIS112A*

family. The *AtHARB3* and *AtHARB4* come together while *AtHARB1* and *AtHARB5* make a same branch. The *Brassica* Harbingers grouped together in the same clade without any clustering of species-specific group within genus *Brassica*. The transposase of both *Brassica* Harbingers have shown high homology in their transposase in their entire lengths. *BoHARB3* and *BrHARB4* from *Brassica oleracea* and *Brassica rapa* respectively constitute the same branch indicating the similar pattern of this Harbingers evolution. Similarly *Brassica oleracea* Harbinger *BoHARB5* and *BoHARB6* comes together. *Brassica napus* Harbinger (*BnHARB10*) also grouped with *Brassica rapa* and *Brassica oleracea* Harbingers. The clustering of monocot and dicot Harbingers in separate clades and genus-specific groups within *Brassica* and *Arabidopsis* was observed suggesting their common ancestry (Figure 5.11).

Table 5.6: Size and protein domain organizations of *Brassica* and other plant Harbingers. The known Harbingers were collected from Repbase database. The letters before the HARB represent the generic and species names. The double prime (″) represent 'as above'.

No.	Element Name	Plant Species	Size	Domains (5′-3′)	Reference
1	*BoHARB1*	*Brassica oleracea*	3837	SANT	Present Study
2	*BoHARB2*	*Brassica oleracea*	3749	TRX-ATP11	″
3	*BoHARB3*	*Brassica oleracea*	4057	SANT-GPCR-TNP	″
4	*BrHARB4*	*Brassica rapa*	3521	SANT-NAM-TNP	″
5	*BrHARB5*	*Brassica rapa*	2672	SANT-NAM	″
6	*HARBINGER*	*Arabidopsis thaliana*	5382	TNP	Repbase database (Jurka *et al.*, 2005)
7	*ATIS112A*	*Arabidopsis thaliana*	5099	TNP	″
8	*HARB-3_STu*	*Solanum tuberosum*	4212	SANT-TNP	″
9	*Harbinger-1_VV*	*Vitis vinifera*	4378	SANT-TNP	″
10	*MTISI12A*	*Medicago truncatula*	3914	SANT-TNP	″
11	*HARB-1_Mad*	*Malus domestica*	2818	TNP	″
12	*HARB-2_ZM*	*Zea mays*	6231	TNP-NAM	″
13	*HARB-1_TA*	*Triticum aestivum*	2161	TNP	″
14	*HARB-1_OS*	*Oryza sativa*	5166	SANT-NAM-TNP	″
15	*HARB-10_SBi*	*Sorghum bicolor*	5934	TNP-SANT-CVV	″

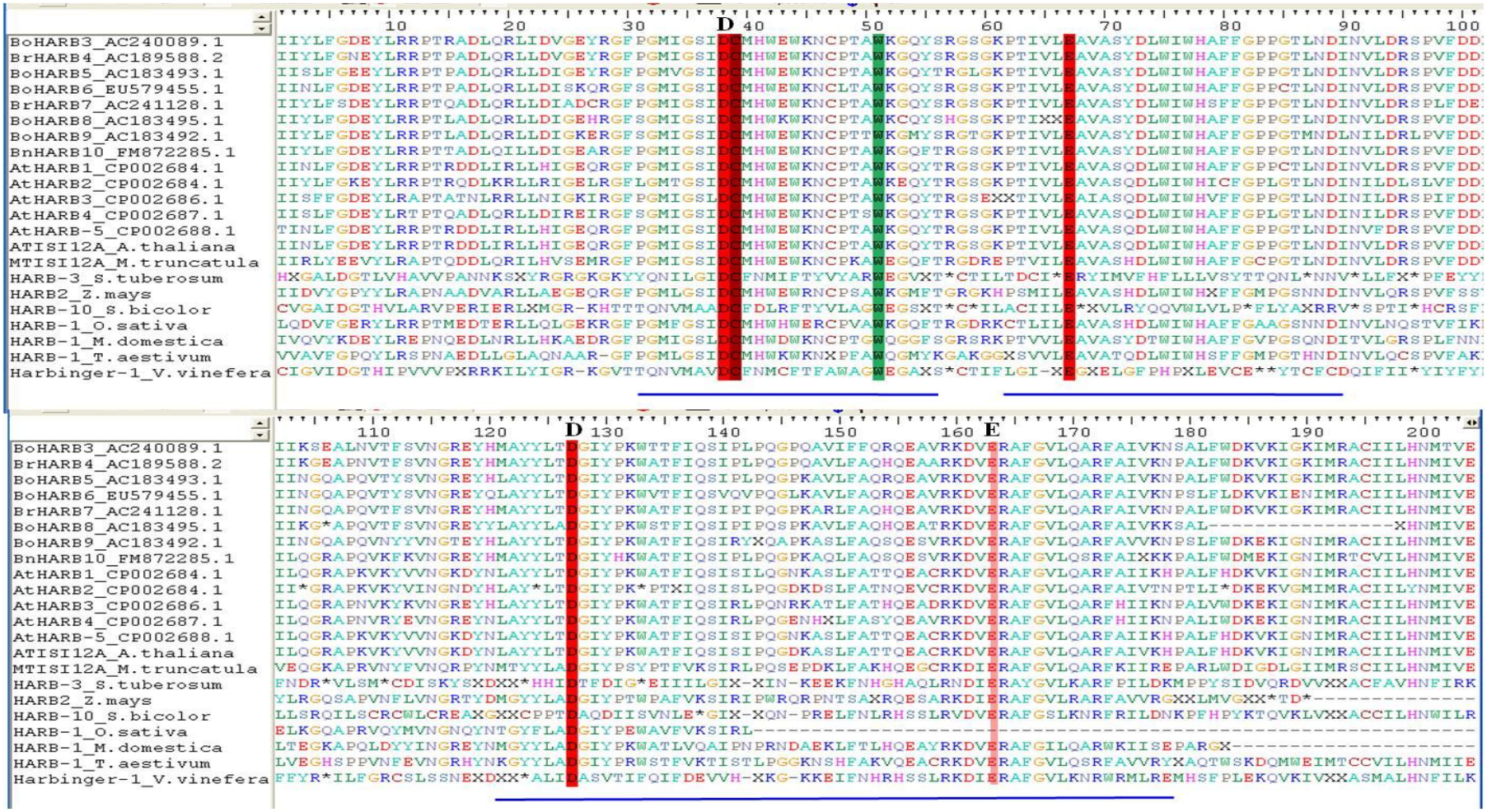


Figure 5.10: Multiple alignment of transposase encoded by *Brassica* and other plants Harbinger elements. The conserved DDE triad is indicated by the letters at the top. The most conserved regions are underlined with blue colour. The nucleotide sequences were aligned with CLUSTALW, translated to amino acids and ~205 aa region around DDE triad is represented. Small insertions were deleted without altering the frame. Asterisks show the stop codons. The letter X indicates an incomplete codon and dashes represent gaps or ends of incomplete sequences.

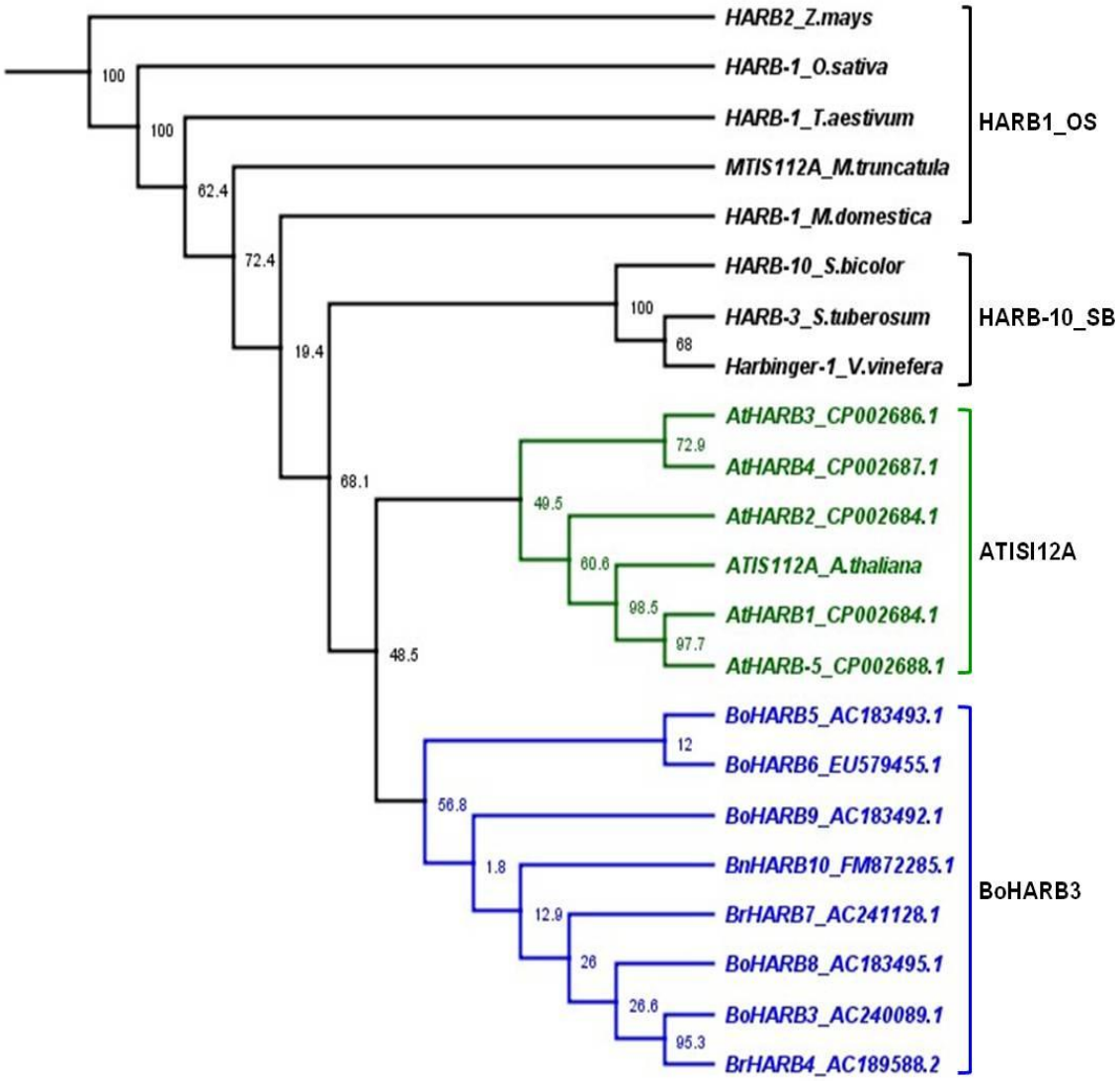


Figure 5.11: Neighbour-Joining tree of *Brassica* Harbinger and other PIF/Harbinger-like elements. Bootstrap (1000 replicates shows as %) Neighbour-Joining tree of *Brassica* Harbinger and other PIF/Harbinger-like elements based on a fragment of the deduced amino acid sequences constituting the DDE domain (~210) of the transposase. The names show BAC accessions or for non-*Brassicaceae* elements, species names; family names (right) on the basis of the well known Harbinger.

5.4 Discussion

5.4.1 CACTA transposons are actively proliferating in *Brassicaceae* genomes

In the present study, 42 CACTA elements and several analogues were detected proliferating in the *Brassica* genome (Figure 5.1; Table 5.1). It was found that the first identified element in this study has shown 100% homology to the *Bot1* element (Alix *et al.*, 2008), due to which all *Brassica* CACTA identified in present study were placed in

Bot1 family. The *Bot1* family of *Brassica* were also investigated in *Arabidopsis*, where ~110 copies of *Bot1*-like transposase were isolated suggesting their abundance and proliferation in *Arabidopsis* genomes. This predicts the common origin of *Brassica* and *Arabidopsis* transposons from a common ancestor. This was confirmed by computational based comparative analysis of *Brassica* and *Arabidopsis*, indicating that both share the same collection of TEs but in varied proportions, the number being greater in *Brassica oleracea* due to three fold larger genome than *Arabidopsis* (Zhang *et al.*, 2004). The present study indicates that the CACTA elements from A and C-genome specific *Brassica* have shown a very high homology with each other especially in TIRs (98-100%). The homology within CACTA sequences remained consistent among *Brassica* and *Arabidopsis*. This is in accordance to (Zhang *et al.*, 2004), who analyzed the evolutionary relationship of CACTA transposons in *Brassica* and *Arabidopsis* and showed high intra-family homology of *Brassica oleracea* CACTA with a close relation to *Arabidopsis*. The molecular analysis of CACTA investigated in present study revealed that it encodes two transposase (TNPD, TNPA) and 1-3 additional proteins. Similar additional proteins were observed in *Casper* family in *Triticeae* (Wicker *et al.*, 2003).

5.4.2 CACTA are diverse and an abundant superfamily of transposons

The identification of several autonomous CACTA and their non-autonomous copies revealed their abundance in *Brassica* genomes. Out of 42 *Brassica* CACTA characterized, 35 are autonomous and remaining 7 are non-autonomous. All the 35 elements encode transposase TNPD but 15 *Brassica* CACTA lack the transposase TNPA, but still are active in their mobilization. Although few of these autonomous CACTA have frameshift mutations or in-frame stop codons within their coding regions, but all those elements were considered as autonomous or intact elements which have TSDs, TIRs and a transposase. Out of 35 intact elements, 14 are from *Brassica rapa*, 19 from *Brassica oleracea* and 2 from *Brassica napus*. The mechanism of transposition of non-autonomous elements is not clear but it is likely that they utilize transposase molecules of autonomous CACTA elements.

The copy numbers of CACTA elements were estimated and it was found that *Brassica oleracea* harbours ~3085 copies, whereas ~205 copies were estimated for *Brassica rapa* whole genomes. The high copy numbers in *Brassica oleracea* suggests its successful

proliferation and distribution in C-genome. Our results based on genome-wide bioinformatic sampling is notably in accordance to Alix *et al*., (2008), where ~395-910 and 3000 copies of CACTA elements were estimated in A and C-genomes respectively by hybridization to BAC arrays. The diversity and abundance of *Bot1* family of CACTA was investigated in *Brassica oleracea* genome, where large sized (9.3-11 kb) *Bot1* elements played a vital role in *Brassica rapa* and *Brassica oleracea* genome divergence by proliferating in *Brassica oleracea* (Alix *et al.*, 2008). The new results show parallels with results in the Solanaceae, where the diversity of CACTA elements was investigated (Proels and Roitsch, 2006). A CACTA insertion found in intron I of tomato (*Solanum* (*Lycopersicon) esculentum*) extracellular invertase gene Lin5 showed high homology to the transposase of *Antirrhinum majus Tam1* element. Based on these findings, a consensus primers pair from transposase (Tpase) was used for PCR amplification and analysis of TPase-like sequences from *Solanum tuberosum, Nicotiana tabacum,* and *Datura stramonium,* showing the distribution and indicating high sequence conservation throughout the family *Solanaceae* (Proels and Roitsch, 2006). The soybean genome harbours several CACTA elements in their genomes, where nine CACTA elements designated as *Tgm1*, *Tgm2*, *Tgm3*, *Tgm4*, *Tgm5*, *Tgm6*, *Tgm7*, *Tgm-Express1*, and *Tgmt**, have been reported (Zabala and Vodkin, 2008). The monocot genome is also a hotspot for CACTA identification (Wickler *et al*., 2003). Approximately, 600-700 copies of Rim2 were reported from rice genomes, while 347 Rim2 elements from the 230 MB of rice genome were identified by data mining and cloning of the amplified genomes. The diversity and distribution of the CACTA elements incorporating the gene fragments were investigated in maize genome (Qing *et al*., 2008), with 69 elements representing 0.01% of the genome being distributed on all 10 chromosomes.

5.4.3 Terminal inverted repeats are conserved in *Brassica* and other plant CACTAs

The number and conserved pattern of TIRs specify a DNA transposon superfamily. The TIRs in *Brassica* CACTA are 15 bp and highly conserved (5′-CACTACAAGAAAACA-3′) with the exception of 1 autonomous *BoCACTA24* and a non-autonomous *Br-N-CACTA6*, where single and two nucleotide insertions were detected upstream to the 3′-termini (CA) of TIRs. All the autonomous and non-autonomous CACTA identified among *Brassica rapa*, *Brassica oleracea* and *Brassica napus* genomes display the 15 bp highly conserved TIRs (5′-CACTACAAGAAAACA-3′) among all CACTA studied in

Brassica (Figure 5.12). Similar 15 bp TIRs were generated by CACTA investigated from *Brassica rapa* and *Brassica napus* (Alix *et al.*, 2008). The TIRs of *Brassica* CACTA elements were compared with the TIRs of other plant CACTA transposons collected from Repbase database of transposable elements (Jurka *et al.*, 2005). The known CACTA element *BRENSPM1* from *Brassica rapa* also possess similar 15 bp TIRs (5′-CACTACAAGAAAACA-3′). The closest genera (*Arabidopsis thaliana*) of *Brassica* have shown more or less similar TIRs. The 8216 bp large *Chester-1* from *Arabidopsis thaliana* also displays 15 bp TSDs (5′-CACTACAAGAAATAT-3′), of which 13 bp are similar to *Brassica* CACTA TIRs. In contrast the element *CAC1* from *Arabidopsis thaliana* generates the shortest TIRs (5′-CACTACAA-3′), which are completely similar to 5′ termini of TIRs. The similarity of TIRs in *Brassica* and *Arabidopsis* suggests their common origin from the same ancestral CACTA sequence before the separation of both genera. However, the CACTA superfamily is evolutionarily much older: the TIRs generated by other dicotyledonous plants are nearly homologous with *Brassica* CACTA TIRs. A 5251 bp CACTA named *TDC1* exhibits 18 bp TIRs; *PSL* element from *Petunia hybrida*, *EnSpm-13_VV* element from *Vitis vinifera* and *EnSpm_MT* elements from *Medicago truncatula* exhibit 12, 13 and 14 bp TIRs respectively. A large CACTA element *EnSpm-2_Stu* (17800 bp) from *Solanum tuberosum* generated 16 bp TIRs, while TIRs of *Tgm1* (Xu *et al.*, 2010) from Soybean showed 30 bp TIRs with first 14 nucleotides similar to other plants (Table 5.2).

The structural features of CACTA TIRs from monocotyledonous plants revealed less homology to *Brassicaceae* members. The closest TIRs were observed in an 8287 bp *EN1* element from *Zea mays*, where 13 bp TIRs (5′-CACTACAAGAAAA-3′) are fully homologous to *Brassica* CACTA TIRs. *Zea mays Dopia4_ZM* displays 32 bp while *EnSpm-10_ZM* displays 29 bp TIRs. The largest TIRs (33 bp) were observed in *Oryza sativa* CACTA *EnSpm_OS*. A large CACTA element *RIM2-569* (20352 bp) exhibits 16 bp TIRs (5′-CACTGGTGGAGAAACC-3′), similar to *Brassica* CACTA TIRs. The TIRs of *EnSpm1_TM* (*Triticum monococcum*) and *EnSpm-1_TA* (*Triticum aestivum*) are 25 and 19 bp respectively. *EnSpm-1_HV* elements from *Hordeum vulgare* and *EnSpm-15_SB* from *Sorghum bicolour* displays 14 and 9 bp TIRs. The overall review of plant CACTA revealed that first 5-13 bp of TIRs are highly conserved in almost all plants CACTA (Figure 5.12); the 'CACTA' signature is the most conserved motif present in all the CACTA elements identified from plants (Table 5.2).

Figure 5.12: Pictrogram showing the conserved TIRs based on 42 *Brassica* CACTA. The CACTA motif is highly conserved and observed in all 100% of TIRs. The height of nucleotides (0 to 2) is proportional to their conservation.

5.4.4 Harbinger transposons are less active in *Brassica* genomes

In present study, 5 Harbinger transposons with or without active transposase were identified, which showed very less activity as compared to *Brassica* CACTA elements (abundant elements). The first identified element *BoHARB1* is 3843 bp large in size with TAA TSDs and 42 bp TIRs. The 60% AT rich *BoHARB1* showed high AT content (75%) in the first 350 bp downstream to 5′ TIR, while several sub-terminal tandem repeats are present at both ends. It showed structural similarity to a 2.5 kb Harbinger named *DcMaster1* from *Daucus carota*, which was found inserted in the first intron of carrot vacuolar acid invertase isozyme-II gene. The insertion was characterized by TTA TSDs and 22 bp TIRs and 43 bp imperfect sub-terminal regions with AT rich region (80%) in 640 bp towards the 5′ terminal end. Due to the lack of any transposase coding protein, the element is considered as a non-autonomous Harbinger element (Grzebelus *et al.*, 2006).

5.4.5 Harbinger capture additional protein domains

All the major Harbinger encode two proteins as found in several eukaryotic genomes but few encode a third protein domain (Kapitonov and Jurka, 2004). *Brassica BoHARB1* only encodes a SANT protein. *BoHARB2* lack transposase but captures two other proteins named TRX and ATP11 protein families. *BoHARB3* and *BrHARB4* encode transposase and SANT proteins with one additional protein GPCR and NAM respectively, while *BrHARB5* only encode a SANT and NAM domains. The variability and evolutionary time of origin of these rearranged models will be interesting to study and may assist in understanding the origins of the tetraploid *Brassica* species. The domain organization of Harbinger elements from other species were investigated and found a similar range of variation in number and nature of ORFs. Examples include the 5.3 kb *HARBINGER* and

5.0 kb *ATIS112* element from *Arabidopsis thaliana* and 2.8 kb *HARB-1_Mad* from *Malus domestica* that only encode a transposase in their molecules. The *HARB-3_Stu* from *Solanum tuberosum*, Harbinger-*1_VV* from *Vitis vinifera*, and *MTISI12A* from *Medicago truncatula* encode transposase and SANT protein domains, which are present in majority of plant Harbinger. A 6.2 kb *HARB-2_ZM* from *Zea mays* encodes a transposase and NAM family of proteins. A 2.1 kb large *Triticum aestivum* Harbinger named *HARB-1_TA* only encodes a transposase. The *HARB-1_OS* from *Oryza sativa* and *HARB-10_SBi* from *Sorghum bicolor* encode a SANT domain and a transposase. Besides these typical domains, *HARB-1_OS* and *HARB-10_SBi* encode additional domains as NAM and CVV respectively, where CVV is Caulimovirus viroplasmin protein family (Table 5.6).

5.5 Conclusion

The CACTA and Harbinger superfamilies are ancient, abundant and evolutionary active components of the *Brassica* genome. Our detailed characterization in *Brassica* shows the diversity in structure of TEs i.e. TSD size and sequence, TIR sizes, length and ORF composition, which are characteristic of TE superfamilies and parallel the structures found in other well-analysed groups such as the *Triticeae* and *Solanaceae.* Notably, CACTA elements represent a 10-fold greater proportion of the *Brassica oleracea* (genome size 694 Mbp) than the A and B-genome *Brassicas* (527 and 633 Mbp), or the proportion in *Arabidopsis thaliana*. Since both CACTA and Harbinger elements found here have been shown to capture extra ORFs (such as the ATHILA-ORF1 protein) and subsequently amplify, they can affect genome evolution by their high copy number, by disruption at the site of insertion, or though the amplication of captured genes. The genome specificities of some of the CACTA (A and C) and Harbinger elements suggest that they will be valuable as probes for *in situ* hybridization to identify chromosome introgression and recombination events in hybrids (like the C-genome CACTA of Alix *et al*., 2008), but with the prospect of greater specificity and to the genomes. Since the PCR amplifications from different accessions within single species are sometimes showing polymorphisms, there is the potential to exploit these robust PCR markers for varietal identification, and perhaps for transposon-tagging of genes in appropriate populations as in systems based on *En/Spm* and *Ac/Ds* elements.

CHAPTER 6
NON-AUTONOMOUS DNA TRANSPOSONS & NOVEL INSERTIONS IN *BRASSICA* CROPS: DIVERSITY AND ABUNDANCE

Summary

Fifteen hAT transposon families were identified in the present work and estimated ~6505 and ~4664 copies of them from *Brassica rapa* and *Brassica oleracea* genomes respectively. The Mutator-like elements were few, mostly defective and have shown least activity in the *Brassica* genomes. The study revealed that non-autonomous DNA transposons are abundant compared to their autonomous counterparts. Transposon insertional polymorphism (TIP) markers were developed to study the insertion preference of transposons in diverse *Brassica* genomes and found that many elements are polymorphic across *Brassica* accessions. Some elements were A or C-genome specific, while most of them are present in *Brassica* diploids and allopolyploids. Several mobile insertion/deletions were also identified with or without TIRs and TSDs of varied sizes, not common to known superfamilies of transposons with internal non-coding regions. The detailed study of these insertions revealed that they are novel mobile insertions, which although less in number and small, are playing a role in genome size evolution.

6.1 Introduction

Transposable elements (TEs) are fundamental agents of genome evolution and diversification with the acquisition of functions independent of transposition in genomes. The abundance of whole-genome sequence data and advanced sequencing projects has increased our ability to identify and characterize complete complement of transposons within genomes by variable computational, molecular and cytogenetics studies, leading to a better understanding of the origins of transposons and their relationships within the genomes they reside. The classification of TEs into two major classes; retrotransposons and DNA transposons is quite universal, as first proposed by Finnegan, 1989 and updated later (Kumar and Bennetzen, 1999; Hansen *et al.*, 2005; Wicker *et al.*, 2007). The DNA transposons are further classified into superfamilies, of which Tc1-Mariner, hAT, CACTA, Mutator, Harbingers and P are common in plants, while others are common in animal genomes (Finnegan, 1989; Wicker *et al.*, 2007; Deragon *et al.*, 2008). The hAT

transposons constitute a large superfamily both in their numbers and diversity and there is an increasing interest to investigate their role played in the evolution of the plant species. A number of different active members of this superfamily have been discovered, and much remains to be learned about the activity and regulation of hAT elements, particularly when active forms are introduced into new hosts. In several previous investigation, the point of focused was on the transposons sequences closely related to the *Ac*, *hobo*, *Tam3, Tol2* and *Hermes* elements and the conclusion was that this superfamily is very ancient (Kempken and Windhofer, 2001; Rubin *et al.*, 2001). However recent analysis of hAT elements detected in 12 *Drosophila* species concluded that four clades (or families) of hAT elements could be identified (de Freitas Ortiz *et al.*, 2010). The hATs are recently investigated in several plants including maize (Du *et al.*, 2011), cereal grass (Muehlbauer *et al.*, 2006), sugar beet (Menzel *et al.*, 2012) and *Arabidopsis* (Bundock and Hooykaas, 2005)

Mutator (Mu) transposable elements are one major superfamily of DNA transposons and display a two- component system, one autonomous MULEs and many non-autonomous *Mu* elements, which have a mutagenic effect in maize (McCarty *et al.*, 2005). Eight Mutator-like elements (MULEs) are well characterized from maize (*Mu1*, *Mu2*/*Mu1.7*, *Mu3*, *Mu4*, *Mu5*, *Mu6/7*, *Mu8*, *MULE*) and three are classified as Mutator on the basis of 9 bp TIRs. MULEs represents the autonomous group of Mutator system with ~170-220 bp TIRs, while a few MULEs are characterized by 9 bp TSDs, 50-200 bp TIRs, heterogeneous and unrelated internal sequences. Mutator-like elements also named MULEs are identified from *Arabidopsis*, rice and other crops (Jiang *et al.*, 2011). The non-autonomous elements from other DNA transposon superfamilies are deletion derivatives of autonomous elements, but the internal sequences of *Mu* elements are often unrelated to their progenitors and showed high similarity to their host genome, suggesting a possible gene capture. These *Mu* elements were classified as Pack-MULEs, non-autonomous components of the Mutator transposon superfamily. The presence of Pack-MULEs was studied in plants which revealed the presence of 2853 elements in rice, 275 in maize, while only 46 Pack-MULEs copies were identified from *Arabidopsis* (Jiang *et al.*, 2004a; Jiang *et al.*, 2011). Despite of identification of several families of TEs, there are several others which need to be explored and sequencing will help to explore novel insertions.

The present study focussed on the identification and characterization of novel small non-autonomous elements or mobile insertions, which otherwise are not easy to identify by routine computational and molecular analysis i.e. based on homology or amplified by universal primers designed from conserved domains (like transposase, RT, INT). The present study focussed on the identification of small insertions, with or without protein domains in their internal regions.

6.2 Results and Discussion

6.2.1 hAT Elements

6.2.1.1 Identification of non-autonomous hAT transposons in *Brassica*

The non-autonomous hATs were identified by comparison of homoeologous BAC/genomic sequences. Four BAC pairs were the source of 15 varied types/families of hATs characterized in this study. The BAC pairs were 1) *Brassica rapa* clone KBrB028I01 (AC189298.1) x *Brassica oleracea* clone BoB028L01 (EU642504.1), 2) *Brassica rapa* clone KBrH004D11 (AC155341.2) against its homologue *Brassica oleracea* clone BOT01-64A15 (AC240089.1), 3) *Brassica rapa* clone KBrH080A08 (AC155344.2) x *Brassica oleracea* clone BOT01-121H07 (AC240081.1), 4) *Brassica oleracea* (EU579455) x *Brassica rapa* (CU984545) (see Conclusion; Figure 10.1-10.3). The maximum activity of hATs was observed in *Brassica oleracea* (AC240081.1), where 6 non-autonomous hATs were identified. Three hATs were identified from *Brassica oleracea* (AC240089.1) and three from *Brassica rapa* (AC155341.2). Two elements were identified from *Brassica rapa* accession 'AC189298.1' and one from *Brassica oleracea* 'EU642504.1'. One element was identified from *Brassica oleracea* (AC149635.1) accession. The results suggested that the hATs are dispersed in both A and C-genome *Brassica* and are actively proliferating in their genomes.

6.2.1.2 General features of *Brassica* hATs

The elements range in sizes from 402-3695 bp with canonical features of hATs i.e. 8 bp TSDs and TIRs ranging from 9-24 bp. The smallest hAT was a 402 bp element (*BoN-hAT6*) identified from *Brassica oleracea* (EU642504.1), while a 3695 bp large hAT-like

element was identified as insertion in a *Brassica oleracea* (AC240089.1) accession. All the other hATs range in sizes from 500-1000 bp. Majority of the elements have generated 8 bp TSDs, while few elements (*BoN-hAT7, BoN-hAT8, BoN-hAT13*) have produced 6-7 bp TSDs upon integration to the host site. The TSDs are AT rich in all the hATs, while TIRs are different and variable in numbers. The shortest TIRs were identified from *BoN-hAT11-2* (9 bp), while *BrN-hAT3* is flanked by 24 bp TSDs. *BoN-hAT6* elements are flanked by ~22 bp TIRs. The average size of TIRs among various *Brassica* hATs is between 9-15 bp. The detailed structural analysis of the elements revealed that no internal transposase related proteins were identified from any hAT suggesting their non-autonomous nature (Figure 6.1; Table 6.1).

6.2.1.3 Diversity and abundance of *Brassica* hATs

The sequences were used as query against *Brassica rapa* and *Brassica oleracea* Nucleotide Collection (nt/nr) database available in NCBI and number of complete copies were identified. It was found that the complete non-autonomous copies were very less as compared to their deleted copies and fossil remnants. The complete copies for each family were counted, which concluded that some elements are highly active as compared to the other hATs. The maximum numbers of complete copies were counted for *BoN-hAT7* family, which showed 94 and 4 copies of *Brassica rapa* and *Brassica oleracea* with an estimated copies of 962 and 553 in *Brassica rapa* and *Brassica oleracea* whole genome respectively. The second largest family is *BoN-hAT11-1*, where 951 and 415 copies were estimated for *Brassica rapa* and *Brassica oleracea* respectively (Table 6.1). Few hATs family were found to be genome specific such as *BrN-hAT1*, *BrN-hAT3*, *BrN-hAT4* and *BrN-hAT5* were detected only in *Brassica rapa*, while *BoN-hAT14* was only identified among *Brassica oleracea* accessions. This suggests that some elements are totally genome specific, or showed less activity in one genome but high proliferation in other. Some hATs were middle copy number, while others are low copy number families. The estimated number of copies for *BrN-hAT5* and *BoN-hAT15* in *Brassica rapa* were only 10, in contrast to 962 copies of *BoN-hAT7* (high copy number family). A total of 6505 and 4664 non-autonomous hATs were estimated from *Brassica rapa* and *Brassica oleracea* respectively belonging to 15 families. The high numbers in *Brassica rapa* is due to the presence of additional families, absent in *Brassica oleracea*, otherwise most of the families are common in both A and C-genomes (Table 6.1).

Table 6.1: List of non-autonomous hAT transposons with accessions numbers of *Brassica*, sizes, TSDs, TIRs, positions in the BAC sequences and estimated copy numbers (ECN) in *Brassica rapa* and *Brassica oleracea* genomes. The asterisks followed by TSDs are indicating mismatched TSDs, which are shown in small letters in TSD sequences. The total numbers of estimated hATs are mentioned at the end. Nucleotide sequences of hAT elements are available in Appendices (attached CD).

Family Name	BAC Accessions	Species	Size (bp)	TSD	TSD sequence	TIR (5´-3´)	Position	ECN in *B. rapa*	ECN in *B. oleracea*
BrN-hAT1	AC189298.1	*B. rapa*	670	8	AGTATTTT	*CTAGGCCTGGGCATT*	1727-2396	215	---
BrN-hAT2-1	AC189298.1	*B. rapa*	716	8	GTGTGGAC	*TCTATGTTACATGGAA*	32998-33712	511	---
BoN-hAT2-2	EU642504.1	*B. oleracea*	702	8*	TttCTCAT	*CATACGCGGAA*	68287-69477	501	10
BrN-hAT3	AC155341.2	*B. rapa*	620	8*	TaAAATTT	*GGCTTTGATTGGTAACATGTAGTA*	46457-47076	184	---
BrN-hAT4	AC155341.2	*B. rapa*	786	8	ATATTAGC	*TAGGGGTGGG*	105234-106019	82	---
BrN-hAT5	AC155344.1	*B. rapa*	979	8	GATAGACA	*CAGTGCCGGTCCGG*	34806-35784	10	---
BoN-hAT6	EU642504.1	*B. oleracea*	402	8	AATTGGAG	*CAGTGTTTTTAAAACCGGACCG*	88339-88740	256	100
BoN-hAT7	AC240081.1	*B. oleracea*	701	6*	TTTCgg	*TAGGGCTGGG*	4835-5532	962	553
BoN-hAT8	AC240081.1	*B. oleracea*	998	7	GGAATAC	*TATAATTTTTATT*	65232-66229	880	1096
BoN-hAT9	AC240081.1	*B. oleracea*	588	8	CCTAGTGT	*TAGGCCTGGGAC*	66529-67116	235	830
BoN-hAT10	AC240081.1	*B. oleracea*	570	8	CTAATAAC	*CCCGGTTCGGAAAAC*	71911-72481	51	277
BoN-hAT11-1	AC240081.1	*B. oleracea*	724	8	TAAAAATG	*TAGGGGTGGG*	87115-87838	931	415
BoN-hAT11-2	AC240081.1	*B. oleracea*	688	8	TACAAATG	*TAGGGGTGG*	106545-107232	951	415
BoN-hAT12	AC240089.1	*B. oleracea*	924	8	CCTACTCT	*TAGGGCCGTTCAATATGG*	8543-9466	501	415
BoN-hAT13	AC240089.1	*B. oleracea*	629	7	GCTTAGA	*GCATCTCCAA*	24325-24980	225	277
BoN-hAT14	AC240089.1	*B. oleracea*	3695	8*	CTTAAAct	*TTTGAATGGTAACCA*	59770-63464	---	138
BoN-hAT15	AC149635.1	*B. oleracea*	595	8	TTTGTAAC	*CAGTGTTCTAAAA*	52416-53003.	10	138
							Total estimated copy numbers:	**6505**	**4664**

6.2.1.4 Structural characterization of *Brassica* hAT transposons

The first identified hAT *BrN-hAT1* was found inserted in *Brassica rapa* (AC189298.1) accession from 1727-2396 bp. The element was 670 bp in size including 8 bp TSDs and 15 bp TIRs (GC rich), while the internal regions are AT rich (Figure 6.1). Blast searches yielded 21 complete copies, depending on which 215 copies were estimated for *Brassica rapa*. Another insertion was present in *Brassica rapa* (AC189298.1) accession starting from 32998-33712 in BAC sequence. The element is designated as *BrN-hAT2-1*, which display a size of 716 bp including 8 bp TSDs and 16 bp TIRs. A total of 50 copies were retrieved from database searches with an estimation of 511 copies from whole *Brassica rapa* genome. Another analogue of *BrN-hAT2-1* was identified from *Brassica oleracea* (EU642504.1) accession, which is named *BoN-hAT2-2* (702 bp). *BrN-hAT3* and *BrN-hAT4* were 620 and 786 bp large in sizes with 8 bp TSDs. The largest TIRs (24 bp) in *Brassica* hATs were identified in *BrN-hAT3*. *BrN-hAT5* is 1 kb element including 14 bp TIRs, identified in *Brassica rapa* (AC155344.1) (Table 6.1; Figure 6.1 & 6.2).

BoN-hAT6 is the smallest hAT element identified from *Brassica oleracea* (EU642504.1). It is 402 bp in size including 8 bp TSDs and 22 bp TIRs. *Brassica oleracea* accession 'AC240081.1' harbour six non-autonomous hATs named as *BoN-hAT7, BoN-hAT8, BoN-hAT9, BoN-hAT10, BoN-hAT11-1* and *BoN-hAT11-2* with a size range of 570-1 kb. All the elements generate perfect 8 bp TSDs, but *BoN-hAT7* generates 6 bp imperfect TSDs (Table 6.1; Figure 6.1). The detailed analysis of *Brassica oleracea* BAC accession 'AC240089.1' led to the identification of 3 hAT elements. *BoNhAT12* and *BoNhAT13* are 924 and 629 bp elements including hAT specific 8 bp TSDs. The TIRs are variable and no homology is observed between two elements representing two different hAT families. The largest hAT-like insertion was investigated in *Brassica oleracea* accession 'AC240089.1' from position 59770-63464. The element named *BoN-hAT14* is 3.7 kb in size with imperfect TSDs (6-8) and flanked by 15 bp TIRs. The internal region showed no coding regions for transposase or any other known proteins. The high AT content (70%), 6-8 bp TSDs and TIRs of the *BoN-hAT14* bring it under hAT superfamily due to similar characteristics with hATs. *BoN-hAT15* was found inserted in *Brassica oleracea* (AC149635.1) accession from 52416-53003 bp. The element is 595 bp in size including the 8 bp TSDs and 13 bp TIRs with non-coding internal region (Table 6.1; Figure 6.1).

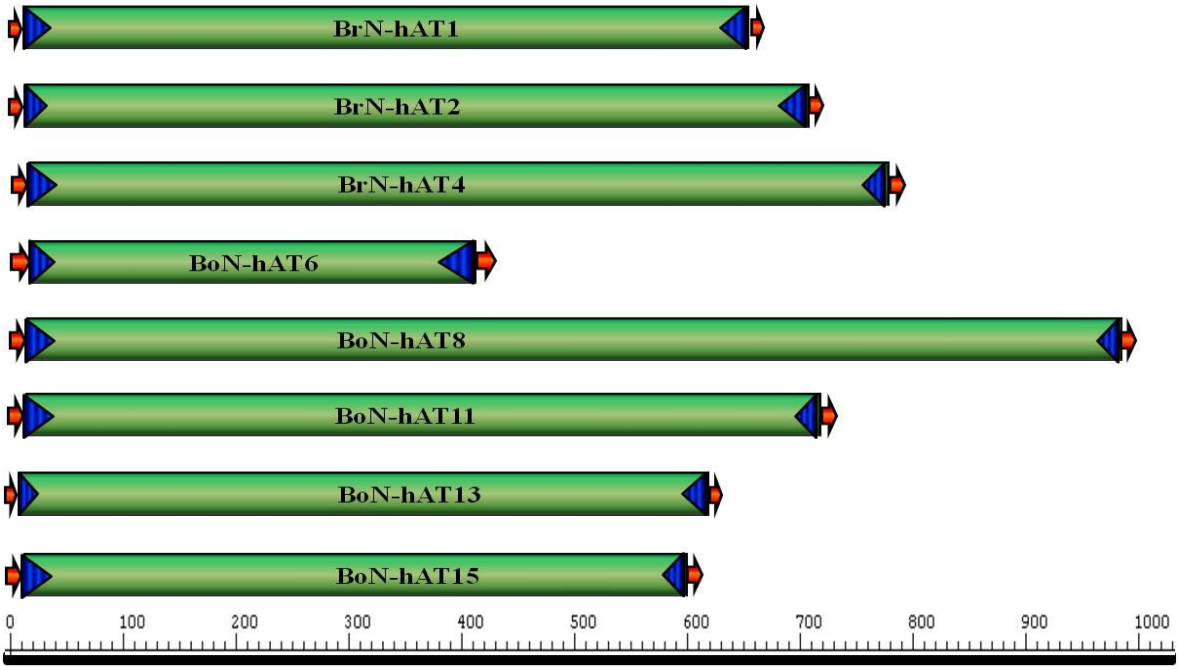


Figure 6.1: Schematic representation of non-autonomous hATs in *Brassica*. Red arrows indicate 8 bp TSDs and blue triangles represent TIRs. The internal regions of the *Brassica* hATs were highly heterogeneous without any protein coding regions.

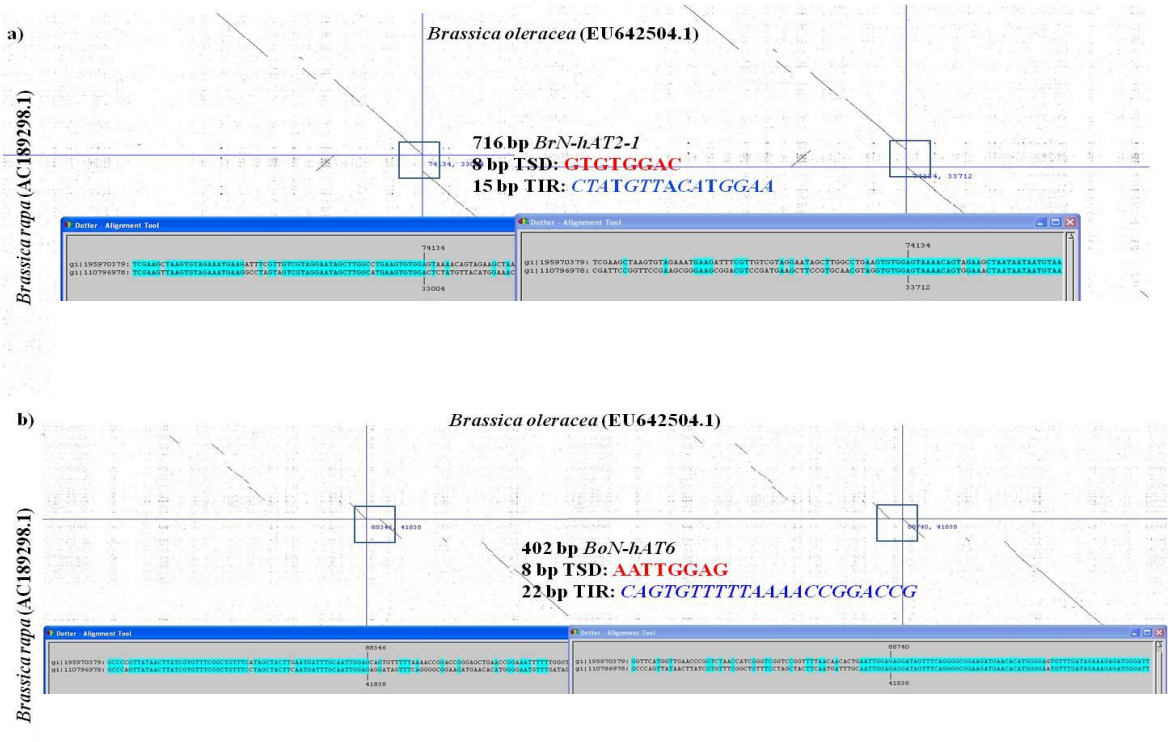


Figure 6.2: Dot plot comparison of homoeologous BAC clones *Brassica oleracea* (EU642504.1) (horizontal) against *Brassica rapa* (AC189298.1) (vertical) showed a) *BrN-hAT2-1* and b) *BoN-hAT6* insertion sites in *Brassica rapa* and *Brassica oleracea* respectively. The dot plots are shown twice (left and right) with large crosshair showing insertion points with the size, number and sequences of TSDs and TIRs indicated (insets) in the base pair alignments at the termini of the insertion sites.

6.2.1.5 TIP markers to study the hAT polymorphisms in *Brassica* accessions

Transposon insertional polymorphism (TIP) markers are appropriate to study biodiversity and evolution. Six primer pairs (Table 6.4) were designed from the flanking regions of the insertions to amplify the insertions (higher bands) or the flanking regions without insertions (lower bands). BrNhAT1F 5′-GCTACGTACATAGCAAAGGTG-3′ and BrNhAT1R 5′-CGTCAGACGGTTCTGTAAAAG-3′ were designed flanking the *BrN-hAT1* insertion in *Brassica rapa*. The insertional loci were amplified from all the six *Brassica rapa* diploid accessions (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons) and allopolyploids with A-genome in them as *Brassica juncea* (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna), *Brassica napus* (New, Mar, Last and Best, Fortune, Drakker, Tapidor) and four allohexaploid *Brassica*. No amplification was detected in C-genome *Brassica oleracea* and its allopolyploids. This suggests that *BrN-hAT1* is A-genome specific and is contributed to its allopolyploids by hybridization of A with B or C-genomes (Figure 6.3a). This was confirmed by blast searches, where no hits were found for *Brassica oleracea* but several complete copies were collected from *Brassica rapa*. The primer pair BrNhAT4F 5′-CAAGAAAGCTCAGATTCTTG-3′ and BrNhAT4R 5′-CAGGGAAACAAATAATACCC-3′ were designed to amplify *BrN-hAT4* (786 bp) insertion sites to amplify 925 expected product. The primers amplified the longer bands (product) from *Brassica rapa* and its allotetraploid and allohexaploid *Brassica* (Figure 6.3c).

Many other C-genome specific TIP markers were used to detect the amplification of hAT elements among 6 species comprising 40 accessions. The polymorphism pattern of *BoN-hAT2-2* and *BoN-hAT6* was nearly similar, but the two elements showed no homology. To amplify the *BoN-hAT2-2*, the primer pair BoNhAT2F and BoNhAT2R (Table 6.4) was designed to amplify ~1.1 kb product. The product was amplified from all *Brassica oleracea* diploid accessions plus allotetraploid and allohexaploid accessions. The bands in *Brassica napus* (AACC) were weak as compared to other species. Only upper bands (1.1 kb) were amplified from them, while lower bands with empty sites were amplified from *Brassica rapa* (AA), *Brassica juncea* (AABB), *Brassica napus* (AACC) and allohexaploid species (AABBCC) indicating that the short band is contributed by A-genome (Figure 6.3b). *BoN-hAT6* was amplified by primer pair BoNhAT6F and BoNhAT6R (Table 6.4) to

amplify a 775 bp hAT insertional site. Amplification was detected in *Brassica oleracea* and its allotetraploid and allohexaploid genomes (Figure 6.3d).

The *BoN-hAT10* was amplified with primer pair BoNhAT10F and BoNhAT10R (Table 6.4). The expected product size was 643 bp including the 570 bp hAT insertion. It was amplified from *Brassica oleracea* (De Rosny, Kai Lan, Early Snowball, Cuor Di Bue Grosso, Precoce Di Calabria, GK97361), *Brassica napus* (New, Mar, Last and Best, Fortune, Drakker, Tapidor), *Brassica carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK) and 4 allohexaploid *Brassica*. No amplification in *Brassica rapa* and its hybrids showed its absence from the A-genome and post divergence mobility of the *BoN-hAT10* element (Figure 6.3e). The insertion polymorphism of *BoN-hAT13* was observed by designing the primer BoNhAT13F and BoNhAT13R (Table 6.4). The expected product of ~650 bp was detected only from 3 *Brassica oleracea* (Kai Lan, Precoce Di Calabria, GK97361) and 4 *Brassica napus* (New, Mar, Last and Best, Fortune). No amplification from any of A-genome or allopolyploids (AABB, BBCC, AABBCC) suggests its recent bursts after the separation of *Brassica oleracea* crops (Figure 6.3f).

The results show members from hAT superfamily of DNA transposons are abundant in *Brassica* species, with more non-autonomous hATs than autonomous partners. The present study revealed ~6505 and 4664 copies of non-autonomous hATs from 15 hAT families from *Brassica rapa* and *Brassica oleracea* respectively (Table 6.1). By using the sequences as query in GenBank databases, the hits failed to retrieve their autonomous partners. This suggested that their autonomous copies are very low as compared to non-autonomous fellows. In a recent study, a total of ~610 complete or truncated non-autonomous hATs were detected in sugar beet in comparison to 81 hATs with their transposase coding regions. This suggest the abundance of non-autonomous hATs in *Beta vulgaris* genome (Menzel *et al.*, 2012). Several other analysis in *Drosophila* and other animals have shown the dominance and high copy numbers of non-autonomous hATs over their autonomous partners (de Freitas Ortiz *et al.*, 2010). The present study confirms the activity and mobility of the non-autonomous hATs by PCR analysis using 40 *Brassica* lines. The TIP markers showed different polymorphism patterns from different hAT families. The activity and mobility of the non-autonomous hATs indicates that their autonomous partners are also residing in their close premises, which provide them their

enzymatic machinery for their transposition and mobilization. The hATs are evolutionary an old family, so several degraded or partial fragments can be found dispersed in genomes (Rubin *et al.*, 2001).

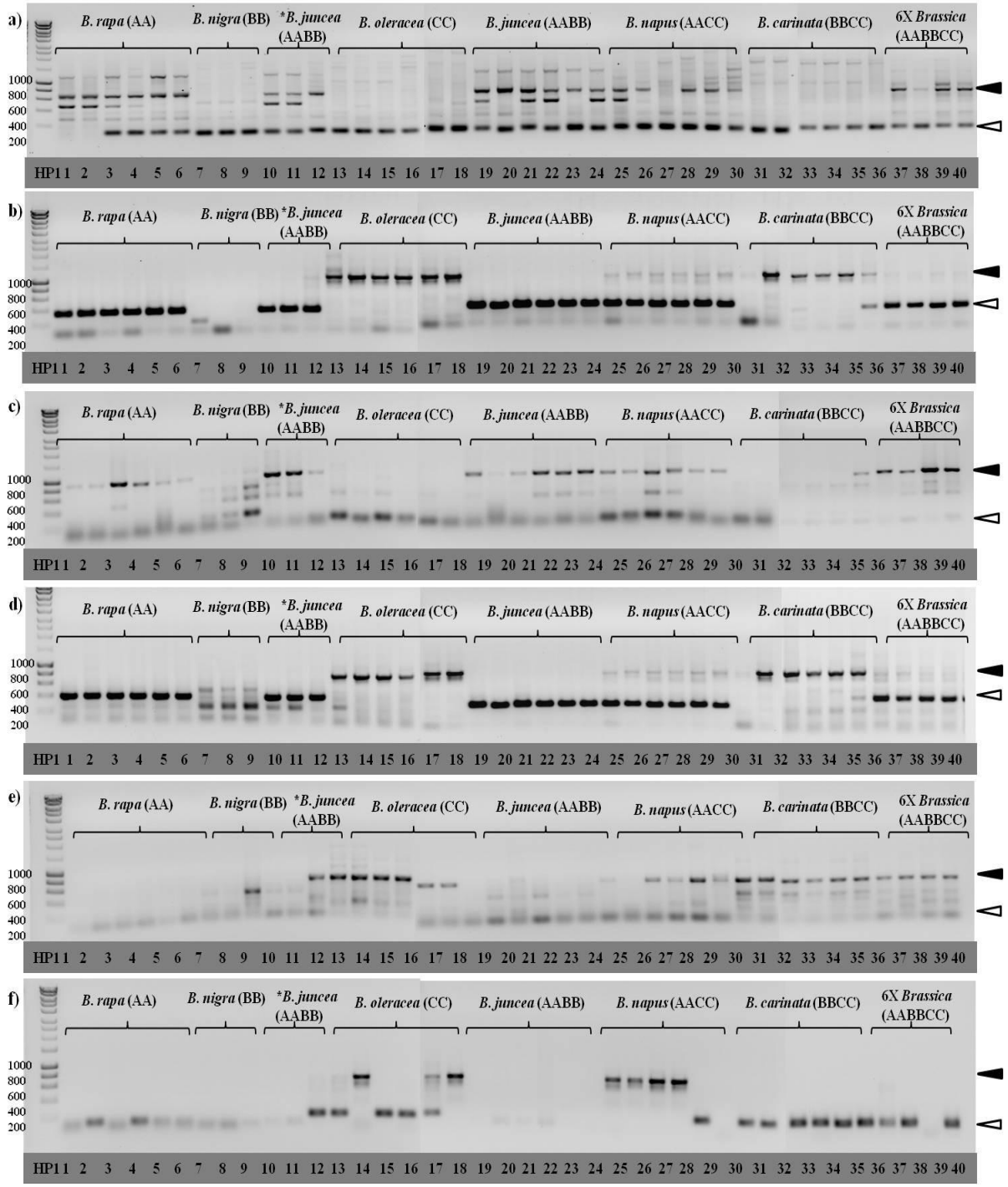


Figure 6.3: Transposon insertional polymorphisms of *Brassica* hATs. a) *BrN-hAT1*; b) *BrN-hAT2-2*; c) *BrN-hAT4*; d) *BoN-hAT6*; e) *BoN-hAT10*; f) *BoN-hAT13* insertion sites in various *Brassica* accessions. Long bands indicated by filled arrowheads (right) indicate amplified *hAT* insertions and short bands amplify the flanking sequences only (open arrowheads). All PCR figures show inverted images of size-separated ethidium bromide stained PCR products following agarose gel electrophoresis; numbers below the lanes identify each cultivar listed in table 2.1 and ladders indicate sizes in bp.

6.2.2 Mutator-like elements

6.2.2.1 Identification of Mutator-like elements (MULEs)

The first Mutator-like element (*BrN-MULE1-1*) in *Brassica* was identified by the comparison of two homoeologous BAC pairs *Brassica rapa* (A-genome; AC189298.1) against *Brassica oleracea* (C genome; EU642504.1). The element was 2781 bp long, flanked by 9 bp TSDs, 76 bp imperfect TIRs, no internal coding region suggesting a non-autonomous Mutator-like elements (MULE). After the identification of this element, the comparative analysis of *Brassica rapa* (AC155342.2) against its homologue *Brassica rapa* (AC146875.2) led to the discovery of a similar element (*BrN-MULE1-2*) with 9 bp TSDs and 76 bp TIRs. The elements showed high AT contents in their internal regions, which are not observed in other MULEs investigated in rice, maize and *Arabidopsis* plants. The studies revealed that few MULEs exhibit TIRs from 50-200 bp, so it showed typical features of MULEs by exhibiting 9 bp TSDs and 76 bp TIRs. The elements were used as a query against Nucleotide Collection and Whole-genome shotgun sequences databases in GenBank and retrieved only 6 complete sequences suggesting the low copy number family of MULEs in *Brassica*. The family of *Brassica* Mutator elements was named *Shahroz*.

6.2.2.2 Molecular characterization of *Shahroz* family of MULEs

The structural features of MULEs identified from *Brassica* revealed that the elements range in sizes from 2734-3160 bp. The first element was identified from *Brassica rapa* accession (AC189298.1) from 6034-8806 bp and is named *BrN-MULE1-1*. The element is 2781 bp in size with 9 bp TSDs 5´-GACCAACGA-3´ and 76 bp imperfect TIRs (Figure 6.4 & 6.5). The internal region of the element is AT rich with only 37% GC content. Another element (*BrN-MULE1-2*) was found inserted in *Brassica rapa* accession (AC146875.2) from 7268-10179 bp. *BrN-MULE1-2* is 2920 bp large in size including 9 bp TSDs (ATCCAGAAG) and 76 bp imperfect TIRs (Figure 6.4). *BrN-MULE1-1* and *BrN-MULE1-2* were compared and they have shown 52% homology in their entire lengths but the TIRs showed the highest homology. BLASTN searches resulted in the collection of 4 other homologous of similar sizes (~2.7-3.1 kb). *BrN-MULE1-3* is homologous (99%) to *BrN-MULE1-2* except TSDs and was identified from *Brassica rapa* accession (AC189583.2) from 42105-45024 bp. It is flanked by 9 bp TSDs (TAATTTCAA) and 76 bp TIRs, which are highly conserved among the two sequences (Figure 6.4; table 6.2).

BrN-MULE1-4 is 2734 bp large in size including the 9 bp TSDs and 76 bp imperfect TIRs and residing in *Brassica rapa* accession AENI01006341.1 from 227056-229789 bp and sequence retrieved from Whole-genome shotgun sequences database of GenBank. This revealed the presence of *BrN-MULE1-4* on chromosome number 7 and 8 of *Brassica rapa*. Two elements designated as *BrN-MULE1-5* and *BrN-MULE1-6* of 3160 bp sizes were identified from AENI01003197.1 and AENI01009183.1 accessions indicating their presence on chromosome 4 and 10 respectively (Table 6.2). Both elements are flanked by 9 bp TSDs and 76 bp TIRs (Figure 6.4).

6.2.2.3 *Shahroz* is A-genome specific non-autonomous *Brassica* MULEs family

The transposons are sometimes genome specific and showed high proliferation in one organism but not in their relative species. Our data confirms the distribution of *Shahroz* family of MULEs in A-genome *Brassica*. This was confirmed by the BLASTN and molecular analysis, where output imported 6 copies from *Brassica rapa* only and PCR yielded the products in A-genomes and its polyploids (AABB, AACC, AABBCC). To amplify a 3.2 kb product, the element was splitted into two parts and designs the primers; one from flanking region and other from central region of the element to amplify a 1528 and 1514 bp products from 5´ and 3´ ends respectively (Figure 6.6). The primers successfully amplified the 1528 and 1514 bp product from *Brassica rapa* and its allotetraploids (AABB, AACC, AABBCC), but no amplification in C-genome suggested its absence in *Brassica oleracea* and its allotetraploids (BBCC). A total of 40 *Brassica* cultivars were used to amplify the *Shahroz* MULE family. Out of these 40 lines, 6 were each from A and C-genomes, 3 were from B-genome, 9 were AABB, 6 were each AACC and BBCC and 4 were from AABBCC genomes. The 5´-1528 and 1514-3´ end products were amplified from two *Brassica rapa* (Chinese Wong Bok, San Yue Man) accessions among A-genomes. All the six *Brassica rapa* cultivars generates 2 additional bands of ~450 and 380 bp. *Brassica nigra* accessions showed no amplification. Nine *Brassica juncea* cultivars (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna) yielded amplification of both 5´-1528 bp and 1514-3´ end products. Among six *Brassica napus*, 2 cultivars (New, Fortune) amplified the bands. From the four synthetic allohexaploids (AABBCC), 3 yielded weak bands of expected sizes. The insertional polymorphisms of *Shahroz* family suggest its distribution in A-genome diploids and their allotetraploids but lacking in C-genome (Figure 6.6).

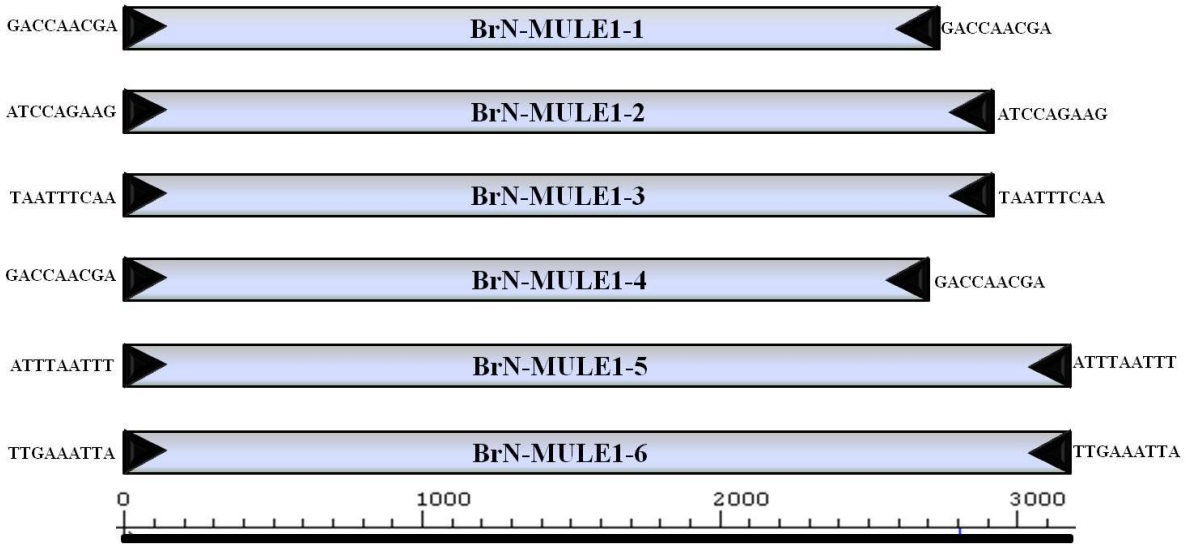


Figure 6.4: Schematic representations of *Brassica* non-autonomous Mutator-like elements from *Shahroz* family. Black triangles represent 76 bp TIRs. The TSDs are shown at both termini.

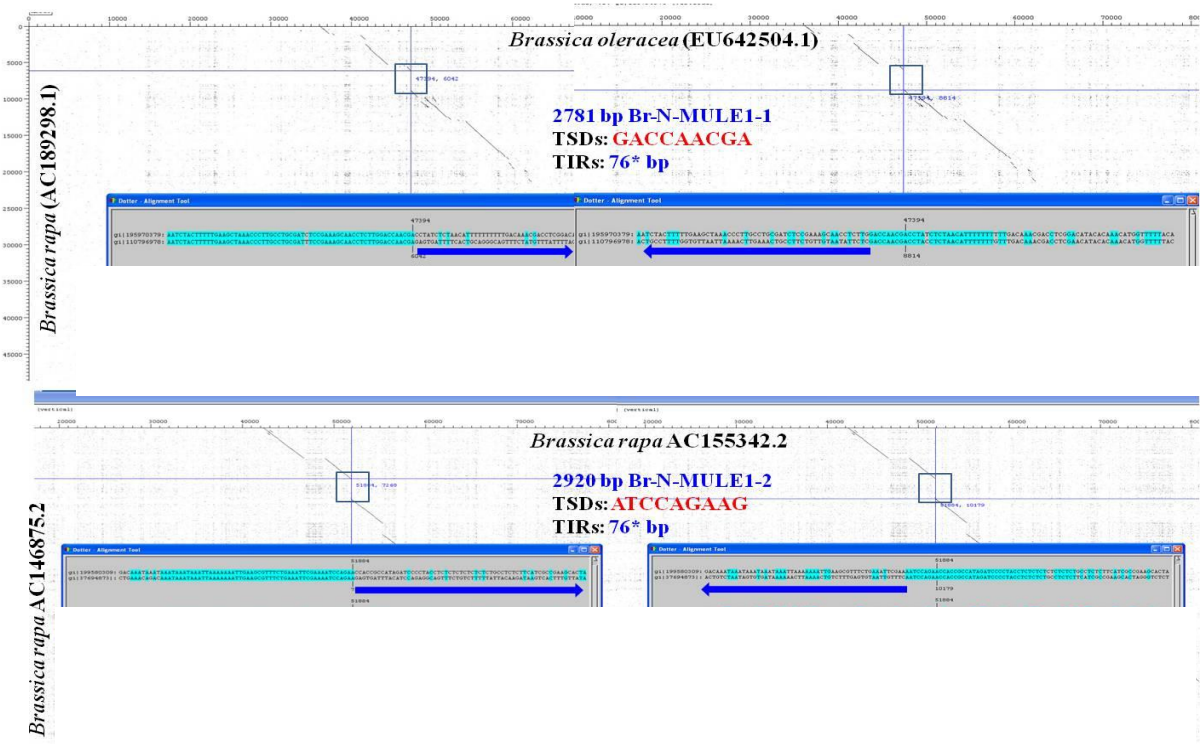


Figure 6.5: Dot plot comparison of homoeologous BAC clones of *Brassica* identified a) *BrN-MULE1-1* b) *BrN-MULE1-2* insertion sites in *Brassica*. The size, TSDs and TIRs are also indicated. The opposing arrows are indicating the TIRs on the sequence alignment insets. Asterisks after the 76 bp TIR indicate the imperfect TIRs.

6.2.2.4 *Shahroz* is an ancient, defective and low copy number family of MULEs

Shahroz is an ancient family of MULEs and passing the evolutionary stages of its degradation. The comparison and analysis of all the six *Brassica* MULEs showed very high homology (60-97%) in their TIRs in comparison to their internal regions (52-55%). There are many substitutions in their TIRs and conserved and varied regions within TIRs of these elements were found. The 3 bp termini (GAC) at 5´ and 9 bp termini at 3´ ends are highly conserved with some internal conserved T rich regions (Figure 6.7). No internal coding regions indicate them non-autonomous or defective elements. It is thought that the high variability and low degrading copies indicate the ancient nature of any family. The substitution rates are also used to count the ages of the elements. Considering this hypothesis, higher the variability within sequences, more ancient the family will be. It is thought that the TIRs of MULEs are responsible for the transposition activity of the elements. This means that the mutated TIRs with small insertions or substitutions indicate the inactive or defective elements (Jiang *et al.*, 2011). The *Shahroz* is a low copy number family with only 6 elements identified from *Brassica* databases in GenBank. Seventy copies of MULEs were estimated in *Brassica rapa* genomes. The low copy numbers of Pack-MULEs in plant genomes by Jiang *et al*., (2011) is supported by our analysis of presence of low copies of MULEs in *Brassicaceae*. The short, degrading nature of TIRs, low copy numbers and lack of non-coding regions indicate that *Shahroz* family is an ancient family in their degradation phase.

Table 6.2: List of non-autonomous Mutator transposons with accessions numbers, sizes, TSDs, TIRs, positions in the BAC sequences.

Name	Family	BAC Accession	Species	Size	TSD	TSD sequence	TIR	Position
BrN-MULE1-1	*Shahroz*	AC189298.1	*B. rapa*	2781	9	GACCAACGA	76	6034-8806
BrN-MULE1-2	*Shahroz*	AC146875.2	*B. rapa*	2920	9	ATCCAGAAG	76	7268-10179
BrN-MULE1-3	*Shahroz*	AC189583.2	*B. rapa*	2920	9	TAATTTCAA	76	42105-45024
BrN-MULE1-4	*Shahroz*	AENI01006341.1	*B. rapa*	2734	9	GACCAACGA	76	227056-229789
BrN-MULE1-5	*Shahroz*	AENI01003197.1	*B. rapa*	3160	9	ATTTAATTT	76	23051-26210
BrN-MULE1-6	*Shahroz*	AENI01009183.1	*B. rapa*	3160	9	TTGAAATTA	76	751-3911

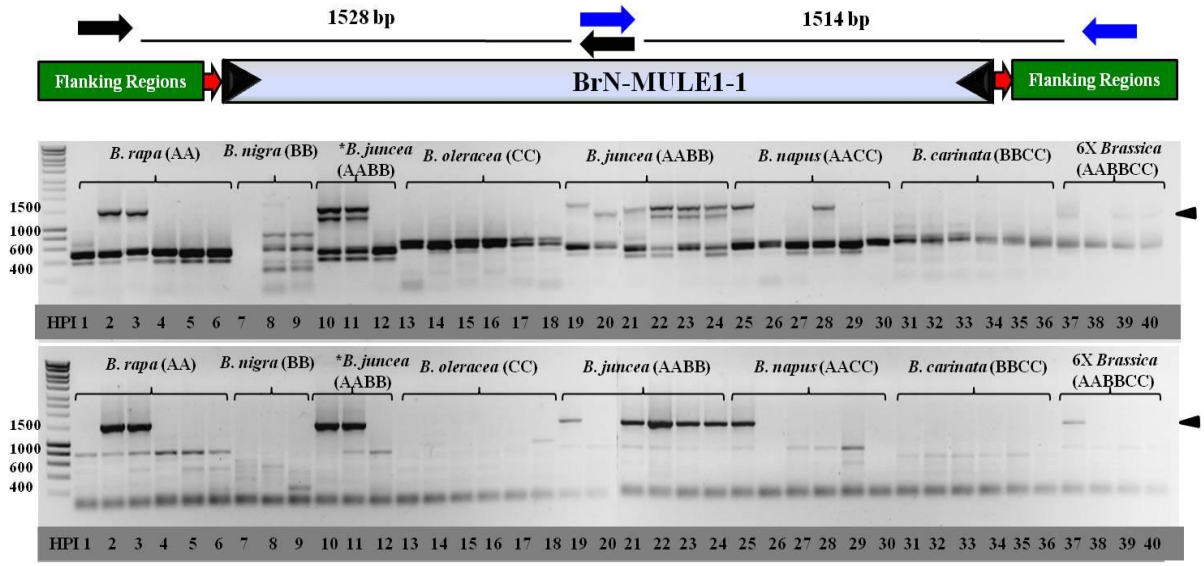


Figure 6.6: Insertion polymorphisms of Mutator elements in *Brassica*. a-b) The amplification of *BrN-MULE1-1* insertion sites by primers in various *Brassica* accessions*:* Long bands (1528bp/1514) show the amplified element and short bands amplify the pre-insertion (empty) sites only. Many polymorphisms between accessions with the same genome constitution are evident.

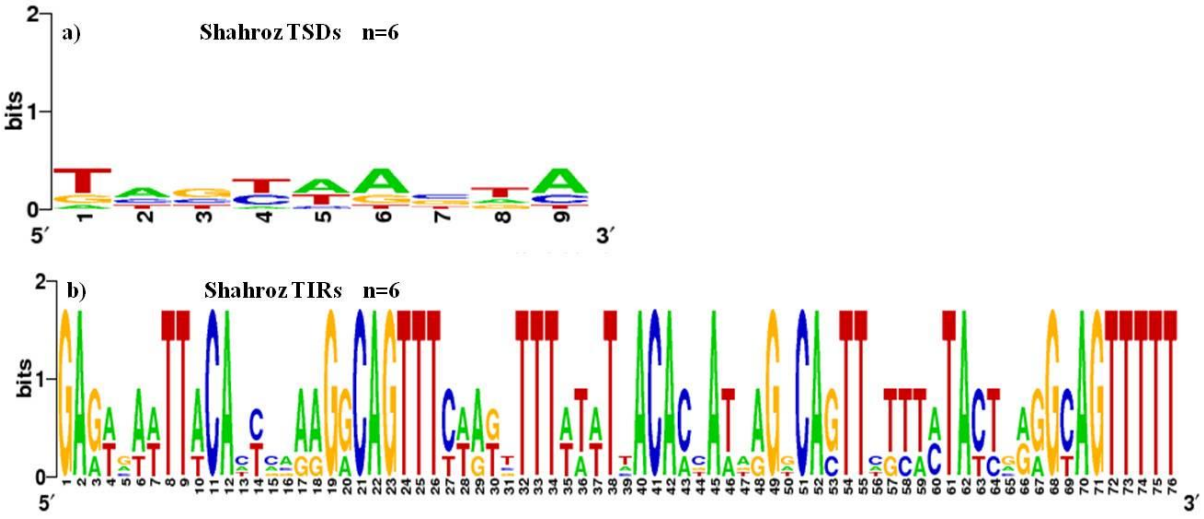


Figure 6.7: Pictograms showing the information content of *Shahroz* family of Mutator DNA transposons indicating a) weak conservation of TSDs; b) strong conservation of half the bases (36/76) in TIRs, with the T conserved motifs dispersed in the TIRs. The height of nucleotides represents the proportion conserved (0 to 2).

6.2.3 Mobile insertions of unknown superfamilies

The dot plot analysis of homoeologous BAC sequences for the identification of various TEs yielded several insertions, which displayed structural features not observed in known superfamilies of TEs. The mobile insertions were detected in the same way as autonomous

and non-autonomous transposons related to known families were identified. The insertions generated perfect TSDs and some are flanked by TIRs, but others lack any TIRs (and hence could be deletions unrelated to transposons). The BLAST results showed that several other copies of these elements are dispersed in *Brassica* genomes flanked by TSDs only and preliminary, computational, molecular and genetic analysis suggest their mobile nature. They were named *BrAT* and *BoAT* in *Brassica rapa* and *Brassica oleracea* respectively, where first letter represent genus, second indicate species, letter 'AT' indicate associated transposon. These mobile insertions were detected from several BAC sequences predominating in *Brassica oleracea* accession 'AC240081.1', where several other hATs and known TE superfamilies were identified. The sizes of the insertions range from 174-2252 bp with displaying heterogeneous sequences. The TSDs starts with 2 bp to 9 bp, but TIRs are either missing or if present are very short (3-4 bp). *Brassica rapa* accession 'AC155344.1' harbour 2 insertions; a 1423 bp *BrAT1* and a small insertion of 174 bp (*BrAT2*), with 4 and 2 bp TSDs an no TIRs flanking the insertions. A large insertion named *BrAT4* (2045) was detected from *Brassica rapa* accession 'AC155342.2' from 73917-75961 bp. The insertion is flanked by 8 bp TSDs but no TIRs were identified. Two elements with a size of 1422 and 1406 were identified with TAA TSDs, but no TIRs restrict their characterization (Table 6.3).

Two insertions were found embedded in *Brassica oleracea* accession 'EU642504.1' with a size of 1284 (*BoAT8*) and 242 bp (*BoAT9*) and flanked by 7 and 5 bp TSDs with no structural similarities to known TEs. *Brassica oleracea* accession 'EU579455.1' harbour 2 insertions; 958 bp (*BoAT10*) and 531 bp (*BoAT11*) with 4 and 7 bp TSDs and no TIRs. The highest mobility of small insertions is observed in *Brassica oleracea* accession 'AC240081.1', where six novel insertions of unknown superfamilies were found in addition to several other DNA transposons. The insertions (*BoAT12- BoAT17*) range in size from 183-2067 bp. A 1684 bp insertion including 2 bp TSDs was only element flanked by 4 bp TIRs, the remaining insertions lack any TIRs or internal coding regions. Three other insertions were detected from *Brassica oleracea* accession 'AC240089.1', with sizes ranging from 184-2252 bp. A 720 bp insertion is flanked by 9 bp TSDs, while the largest insertion (2252 bp) only generated 6 bp TSDs. No TIRs from any of the three insertions indicate that the insertions are novel, and there are several other mobile insertions other than known superfamilies which need further classification (Table 6.3).

Table 6.3: List of various mobile insertions of unknown superfamilies with accessions numbers, sizes, TSDs, TIRs, positions in the BAC sequences.

Sr.No.	BAC Accession	Species	Size	TSD sequence	TIRs	Position	Super-family
BrAT1	AC155344.1	*B. rapa*	1423	TATC	ND	41352-42774	Unknown
BrAT2	AC155344.1	*B. rapa*	174	TT	ND	46199-46372	Unknown
BrAT3	AC155341.2	*B. rapa*	446	GCTTCTTA	ND	36040-36485	Unknown
BrAT4	AC155342.2	*B. rapa*	2045	AAAACATA	ND	73917-75961	Unknown
BrAT5	AC155342.2	*B. rapa*	643	AGACT	ND	65743-66385	Unknown
BrAT6	AC146875.2	*B. rapa*	1422	TAA	ND	50164-51585	Unknown
BrAT7	AC189335.2	*B. rapa*	1406	TAA	ND	34792-36197	Unknown
BoAT8	EU642504.1	*B. oleracea*	1284	GTTTTTT	ND	89403-90686	Unknown
BoAT9	EU642504.1	*B. oleracea*	242	CTAAT	ND	107526-107823	Unknown
BoAT10	EU579455.1	*B. oleracea*	958	CTCA	ND	11695 -12649	Unknown
BoAT11	EU579455.1	*B. oleracea*	531	CCTATAA	ND	27245-27769	Unknown
BoAT12	AC240081.1	*B. oleracea*	1684	GC	AAAC	16577-18260	Unknown
BoAT13	AC240081.1	*B. oleracea*	2067	CTT	ND	26649-28715	Unknown
BoAT14	AC240081.1	*B. oleracea*	1113	TTGTT	ND	75213-76325	Unknown
BoAT15	AC240081.1	*B. oleracea*	183	GA	ND	77359-77541	Unknown
BoAT16	AC240081.1	*B. oleracea*	794	TA	ND	8243-83228	Unknown
BoAT17	AC240081.1	*B. oleracea*	1037	ATCTTTTAA	GGG	98654-99690	Unknown
BoAT18	AC240089.1	*B. oleracea*	184	TGT	ND	10661-10884	Unknown
BoAT19	AC240089.1	*B. oleracea*	720	AATAGAAAT	ND	48851-49570	Unknown
BoAT20	AC240089.1	*B. oleracea*	2252	TTAGAC	ND	57348-59599	Unknown

Table 6.4: List of *Brassica* hAT and Mutator primers with size of the elements, size of the expected products, names and sequences of primers.

Sr. No.	Super-family	TE family	TE Size	Product Size	Primer Name	Primer Sequence
1	hAT	*BrN-hAT1*	670	863	BrNhAT1F BrNhAT1R	GCTACGTACATAGCAAAGGTG CGTCAGACGGTTCTGTAAAAG
2	hAT	*BoN-hAT2-2*	715	1090	BoNhAT2F BoNhAT2R	GGGCAAAGGCCATCTATGCA ATGTACGACTCCGTCAATGA
3	hAT	*BoN-hAT4*	786	925	BoNhAT4F BoNhAT4R	CAAGAAAGCTCAGATTCTTG CAGGGAAACAAATAATACCC
4	hAT	*BoN-hAT6*	402	775	BoNhAT6F BoNhAT6R	GTGAAAATGGTGGCCAGTCT TTTGGAGGTTTTGGTGAAGG
5	hAT	*BoN-hAT10*	570	643	BoNhAT10F BoNhAT10R	GACTTTTCAAGTCAAAGCAA CTTTAACATTGATGAGCTGC
6	hAT	*BoN-hAT10*	629	749	BoNhAT13F BoNhAT13R	CTTCTCCCGTGTAATGAATG CACACAACCTGCACAAATAG
7	Mutator	*BrN-MULE1*	2781	1528	BrNMULE1F BrNMULE1R	GAACATGGTCACCTTCACTG CATGGTTAGAAACCGTGTGG
8	Mutator	*BrN-MULE2*	2781	1514	BrNMULE2F BrNMULE2R	CCACACGGTTTCTAACCATG ACGGGGAAATGAAACTGTAG

6.3 Conclusions

With sequencing data, our knowledge of new families of TEs is increasing. It is obvious that some elements are abundant in one species and less active in others, or even active in some specific chromosomal regions and less active in others. Within the genomes, several mobile insertions are actively proliferating. Mobile insertions can be inserted to or near genes and can alter their function as observed by small SINE and MITE-like elements (Bennetzen, 2000; Deragon *et al.*, 2008; Feschotte, 2008). In this study, several insertions of variable sizes and structures embedded in genomes were detected, with or without TIRs and generating TSDs not common in known families (Table 6.3). Although fewer in number and small in size, it is concluded that they have a role in the diversity and evolution of the organisms by their duplications, capturing the host genes and transduplicating them. Non-autonomous DNA transposons are abundant and dispersed in *Brassica* genomes, with hAT elements most abundant in both *Brassica rapa* and *Brassica oleracea*, while Mutator-like elements were rare with degraded copies. Several mobile insertions were identified, which have no structural features of known superfamilies of TEs. The hAT specific TIP markers are highly informative to study the biodiversity and evolution of plant genomes. Overall, this strategy helped us in identification of several non-autonomous DNA transposons, their deletion derivatives and novel mobile insertions with structural features different from known superfamilies of TEs.

CHAPTER 7

POPULATION DYNAMICS OF MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS (MITEs) IN *BRASSICA*

Summary

Miniature inverted-repeats transposable elements (MITEs) are an abundant component of plant genomes, contributing to their evolution and diversification. A dot plot based approach was developed for *de novo* identification of 15 novel families of MITEs in *Brassica* genomes and systematically their homologues were examined for further analysis. Out of the 15 families discovered, 5 are Stowaway-like MITEs with TA target site duplications (TSDs), 4 Tourist-like with TAA/TTA TSDs, 5 Mutator-like with 9-10 bp TSDs and 1 novel MITE named *BoXMITE* flanked by 3 bp TSDs. The TIRs in the members of the same family are highly conserved while showing variability in different families. About 29112 MITE-related sequences were estimated in the diploid *Brassica* species (A and B-genomes). All the MITEs have high AT rich (~74-82%) regions and have insertional preference in high AT regions, mostly as high copy number families. PCR analyses were performed using TIP based degenerate primers designed from flanking sequences of MITE elements. This analysis detected MITE insertional polymorphism at many insertion sites in *Brassica* diploids and polyploids. Many of the BLASTN hits yielded strong hits in various genes suggesting that many MITEs reside in gene regions. The identification of *Brassica* novel MITEs will have broad applications in the analysis and annotation of *Brassica* genomes and their use as DNA markers and mutagens for the genomics of *Brassica* and its related species.

7.1 Introduction

Miniature inverted-repeats transposable elements (MITEs) represent a heterogeneous group of short non-autonomous mobile DNA elements, generates flanking TSDs, TIRs or having varied lengths, exhibit high AT-rich sequences and posses high copy numbers. In recent years, their diversity, abundance and distribution has been investigated in plant genomes. They are divided into two major groups based on structural features: Stowaway-like MITEs that generate TA TSDs, and Tourist-like MITEs, which generate TAA TSDs upon integration to a new site (Jiang *et al.*, 2004b; Zhang *et al.*, 2004).

Recent discoveries revealed that MITEs derived from almost all DNA superfamilies including hAT, CACTA, Mutator, PiggyBac (Benjak *et al.*, 2009; Kuang *et al.*, 2009; Menzel *et al.*, 2012). Full length DNA transposons are thought to be the evolutionary progenitors of MITEs by a mechanism of their truncation where they lost their internal coding regions. They are characterized by their progenitors based on sequence conservation between MITEs and autonomous DNA transposons. The TSDs and TIRs are the best source to identify the MITEs (Feschotte and Mouches, 2000; Yang and Hall, 2003; Jiang *et al.*, 2004a). As non-autonomous short elements <600 bp, MITEs lack any protein coding domains including the transposase protein, necessary for their transposition and integration to a new sites. As derivatives of active DNA transposons, several investigations suggest that MITEs could be cross-activated by their autonomous partners from which they have derived (Wicker *et al.*, 2007). The role of MITEs in the gene expression and diversification of important crops like rice (Lu *et al.*, 2012), barley (Lyons *et al.*, 2008) and other plants have been studied in detail in the recent years.

In the present study, dot plot and BLASTN based approaches were used for the *de novo* identification of different MITE families proliferating in *Brassica* genomes. The identification, characterization, diversity, distribution and amplification patterns of MITEs were systematically investigated. The effects of MITEs on gene expression and their contribution to the diversity of *Brassica* were analyzed.

7.2 Results

7.2.1 Identification of MITE families in *Brassica*

Fifteen novel families of MITEs in *Brassica* genomes by dot plot sequence comparisons have been identified. The comparison of *Brassica rapa* (AC189298) x *Brassica oleracea* (EU642504.1) accessions led to the identification of *BrTOUR3-1* (AC189298), *BoSTOW3-1* (Figure 7.1a) *BoXMITE1-1* and *BoMuMITE5-1* (EU642504.1). The comparative analysis of *Brassica rapa* (CU984545.1) x *Brassica oleracea* (EU579455.1) gave the detection of *BrMuMITE1-1* (Figure 7.1b), *BrMuMITE2-1* (CU984545.1), *BoMuMITE4-2, BoMuMITE5-2* (EU579455.1). The dot plot sequence comparison of *Brassica rapa* (AC155341.2) against its homoeologue *Brassica oleracea* (AC240089.1) revealed the identification of *BoSTOW4-1* (AC240089.1). The comparison of *Brassica rapa* (AC155344.1) accession against its homoeologue *Brassica oleracea* (AC240081.1)

led to the identification of *BrSTOW1-1*, *BoSTOW2*, *BrTOUR1-1*, *BrTOUR2-1*, *BrMuMITE6-1* (AC155344.1), and *BoSTOW5*, *BoTOUR4* (AC240081.1). After the identification of these MITE sequences, *Brassica* MITE sequence dataset was extended using the representatives of each family as query in BLASTN searches against *Brassica* Nucleotide Collection (nr/nt) database in the GenBank. The full length sequences showing >70% coverage and identity were retrieved. Due to very high copy number of MITEs, only 5-10 sequences were collected as representatives of each family for detailed analysis (Table 7.1).

7.2.2 Characterization and classification of *Brassica* MITEs

Based on the structural features (TSDs and TIRs) of the known superfamilies of TEs, the *Brassica* MITEs were classified into Stowaway, Tourist and Mutator-like MITEs. One MITE family exhibiting 3 bp TSDs remained un-classified due to the lack of any clear marks or strong homology to any known superfamily of MITEs and named *BoXMITE*. The sizes of the members from various *Brassica* Stowaway families range in size from 227-580 bp with perfect TA TSDs. Four families of MITEs generate TAA/TTA TSDs, which are typical features of Tourist-like MITEs derived from PIF/Harbinger. The members of *Brassica* Tourist-like MITEs are also small in sizes ranging from 258-413 bp. Five Mutator-like MITE families were identified, where the sequences exceeds the MITEs length limitations (>600 bp) but they retained the MITE structural features such as TSDs, TIRs, non-coding regions, high AT rich sequences and high copy numbers. The TIRs of these elements are long and starts from terminal regions upto the central regions with perfect 9-10 bp flanking TSDs (Table 7.1). These elements are considered as Mutator-derived MITEs due to similarities in TSDs and TIRs of these elements with Mutator transposons, but no homology to any Mutator element was found.

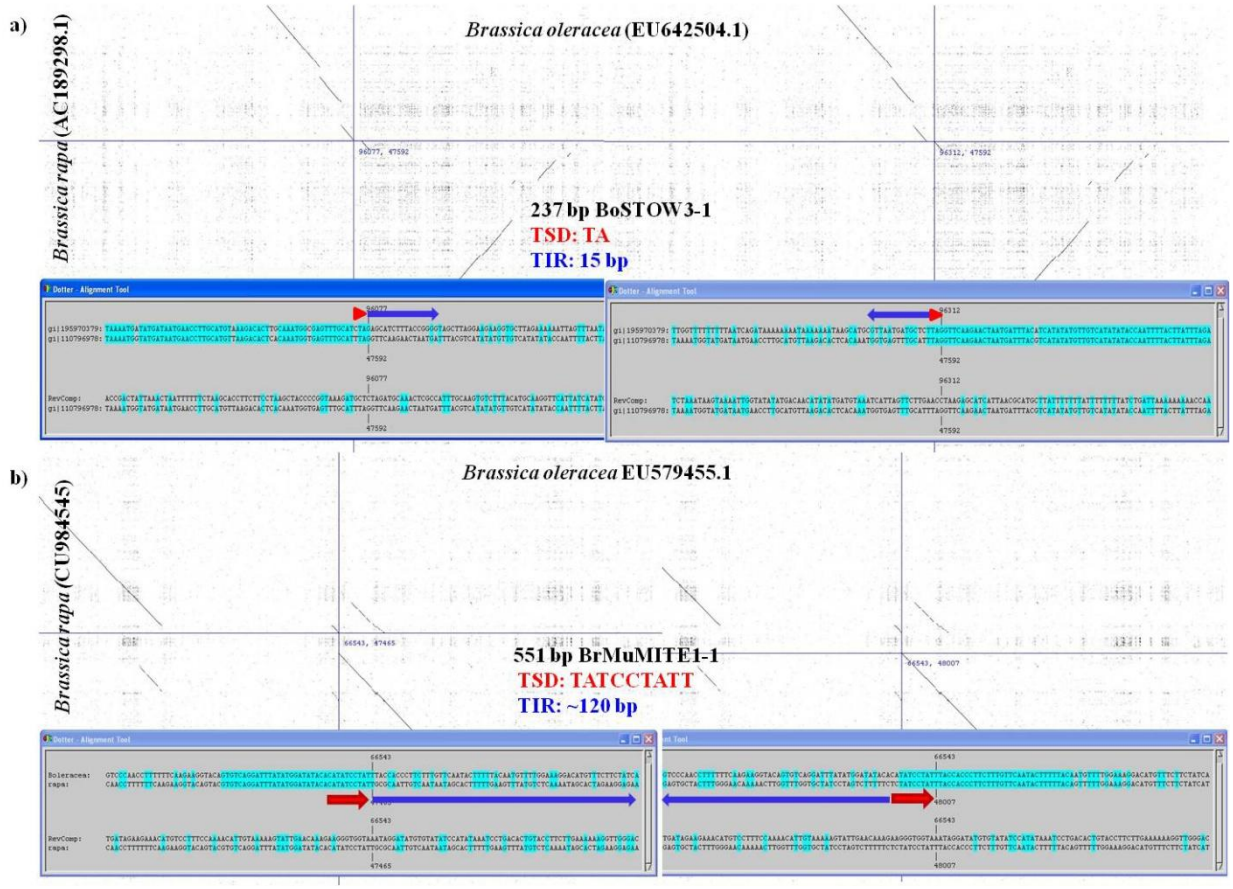


Figure 7.1: Dot plot comparison of *Brassica rapa* (vertical) and *Brassica oleracea* (horizontal) BAC clones for MITE identification. The central diagonal line running from one corner to other indicates the homologous regions between A and C-genomes. The gaps indicate the MITE insertions in *Brassica oleracea* (a) and *Brassica rapa* (b). Associated with the MITE activity, the dot plots show other features of genome structural changes including (a) an inversion and (b) duplications. The inset sequence alignments show the TSDs (red) at both ends of the MITE insertions and the TIRs (blue arrows) from the alignment points (left and right parts of dot plots) with the large cross.

Table 7.1: MITE families identified from *Brassica* BAC sequences with names of elements, sizes, TSDs and TIRs. The asterisks in front of TSDs or TIR indicate a mismatch at 5´ or 3´ TSDs or TIRs. Nucleotide sequences of representative MITEs are available in Appendices (attached CD). ND: Not determined.

Element Name.	BAC Accession	Species	Size	TSDs	Terminal inverted repeats (TIRs)	AT %	MITE Superfamily
BrSTOW1-1	AC155344.1	*B.rapa*	580	TA	*TACCTTTCTGTTCCTAAATATAAGATGTTT*	76	MITE/Stowaway
BrSTOW1-2	AC232537.1	*B.rapa*	329	TA	*GACTCAGGGCCGGCTTACAA*	68	MITE/Stowaway
BrSTOW1-3	AC232534.1	*B.rapa*	329	TA	*GACTCAGGGCCGGCTTACAA*	68	MITE/Stowaway
BrSTOW1-4	AC189530.2	*B.rapa*	328	TA	*GACTCAGGGCCGGCTTACAA*	68	MITE/Stowaway
BrSTOW1-5	AC189319.1	*B.rapa*	324	TA	*GCAGGGCCGGCTCAA*	68	MITE/ Stowaway
BoSTOW2-1	AC240081.1	*B.oleracea*	448	TA	*GGCGCTAGTCG**	70	MITE/ Stowaway
BoSTOW2-2	EU579455.1	*B.oleracea*	460	TA	*GGTGCTAGTCG**	70	MITE/Stowaway
BoSTOW2-3	AC152123.1	*B.oleracea*	442	TA	*GGCGCTAGTCG**	68	MITE/Stowaway
BoSTOW2-4	AC183493.1	*B.oleracea*	436	TA	*GGCGCTAGTCG**	72	MITE/Stowaway
BrSTOW2-5	AC189511.1	*B.rapa*	422	TA	*GGCACTAGTCG**	73	MITE/ Stowaway
BoSTOW3-1	EU642504.1	*B.oleracea*	237	TA	*AGAGCATCTTTACCG*	58	MITE/Stowaway
BoSTOW3-2	AC232493.1	*B.oleracea*	244	TA	*TGAGAGCATCTTT*	66	MITE/ Stowaway
BoSTOW3-3	AC229603.1	*B.oleracea*	243	TA	*GAGCATCTTTAAATA**	58	MITE/Stowaway
BoSTOW4-1	AC240089.1	*B.oleracea*	227	TA	*CTGTTTCCGTTTTACAAAGATATACTTTTT*	81	MITE/Stowaway
BoSTOW4-2	AB180902.1	*B.oleracea*	248	TA	*CTCCCTCCGTTCGTTAATGATAGAATTTTTAG*	78	MITE/Stowaway
BrSTOW4-3	AC189452.2	*B.rapa*	256	TA	*CTCTCTCCGTTTCGAAAAGATATATATTTTAG*	82	MITE/Stowaway
BrSTOW4-4	AC189417.2	*B.rapa*	253	TA	*CTCCTTCCATTTCAAAAAGATAGACTTTTTAGTA*	81	MITE/Stowaway
BrSTOW4-5	AC189322.2	*B.rapa*	251	TA	*CTCCTTCCGTTTCACAAAGATAGACTTTTTAG*	80	MITE/ Stowaway
BrSTOW4-6	AC189444.2	*B.rapa*	251	TA	*CTCCTTCCGTTCCTAAAATATATACTTTTTAG*	80	MITE/Stowaway
BrSTOW4-7	AC232543.1	*B.rapa*	248	TA	*CTCCATCCGTCCTAAAAGATAAATTTTTTAG*	79	MITE/ Stowaway
BrSTOW4-8	AC232514.1	*B.rapa*	245	TA*	*CTCCATCCGTTTAAAAAAGATAGATGTTTT*	79	MITE/Stowaway
BrSTOW4-9	AC189476.2	*B.rapa*	233	TA	*CTCTGTTCTTTAAAAATAGATTTTCTAG*	79	MITE/Stowaway
BrSTOW4-10	AC189492.2	*B.rapa*	218	TA*	*CTCCATTCAACAAAAATATATATTTTA*	82	MITE/Stowaway
BoSTOW5-1	AC240081.1	*B.oleracea*	243	TA	*TATTTCTTCCGTTTCGATTTA*	80	MITE/ Stowaway
BoSTOW5-2	AC240087.1	*B.oleracea*	243	TA	*CTCCCTCCGTTTCATATCA*	74	MITE/Stowaway
BoSTOW5-3	AC183492.1	*B.oleracea*	241	TA	*CTCCATCCGTTTCATATTA*	74	MITE/ Stowaway
BrSTOW5-4	AC232467.1	*B.rapa*	244	TA	*CTCCCTCCGTTTCGATTTA*	76	MITE/Stowaway
BrSTOW5-5	AC189391.2	*B.rapa*	242	TA	*CTCTCTCCGTTTCATTTTA*	74	MITE/ Stowaway
BnSTOW5-6	AJ291500.1	*B.napus*	242	TA	*CTCCCTCTGTTTCATCATA*	74	MITE/Stowaway
BrSTOW5-7	AC241048.1	*B.rapa*	241	TA	*TTCCTTCCGTTTCATTTTA*	76	MITE/Stowaway
BrSTOW5-8	AC189207.2	*B.rapa*	239	TA*	*CTCTCTCCGTTTCATTTTA*	78	MITE/Stowaway
BrSTOW5-9	AC189417.2	*B.rapa*	242	TA	*CTCCCTCCATTTCATTTTA*	72	MITE/ Stowaway
BrSTOW5-10	AC189565.2	*B.rapa*	245	TA	*CTCCCTCCATTTTATAATA*	78	MITE/Stowaway
BrTOUR1-1	AC155344.1	*B.rapa*	413	TTA	*GGGGGTGTTAGTGGGA*	73	MITE/Tourist
BrTOUR1-2	AC232445.1	*B.rapa*	421	TAA	*GGGGGTGTTAGTG*	79	MITE/Tourist
BrTOUR1-3	AC189390.2	*B.rapa*	418	TAA	*GGGTGTTAGTGGGA*	76	MITE/Tourist
BrTOUR1-4	AC189314.1	*B.rapa*	413	TTA	*GGAGGGTGTTAGTGGGA*	76	MITE/Tourist
BrTOUR1-5	AC232479.1	*B.rapa*	412	TTA	*GGGGGTGTTAGTGGGGA*	74	MITE/Tourist
BrTOUR1-6	AC189261.2	*B.rapa*	412	TTA	*GGGGGTGTTAGTAGGGA*	74	MITE/Tourist
BrTOUR1-7	AC189219.1	*B.rapa*	412	TTA	*GGGGGTGTTAGTGGG*	75	MITE/Tourist
BnTOUR1-8	AC236791.1	*B.napus*	412	TAA	*GGGGGTGTTAGTGAGGA*	74	MITE/Tourist
BrTOUR1-9	AC189415.2	*B.rapa*	402	TAA	*GGGGGTGTTAGTGGG*	74	MITE/Tourist
BrTOUR1-10	AC189397.2	*B.rapa*	392	TTA	*TGGGATATGGATTTGTAGTGA*	75	MITE/Tourist
BrTOUR2-1	AC155344.1	*B.rapa*	285	TAA	*GAGACACCCCCATTAGTGAAC*	63	MITE/Tourist
BrTOUR2-2	AC172859.1	*B.rapa*	289	TTA	*GAGCATCCCCATTAGTGAAC*	62	MITE/Tourist
BrTOUR2-3	AC189450.2	*B.rapa*	287	TAA	*GAGCACCCCCATTAGTGAAC*	62	MITE/Tourist
BrTOUR2-4	AC189577.2	*B.rapa*	284	TAA	*GAGCACCCCATTAGTAAAC*	64	MITE/Tourist
BrTOUR2-5	AC232550.1	*B.rapa*	273	TTA	*GAGCACCCCCATTAGTGAAC*	65	MITE/Tourist
BrTOUR3-1	AC189298.1	*B.rapa*	258	TTA	*GGACATCTCCA---(105)*	67	MITE/Tourist
BoTOUR3-2	DQ222849.1	*B.oleracea*	258	TAA	*GGACATCTCCA---(106)*	66	MITE/Tourist
BoTOUR3-3	DQ222850.1	*B.oleracea*	258	TTA	*GAGCATCTCCA---(106)*	66	MITE/Tourist
BnTOUR3-4	FJ384103.1	*B.napus*	258	TAA	*GAGCATCTCCA---(102)*	66	MITE/Tourist
BrTOUR3-5	AC189458.2	*B.rapa*	258	TTA	*GAGCATCTCCA---(102)*	67	MITE/Tourist

Element Name.	BAC Accession	Species	Size	TSDs	Terminal inverted repeats (TIRs)	AT %	MITE Superfamily
BrTOUR3-6	AC172875.2	*B.rapa*	258	TTA	*GAGCATCTCCA---(102)*	66	MITE/Tourist
BrTOUR3-7	AC189299.2	*B.rapa*	258	TTA	*GGGCATCTCCA---(101)*	64	MITE/Tourist
BrTOUR3-8	AC189445.2	*B.rapa*	258	TTA	*GGGCATCTCCA---(103)*	67	MITE/Tourist
BrTOUR3-9	AC189370.2	*B.rapa*	258	TAA	*GAGCATCTCCA---(102)*	66	MITE/Tourist
BrTOUR3-10	AC155339.1	*B.rapa*	259	TTA	*GAGCATCTCCA---(102)*	64	MITE/Tourist
BoTOUR4-1	AC240081.1	*B.oleracea*	267	TAA*	*TACTCACTCTGTTTCATAAATGTCATTCTAACTTTTTT*	76	MITE/Tourist
BrTOUR4-2	AC189192.2	*B.rapa*	332	TTA	*CTCCCTCTTCGTAATTAATTACT*	77	MITE/Tourist
BrTOUR4-3	AC241150.1	*B.rapa*	272	TAA*	*TATACTCTCTCTATTTTATAATAAGTGTCA*	79	MITE/Tourist
BnTOUR4-4	AF136223.1	*B.napus*	272	TAA	*TACTCCATCTGTTTCATATTAAGTGTCATTGTAACA*	79	MITE/Tourist
BrTOUR4-5	AC232552.1	*B.rapa*	272	TTA	*CTACTCCTTCCGTTTCTGAATAAGTGTCATTTT*	78	MITE/Tourist
BrTOUR4-6	AC189299.2	*B.rapa*	271	TAA*	*TACCCTCTCCATTTCTGAATAACTGTCA*	75	MITE/Tourist
BrTOUR4-7	AC189587.2	*B.rapa*	266	TCA*	*TACTCCTTCCGTTTCTAAATAACTGTCA*	81	MITE/Tourist
BrTOUR4-8	AC189218.2	*B.rapa*	264	TAA	*TACTCTTTCTGTTTCTAAATAAATATCACTTTGAAGTTTTT*	79	MITE/Tourist
BrTOUR4-9	AC189322.2	*B.rapa*	261	TTA	*TACTTCCTCCGTTTCATAAAAAATGTCACT*	80	MITE/Tourist
BrTOUR4-10	AC189569.2	*B.rapa*	255	TAA	*TACTCTCTATATTTTTGAAAAAAATATCATTTT*	81	MITE/Tourist
BrMuMITE1-1	CU984545.1	*B.rapa*	551	TATCC TATT	*122/125*	78	MITE/Mutator
BrMuMITE1-2	AC189475.2	*B.rapa*	569	tTTAA TGAA	ND	78	MITE/Mutator
BrMuMITE1-3	AC189340.1	*B.rapa*	559	TAAAA TGAt	ND	78	MITE/Mutator
BrMuMITE1-4	AC232437.1	*B.rapa*	547	TTTAC ATAA	ND	77	MITE/Mutator
BnMuMITE1-5	AC236785.1	*B.napus*	527	TATTT aTTaT	ND	77	MITE/Mutator
BrMuMITE2-1	CU984545.1	*B.rapa*	905	CTTTA GAAAC	*427/435*	81	MITE/Mutator
BrMuMITE2-2	AC189218.2	*B.rapa*	1060	TTATT TAAAT	ND	80	MITE/Mutator
BrMuMITE2-3	AC189224.1	*B.rapa*	1055	TATTT TATTG	ND	80	MITE/Mutator
BrMuMITE2-4	AC189578.2	*B.rapa*	1052	AACAA TATAG	ND	80	MITE/Mutator
BrMuMITE2-5	AC155345.1	*B.rapa*	958	TAAAA CTGTG	ND	81	MITE/Mutator
BrMuMITE3-1	AC232530.1	*B.rapa*	1586	CAAAA AAAAc	*717/689*	77	MITE/Mutator
BrMuMITE3-2	AC189366.2	*B.rapa*	1624	AATAA AATAT	ND	78	MITE/Mutator
BrMuMITE3-3	AC232539.1	*B.rapa*	1575	CATAA TAATT	ND	77	MITE/Mutator
BrMuMITE3-4	AC189401.2	*B.rapa*	1555	GATTT AATAT	ND	77	MITE/Mutator
BrMuMITE3-5	AC189580.2	*B.rapa*	1497	TAAAA AGAAC	ND	79	MITE/Mutator
BrMuMITE3-6	AC232458.1	*B.rapa*	1581	GATTT TCAAG	ND	77	MITE/Mutator
BrMuMITE3-7	AC232562.1	*B.rapa*	1552	AAAAC AAAAC	ND	77	MITE/Mutator
BoMuMITE3-8	EU642504.1	*B.oleracea*	1539	GATTA GATTC	*649/616*	78	MITE/Mutator
BoMuMITE3-9	EU579455.1	*B.oleracea*	886	TTAAA TgTT	*255/243*	78	MITE/Mutator
BoMuMITE4-1	AC149635.1	*B.oleracea*	899	TATAT ATAT	*407/446*	73	MITE/Mutator
BoMuMITE4-2	EU579455.1	*B.oleracea*	766	TTGGa TtGT	*358/351*	77	MITE/Mutator
BrMuMITE4-3	AC172877.1	*B.rapa*	886	CTAAA ATTA	ND	75	MITE/Mutator
BrMuMITE4-4	AC232459.1	*B.rapa*	839	ATTTT TCTTT	ND	75	MITE/Mutator
BrMuMITE4-5	AB257127.1	*B.rapa*	820	TTTTT TTAA	ND	77	MITE/Mutator
BrMuMITE5-1	AC155344.1	*B.rapa*	1159	TTTAT Taga	*354/349*	58	MITE/Mutator
BrMuMITE5-2	AC172882.1	*B.rapa*	1157	TTATT aga	*354/348*	58	MITE/Mutator
BrMuMITE5-3	AENI01009313.1	*B.rapa*	1164	AAgAG AAAT	*353/357*	58	MITE/Mutator
BoXMITE1-1	EU642504.1	*B.oleracea*	402	TTC	*GGCCATGTTCGTTTACGTGTCGCGCGACCTACGACCTGCGAC*	48	MITE/ND
BrXMITE1-2	AENI01000925.1	*B.rapa*	356	CTC*	*TTCATTTACGTATCGCGCGACCTGCGACCTG*	52	MITE/ND
BrXMITE1-3	AC189543.2	*B.rapa*	320	AAT*	*GGCCTGTTCCTTACCTGTCTGGC*	54	MITE/ND
BrXMITE1-4	AENI01003669.1	*B.rapa*	320	AAT*	*GGCCTGTTCCTTACCTGTCT*	54	MITE/ND
BrXMITE1-5	AENI01006359.1	*B.rapa*	308	CAT*	*TCGTTTACGTATCGTGCGACCTGCGACT*	58	MITE/ND

7.2.3 Structural features of *Brassica* Stowaway-like MITEs

Five novel families of Stowaway-like MITEs were detected in *Brassica* genomes. The elements range in sizes from 218 bp (*BrSTOW4-10*) to 580 bp (*BrSTOW1-1*). The average size of the elements range from 230-260 bp. *BrSTOW1-1* is the only member exceeding 500 bp, while its homologues range in sizes from 324-329 bp. The Stowaway-like elements show an insertional preference in AT rich regions and are terminated by TA TSDs. The first family is designated as *BrSTOW1* due to the identification of its first member from *Brassica rapa* accession (AC155344.1) from 31584 to 32161 bp. The element is named *BrSTOW1-1* and display the longest (30 bp) imperfect TIRs. The BLASTN hits retrieved ~52 full length or truncated copies. The total estimated copies from *BrSTOW1* family in *Brassica rapa* and *Brassica oleracea* are 990. The TIRs of the family members ranges between 15-30 bp with average TIRs of 18-20 bp (Figure 7.2; table 7.1). *BoSTOW2* family represents a low copy number family with an estimation of only 382 copies from *Brassica* genomes. The first member of this family (*BoSTOW2-1*) was identified from *Brassica oleracea* accession 'AC2400081', while its other homologue (*BoSTOW2-2*) was observed in another *Brassica oleracea* accession 'EU579455.1'. The elements range in sizes from 436-460 bp, terminated by TA TSDs and 11 bp terminal inverted repeats. *BoSTOW3* family is low copy number family with 230 estimated copies from *Brassica* genomes. The family members range in sizes from 237-244 bp and terminated by TA TSD and terminal inverted repeats of 13-15 bp with conserved 5′-GAG-3′ termini. This family showed strong hits against *Brassica oleracea* sequences only suggesting its proliferation in C-genome (Figure 7.2; Table 7.1).

BoSTOW4 family represents a diverse group of MITEs ranging in sizes from 218-256 bp with average sizes of 251 bp. The representative of the family has 2 bp TSDs and conserved 27-34 bp TIRs with highly conserved 5′-CTC-3′ termini. The TIRs are highly homologous among the member of the elements with the slight variations of one to few bp. BLASTN hits using *BrSTOW4-1* as query sequences yielded 201 sequences as output, of which 120 were considered as full length or truncated elements. Approximately 2294 total copy numbers from *Brassica rapa* and *Brassica oleracea* whole genomes is estimated, making it a high copy number family. The fifth family *BoSTOW5* is characterized by members having TA TSDs and TIRs of 19-21 bp, which are highly conserved among the members of the families. Like other Stowaway investigated in

present study, the TIRs have shown conserved 5′-CTC-3′ termini. The elements range in sizes from 239 (*BrSTOW5-8*) to 245 (*BrSTOW*5-10) bp. The first identified member of the family is *BoSTOW5-1*, which was detected as an insertion residing in *Brassica oleracea* accession (AC240081.1) from 32696-32938 bp. Using this as query sequence in BLASTN searches, 309 hits were returned, out of which 170 were considered full elements or truncated elements. The total copy number estimation revealed that ~3239 copies are actively proliferating in *Brassica rapa* and *Brassica oleracea* genomes.

Table 7.2: List of estimated copy numbers and AT% of *Brassica* MITEs families. The average lengths of elements and their average AT% are given.

Family	TSDs	Length	No. in database	No. in genomes	Average AT%
BrSTOW1-1	TA	324-580	52	990	70
BoSTOW2-1	TA	422-460	20	382	71
BoSTOW3-1	TA	237 -244	12	230	62
BoSTOW4-1	TA	218-256	120	2294	80
BoSTOW5-1	TA	239-245	170	3239	76
BrTOUR1-1	TNA	392-421	85	1624	75
BrTOUR2-1	TNA	273-289	64	1224	63
BrTOUR3-1	TNA	258 -259	205	3918	66
BoTOUR4-1	TNA	255-332	128	2446	78
BrMuMITE1	9 bp	527-569	256	4892	78
BrMuMITE2	10 bp	905-1060	22	420	80
BrMuMITE3	10 bp	1497-1624	28	535	78
BoMuMITE4	9 bp	766-899	312	5964	75
BrMuMITE5	9 bp	1152-1164	6	114	58
BoXMITE1	TTC	308-402	12	1229	53

7.2.3.1 Transposon Insertional polymorphism of Stowaway-like MITEs

To investigate the insertion polymorphisms of *Brassica* Stowaway-like MITEs among 40 *Brassica* lines, three primer pairs were designed against both flanking sequences of each MITE (Table 7.3). Among the 3 primer pairs tested for three Stowaway MITE families, 63 bands with their respective product sizes (MITEs) were amplified from 40 *Brassica* lines. The lower band amplified the flanking regions, which are derived from homeologous regions in the A and C-genomes. The BrSTOW1F and BrSTOW1R primer pair (Table 7.3) was used to amplify the 580 bp *BrSTOW1* MITEs. The expected product size of 682 bp was amplified from *Brassica rapa* (Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons) lines. In the first three lines both upper and lower bands were observed suggesting their heterozygous nature. No upper or lower band was amplified in any of the

three *Brassica nigra* (BB) genome suggesting the B-genome difference against A and C-genomes. No bands were observed in C-genome with the exception of *Brassica oleracea gemmifera* 'De Rosny', where a upper band was amplified only. A very light band of ~550 bp was observed in most of the *Brassica oleracea* lines. The inserional polymorphism in *Brassica juncea* displays the amplification of *BrSTOW1* in *Brassica juncea* 'Megarrhiza' and *Brassica juncea* 'W3' lines while the other four lines only produced the lower band indicating the absence of *BrSTOW1* in these genomes. All the six allotetraploid *Brassica napus* lines (New, Mar, Last and Best, Fortune, Drakker, Tapidor) produced the upper band amplifying the *BrSTOW1*. No *BrSTOW1* amplification observed in any of *Brassica carinata* genomes. All the three except one *Brassica* hexaploid produced the expected product amplifying the *BrSTOW1* in their genomes (Figure 7.3a). This suggests the availability and distribution of *BrSTOW1* in *Brassica rapa* and its corresponding allotetraploids and hexaploids.

The amplification pattern of 237-244 bp *BrSTOW3* was performed by using degenerative primers pair BoSTOW3F and BoSTOW3R (Table 7.3). The primers were designed from flanking regions of MITE shared by both A and C-genomes with a expected product size of 512 bp. Out of 40 *Brassica* lines, 22 produced the higher bands amplifying the *BoSTOW3* insertional loci, while no amplification was observed in other 18 lines, which suggest that MITE *BoSTOW3* is not inserted in this specific locus but might be present on other loci within these lines. All the six *Brassica oleracea* (de Rosny, Kai Lan, Early Snowball, Precoce Di Calabria Tipo Esportazione, Cuor Di Bue Grosso, GK97361) lines produced the larger band only amplifying the MITE *BoSTOW3*. Similarly all the six lines each from allotetraploid *Brassica napus* (AACC) and *Brassica carinata* (BBCC) and 4 hexaploid *Brassicas* (AABBCC) yielded the product of 512 bp amplifying the *BoSTOW3* in their genomes. All these lines also produced the lower band of ~260 bp (Figure 7.3b).

BoSTOW4 insertional polymorphism among the 40 *Brassica* lines gave nearly similar results as observed in the amplification of *BoSTOW3*. Out of 40 *Brassica* lines, 23 amplified the 500 bp *BoSTOW3* insertional loci, while no amplification was observed in other 17 lines. One *Brassica juncea* (NARC-II) amplified the 500 bp band. The *BoSTOW4* insertions were amplified from six *Brassica oleracea,* six *Brassica napus,* six *Brassica carinata* and four 6x *Brassica* lines. All these lines also produced the lower band of ~270 bp suggesting the heterozygous nature of *Brassicas* (Figure 7.3c).

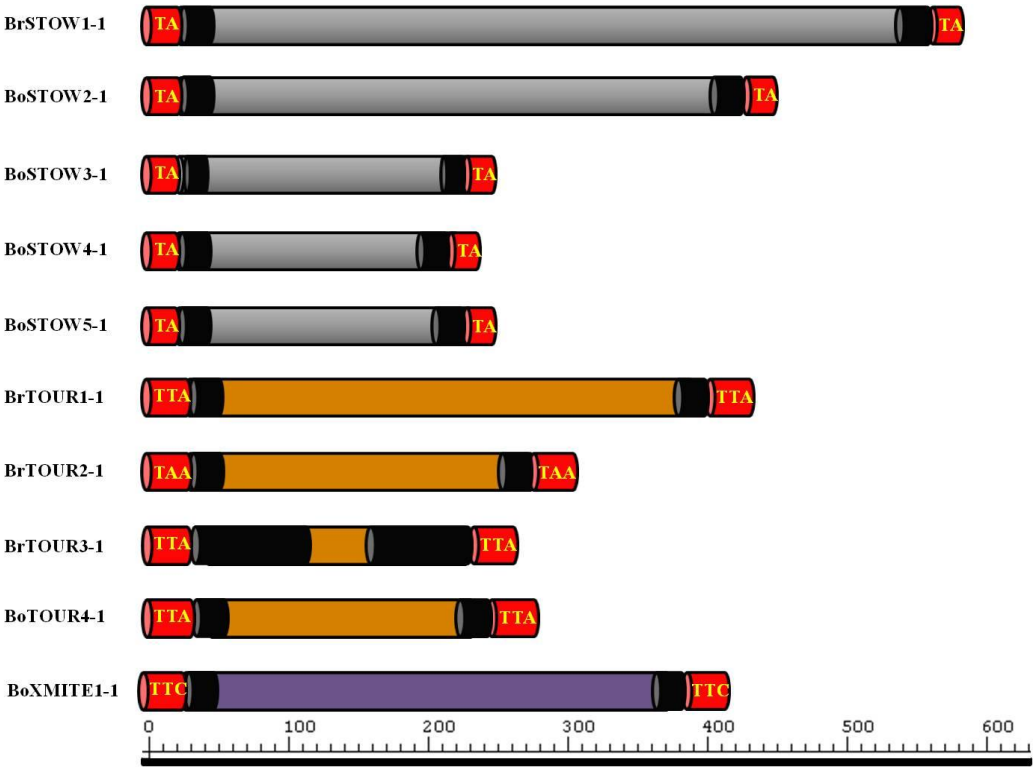


Figure 7.2: Schematic representation of Stowaway, Tourist and novel MITEs in *Brassica***.** Red disks represent 1-3 bp TSDs while black disks of varied sizes represent the TIRs of MITEs. The coloured regions among TIRs indicate the internal non-coding regions. The TIRs and internal regions of all MITEs are AT rich. Scale at base shows sizes in bp.

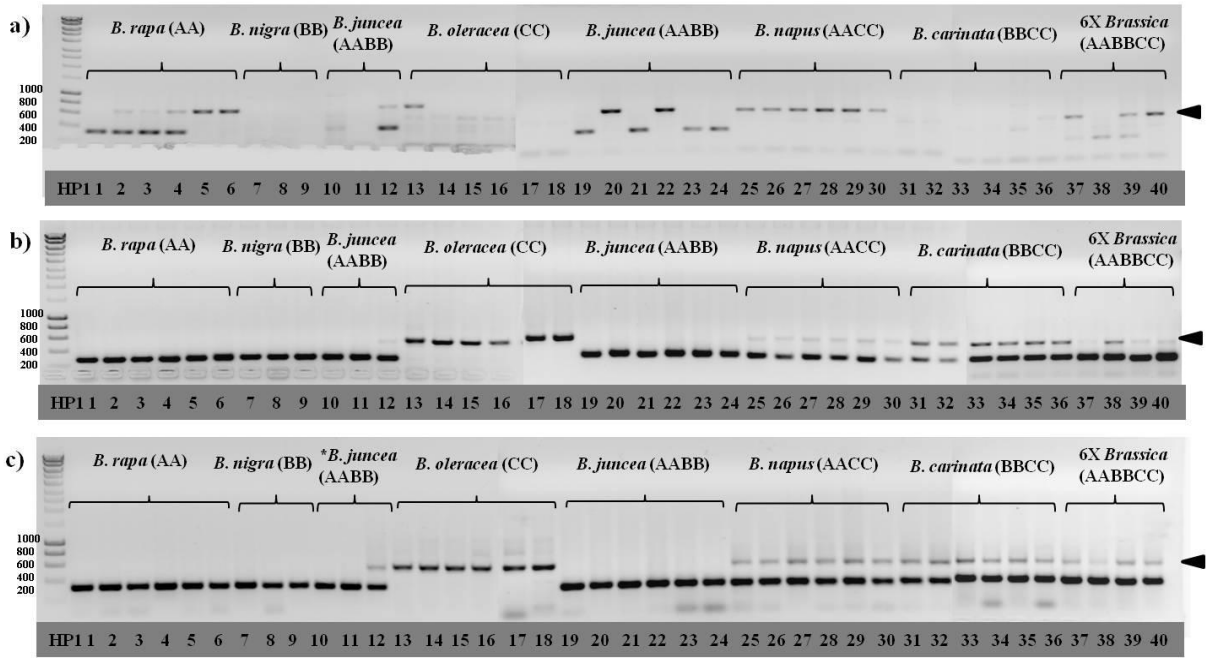


Figure 7.3: Insertional polymorphism of *Brassica* Stowaway-like MITEs. a) *BrSTOW1*; b) *BoSTOW3*; c) *BoSTOW4*. Black arrowheads (right) indicate upper bands with amplified loci having MITE insertions while lower bands are amplified from loci without insertions (see Fig. 7.6). PCR figures show reversed images of size-separated ethidium bromide-stained DNA on agarose gels after electrophoresis; ladders show fragments sizes in base pairs; numbers at the base indicate accessions of the species indicated from Table 2.1.

Table 7.3: List of *Brassica* MITEs primers with size of the elements, size of the expected products, names and sequence of primers.

Sr. No.	Superfamily	Family	Element Size	Product size	Primers	Primer sequence
1	Stowaway	*BrSTOW1*	324-580	682	BrSTOW1F BrSTOW1R	CTTCGTATTCTCTGCAAGAT CGAAATACATAGACGTATAC
2	Stowaway	*BoSTOW3*	237-244	512	BoSTOW3F BoSTOW3R	AGGGTCCAAACATGTGATTA GTTGCAAATAATTGATCGTTG
3	Stowaway	*BoSTOW4*	227	500	BoSTOW4F BoSTOW4R	CAATACCATCCAGTGTTACA TGTTGTCGTCATTAAGGTGA
4	Tourist	*BrTOUR1*	392-421	530	BrTOUR1F BrTOUR1R	GGGGATAATTACACATCTTG CAAATCTCCGACATCAATC
5	Tourist	*BrTOUR2*	273-289	510	BrTOUR2F BrTOUR2R	AGGGTCCAAACATGTGATTA GTTGCAAATAATTGATCGTTG
6	Tourist	*BrTOUR3*	258	564	BrTOUR3F BrTOUR3R	GGACCATACAGTATATCGTT TGGATAACGTTGTTGTTCCC
7	Mutator-like	*BrMuMITE1*	527-569	1016	BrMuMITE1F BrMuMITE1R	CATTGCAGAAGAGCTGGCTGC CAAGATTTTGAGGAGAGATTTG
8	Mutator-like	*BrMuMITE4*	766-899	990	BrMuMITE4F BrMuMITE4R	GATAATTTTTGGGCCATGCA CGATCAGACAAACGACGAAA

7.2.4 Characterization of *Brassica* Tourist-like MITEs

The Tourist-like MITEs are the derivatives of PIF/Harbinger DNA transposons with TAA/TTA TSDs and small TIRs. The Tourist MITEs are actively proliferating in *Brassica* genomes. Four families of Tourist (*BrTOUR1, BrTOUR2, BrTOUR3* and *BoTOUR4*) were detected in *Brassica* genomes. Approximately, 9212 Tourist-like full or nearly full length (truncated) copies was estimated from *Brassica* (A and C) genomes. The highest copy number Tourist family *BrTOUR3* display ~3918 sequences followed by ~2446 copies in *BoTOUR4*. The TIRs generally ranges from 11-21 bp with the exception of *BoTOUR3* family, where 102-106 bp TIRs were observed flanking by TAA target site duplications. The TIRs are highly conserved within the members of the same family but different from the members of the other families. The TIRs in all the elements are GC rich, while the internal non-coding regions are highly AT rich (Table 7.1).

BrTOUR1 is a family of Tourist-like MITEs ranging in sizes from 392 to 421 bp with TTA/TAA TSDs and TIRs of 13-21 bp. The 5′ termini of TIRs are highly conserved with 5′ GGGGG-3′ termini (Figure 7.4). The first identified element of the family was *BrTOUR1-1*, inserted within *Brassica rapa* accession 'AC155344.1' from 43736-44148 bp. The element is 413 bp with 3 bp TSDs (TTA) and 16 bp TIR. The BLASTN hits using this element as query yielded 85 full length elements, which have shown high homology in their entire lengths. The largest element identified from the family is 421 bp *BrTOUR1-*

2, while *BrTOUR1-10* is only 392 bp in length. The average sizes of the elements in this family are 412-413 bp. *BrTOUR2* family proliferating in *Brassica* genomes exhibit ~1224 copies (Table 7.2). The first element discovered from the family was a 285 bp insertion in *Brassica rapa* accession 'AC155344.1' residing from 42921-43204 bp. The other homologues of this family range in sizes from 273-289 bp with 3 bp TSDs and TIRs of 19-21 bp. *BrTOUR2-5* is 273 bp in size and display 20 bp TIRs flanked by TTA target repeats. *BrTOUR3* is high copy number family with ~3918 copies distributed among *Brassica* genomes. All the members of this family exhibit the same size (258-259 bp) with large TIRs of ~102-106 bp. There are insertions of 5 and 7 bp at 11 bp downstream and 11 bp upstream of 5′ and 3′ terminal ends respectively. Like the members of *BrTOUR2*, *BrTOUR3* family members also generate 5′-CATCTCC-3′ conserved termini in their TIRs (Figure 7.4). The members of the family are distributed among *Brassica rapa*, *Brassica oleracea* and *Brassica napus* genomes. *BnTOUR3-4* is a Tourist-like MITE identified from *Brassica napus* accession 'FJ384103.1' (Table 7.1).

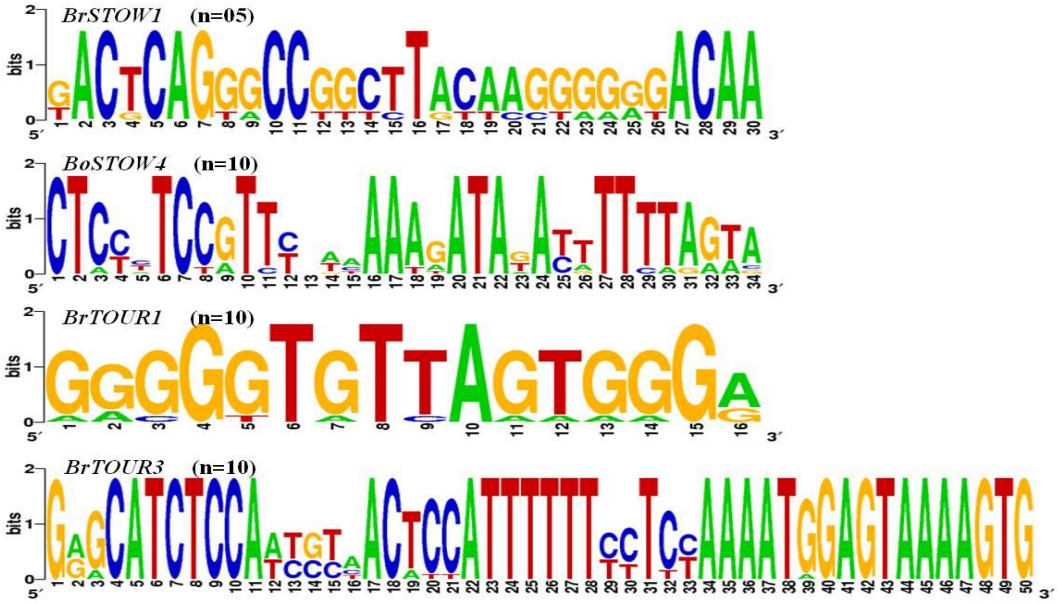


Figure 7.4: Sequence logos (pictograms) of *Brassica* MITE TIRs**.** The logos generated with (n) sequences and letter heights (0 to 2) indicate the information content of nucleotides in the TIRs of a,b) *Brassica* Stowaway (*BrSTOW1*, *BoSTOW4*) and c,d) Tourist-like (*BrTOUR1*, *BrTOUR3*) MITE families. The short height nucleotides represent non-conserved motif or insertion. The 5 bp insertion in *BrTOUR3* (12-16 bp) represents a 5 bp non-homologous sequence or an insertion.

BoTOUR4 is a family of elements distributed among various *Brassica* genomes. The first element (*BoTOUR4-1*) was identified from *Brassica oleracea* accession 'AC240081.1' as an insertion present in this BAC from 32202-32468 bp. *BoTOUR4-1* is a 267 bp element with 3 bp imperfect TSDs and 38 bp imperfect TIRs (Figure 7.2). Using this element as query sequence, several other homologues were collected ranging in sizes from 255-332 bp. Most of the retrieved sequences were from *Brassica rapa*, which might be is the result of high genome size deposition at GenBank database. Approximately, 2446 copies are estimated from *Brassica rapa* and *Brassica oleracea* genomes making it a high copy number family of MITEs in *Brassica* (Table 7.2). The TIRs of the elements rang is sizes from ~23-41 bp, which have shown high similarity among the members of the family with 5′-TACTC-3′ like conserved termini. *BrTOUR4-2* with a size of 332 bp is the largest member of the family with small insertions in it. The average sizes of the elements are ~272 bp, whereas *BrTOUR4-10* is the smallest MITE from this family (Table 7.1).

7.2.4.1 *Brassica* Tourist MITE insertion polymorphism among *Brassica* cultivars

To study the insertional polymorphism among various *Brassica* cultivars/lines (40), the primers were designed from the flanking regions of the MITEs. Among the primer pair BrTOUR1F and BrTOUR1R tested for the amplification of *BrTOUR1* MITE, 15 generated the expected product with MITE insertion while rest 25 yielded the lower bands without insertions (Figure 7.5a). The BLASTN results have shown the distribution of *BrTOUR1* among various *Brassica* species, mostly in *Brassica rapa*. This primer set only amplified *BrTOUR1* MITEs loci from A-genome accession 'Chinese Wong Bok', which produced both higher and lower bands. No PCR amplification of *BrTOUR1* was seen in other A and B-genome species. Out of the six diploid C-genomes, 5 *Brassica oleracea* accessions (Kai Lan, Early Snowball, Precoce Di Calabria Tipo Esportazione, Cuor Di Bue Grosso, GK97361) amplified the band while no band was observed in *Brassica oleracea* 'De Rosny'. Five out of 6 *Brassica napus* genomes (New, Last and Best, Fortune, Drakker, Tapidor) amplified the respective bands (~530 bp). Similarly 3 *Brassica* hexaploid lines amplified the *BrTOUR1* loci. The insertional polymorphism of 237-289 bp *BrTOUR2* among 40 *Brassica* accessions gave various polymorphic bands (Figure 7.5b). The primers BrTOUR2F and BrTOUR2R were used to amplify 510 bp products having MITE insertions and flanking sequences. Out of 3 diploids species (AA, BB, CC), the MITEs insertion loci was only amplified in A-genomes including *Brassica rapa*

accessions (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons). No amplification of these loci with *BrTOUR2* insertion was seen in diploids B and C-genomes. All the 9 lines among *Brassica juncea* (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna), 6 lines from *Brassica napus* (New, Mar, Last and Best, Fortune, Drakker, Tapidor) and 4 lines from hexaploid *Brassicas* yielded the MITE inserted loci.

A 258 bp element *BrTOUR3* was amplified from various *Brassica* lines by using BrTOUR3F and BrTOUR3R primers pair (Table 7.3). The upper bands (~564 bp) showed the amplification of *BrTOUR3* MITE, while lower ~300 bp bands amplified flanking regions without the insertion (Figure 7.5c). Of the 3 diploids species tested, only A-genome *Brassica rapa* accesions (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Suttons) yielded the expected band except *Brassica rapa* (Vertus). All the six A-genomes also produced the lower band of ~300 bp. No *BrTOUR3* MITE containing loci amplified in B and C-genomes, where only lower band with flanking regions are yielded. Among 9 *Brassica juncea* lines tested, only 6 accessions showed expected bands while 3 have shown only lower bands amplifying the flanking regions. All the lines tested from *Brassica napus*, *Brassica carinata* and hexaploid *Brassicas* produced the strong lower bands suggesting the presence of flanking regions without MITE insertions.

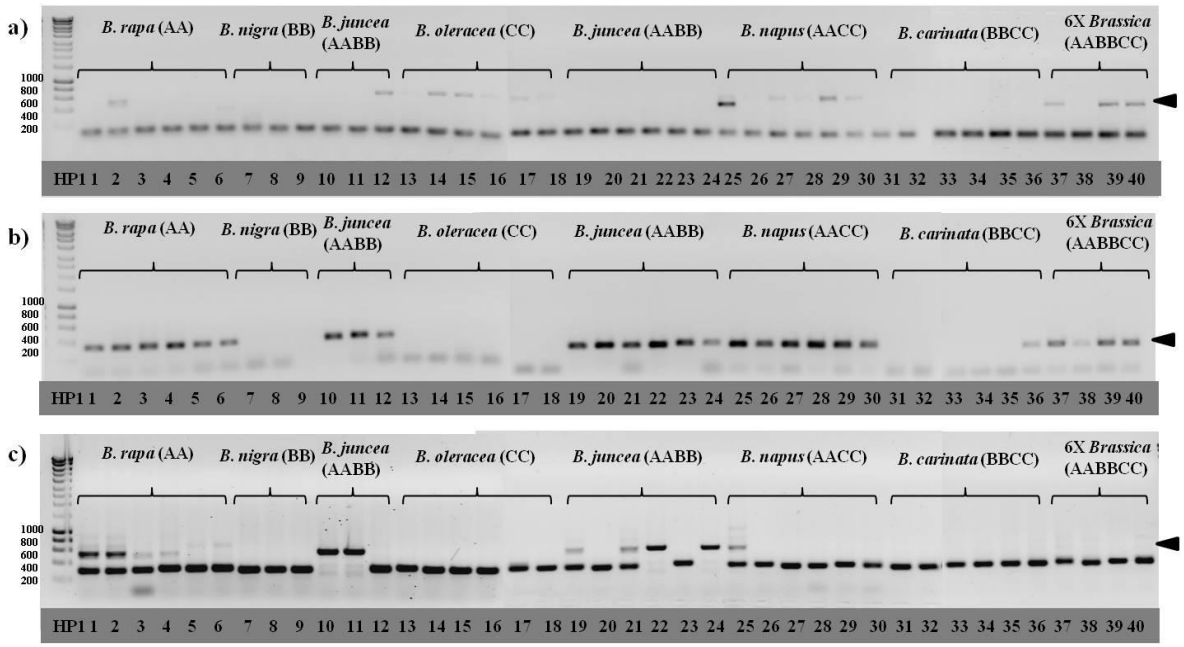


Figure 7.5: Insertional polymorphism of *Brassica* Tourist-like MITEs. a) *BrTOUR1*; b) *BrTOUR2*; c) *BrTOUR3*. The black arrowheads (right) indicate upper bands from loci having MITE insertions (eg ~270 bp in c) while lower bands are amplified from loci without MITE insertion.

7.2.5 Molecular characterization of Mutator-like MITEs in *Brassica*

The deletion of transposase and internal regions from Mutator DNA transposons led to the formation of Mutator-like MITEs. The present study explained the high diversity, distribution and mobilization of Mutator MITEs among various *Brassica* species. Five Mutator-like MITE families were identified designated as *BrMuMITE1, BrMuMITE2, BrMuMITE3, BoMuMITE4,* and *BrMuMITE5*. These MITEs range in sizes from 527-1624 bp with 9-10 bp TSDs. The TIRs among the members of the various families range from ~100-750 bp with varied lengths in 5′ and 3′ TIRs due to insertions/deletions. Approximately 12536 copies of Mutator-like MITEs were estimated from *Brassica rapa* and *Brassica oleracea* whole genomes (Table 7.1 & 7.2). Using the first identified Mutator derived MITEs (reference) as query, other complete elements were collected for each family. Due to very high copy numbers of these MITEs, only five hits with maximum homology were collected and studied in detail.

BrMuMITE1 is a family of Mutator derived MITE, highly distributed among *Brassica* crops. This is a high copy numbers family with an estimated 4892 copies proliferating in A and C-genome *Brassicas*. The first element identified from this family was a 551 bp sequence, generates perfect 9 bp TSDs (TATCCTATT) and TIRs of 122/125 bp. Using this sequence as query, many other copies were collected and characterized from this family. The elements of the family range in sizes from 527-569 bp with 9 bp target repeats with the exception of *Brassica napus* MITE (*BnMuMITE1-5*; 10 bp TSDs). The TSDs of the representatives of family are AT rich and the TIRs have shown high homology with each other. *BrMuMITE2*, a low copy number (~420 estimated copies) family is distributed among *Brassica* genomes with members exhibiting ~905-1060 bp lengths, flanked by 10 bp TSDs and high AT content. TIRs of the various elements are ~400-450 bp in sizes with high homology within the members of the family. *BrMuMITE2-2* is the largest member (1060 bp), while 958 bp *BrMuMITE2-5* is representing the smaller member with high AT content (81%). *BrMuMITE2-3* and *BrMuMITE2-4* are 1055 and 1052 bp elements respectively including 10 bp TSDs (Figure 7.6; Table 7.1 & 7.2).

BrMuMITE3 is a middle copy number family with ~611 estimated copies within *Brassica* genomes (A and C). The representatives of the family range in sizes from 886-1624 bp including 9-10 bp TSDs. The TIRs ranges from 250-750 bp with high homology within

the members of the family in their entire regions. A 886 bp MITE named *BoMuMITE3-9*, was found inserted in *Brassica oleracea* (EU579455.1) from 1113-1983 bp. In a parallel investigation, another similar element (*BrMuMITE3-1*) but larger in size (1586 bp) was identified from *Brassica rapa* (AC232530.1) from 47143-48728 bp. Both sequences showed high homology and suggested the members of the same family. Using these two sequences as query in GenBank database, 7 other full length elements were collected. The largest element was *BrMuMITE3-2*, which is 1624 bp large including 10 bp TSDs. Most of these elements are >1500 bp with long TIRs and very small internal non-coding regions. *BoMuMITE3-8* is a 1539 bp MITE including 10 bp TSDs and 5´-649/616-3´ bp TIRs residing from 38719-40252 bp (Figure 7.6; Table 7.1).

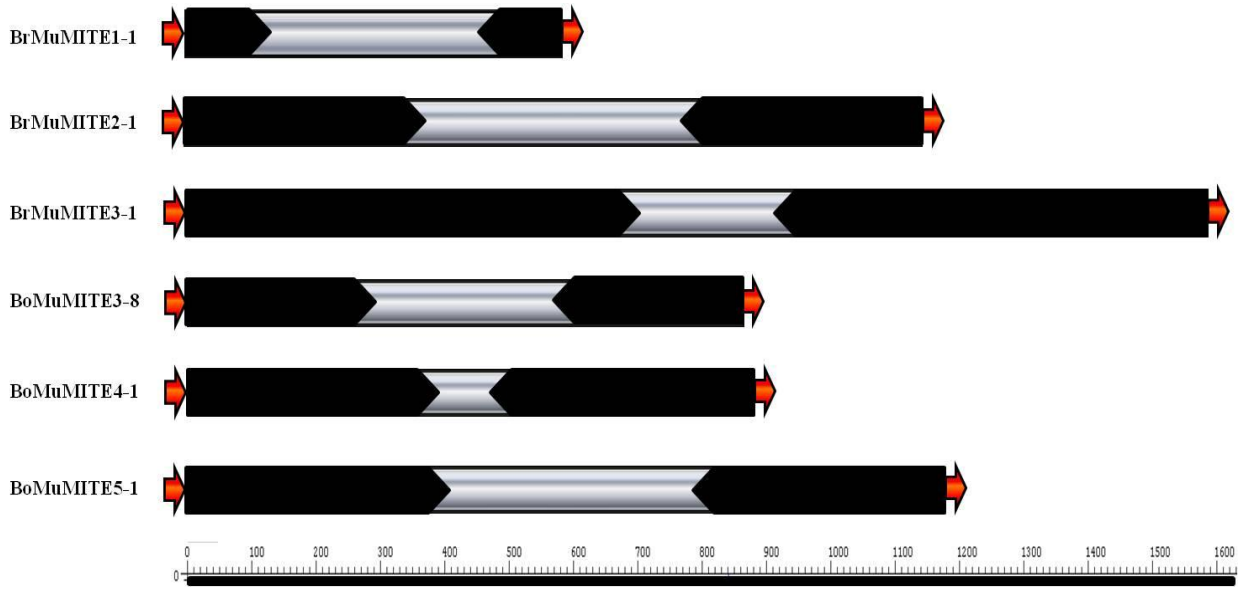


Figure 7.6: Structures of Mutator-like MITEs in *Brassica*. The red arrows represent 9-10 bp TSDs while black filled pentagon of varied lengths represents the long TIRs of Mutator-like MITEs. The light coloured regions between the inverted pentagons indicate the internal non-coding regions. The TIRs and internal regions of all MITEs are AT rich. Scale at base shows sizes in bp.

BoMuMITE4, a family comprising the highest copy numbers (5964) was identified with members distributed among various *Brassica* crops. Using *BoMuMITE4-1* as query returned 2400 hits, of which ~312 full length elements were extrapolated. The first identified element *BoMuMITE4-2* was 766 bp in size with 9 bp imperfect TSDs and TIRs 5´-358/351-3´. Another 899 bp MITE was detected from *Brassica oleracea* accession 'AC149635.1' including 9 bp TSDs and 5´-407/446-3´ bp TIRs. The elements of the family range in sizes from 766-899 bp with 9 bp TSDs and large TIRs (~350-446). *BrMuMITE4-3*, *BrMuMITE4-4* and *BrMuMITE4-5* are *Brassica rapa* MITEs with 9 bp

TSDs and large TIRs with very high AT rich regions. The *BrMuMITE5* represents a low copy numbers family (114 estimated copies), where only 3 copies were collected from *Brassica* Nucleotide Collection and Whole-genome shotgun contigs (wgs) databases of NCBI. The elements range in sizes from 1152-1167 bp with 9 bp imperfect TSDs. The elements have shown high AT content (58%) but comparatively very low as compared to other MITE families investigated in *Brassica*. The first element identified was *BrMuMITE5-1*, which is 1152 bp including 9 bp TSDs and 5′-354/349-3′ bp TIRs. The other two elements *BrMuMITE5-2* and *BrMuMITE5-3* are of the same sizes and showed very high homology with each other (Figure 7.6; Table 7.1).

7.2.5.1 Mutator derived MITEs; Insertional polymorphism among *Brassica*

The transposon insertional polymorphism assay (TIP) was used to detect the abundance of Mutator MITEs in a diverse set of 40 *Brassica* accessions/lines from three diploids (AA, BB, CC), three allotetraploids (AABB, AACC, BBCC) and 2 resynthesized hexaploid *Brassicas* (*B. napus* x *B. nigra*; *B. carinata* x *B. rapa*). PCR oligonuclotide primers flanking a respective MITE yielded a higher and a lower band amplifying with and without MITE Insertional sites respectively. Of the 5 Mutator derived MITE families, two (*BrMuMITE1*, *BoMuMITE4*) have shown high diversity and abundance among *Brassica* genomes. *BrMuMITE1* have shown high diversity and distribution among *Brassica rapa* lines, while *BoMuMITE4* have shown high distribution in C-genome specific lines. To confirm this, degenerative primer pairs were designed from the flanking regions common in A and C-genomes. Using BrMuMITE1F and BrMuMITE1R primers (Table 7.3), 1016 bp product with 551 bp *BrMuMITE1* insertions were amplified from 23 out of 38 *Brassica* lines tested. The MITE amplication polymorphisms displayed that 5 *Brassica rapa* accessions (Pak Choy, Chinese Wong Bok, Hinona, Vertus, Suttons) yielded the ~1016 bp product while accession ’San Yue Man′ failed to generate the expected product but yielded >3 kb band of unknown nature. No amplification from B and C-genomes except *Brassica oleracea* ‘Early Snowball’, suggested the absence of the *BrMuMITE1* in comparisons to A-genome, where clear bands amplifying the MITEs were observed. Amongst the classical allotetraploid *Brassica* species, all the six lines from each *Brassica juncea* and *Brassica napus* generated the product with *BrMuMITE1* insertions. All the six *Brassica carinata* lines showed the footsteps of the MITE element. Very strong product bands observed in the two hexaploid *Brassica* lines suggested the duplication of MITEs from diploid genomes, quite common in polyploids species (Figure 7.7).

The insertional polymorphism of *BrMuMITE4* among 40 *Brassica* lines revealed that19 have shown strong or weak bands, while the other 21 have shown no amplification signals (Figure 4.7b). The primers designed from conserved flanking regions common in A and C-genome *Brassica* were tested with six *Brassica rapa* accessions with no amplification of *BoMuMITE4* insertion but loci without it was amplified from all lines. Only one *Brassica oleracea italica* (Precoce Di Calabria) accession yielded the strong signals of *BrMuMITE4* amplification. Another *Brassica oleracea* line (GK97361) amplified a product of ~1.3 kb, which is ~200 bp larger than the expected product. The amplification pattern among *Brassica juncea* lines revealed the amplification in seven accessions except 'Kai Choy' and 'Tsai Sim' accesions. Here a question arose, if none of A and B-genomes amplified the MITE insertions, then why the allotetraploid (AABB) are generating very strong bands? Although the BLASTN searches using *BoMuMITE4-1* sequence retrieved many strong hits against *Brassica rapa* BAC clones. Their might be possibility that the *Brassica rapa* genomes harbour many *BrMuMITE4*-like sequences but the priming sites from flanking regions failed to yield the specific site. There is another possibility that no *BoMuMITE4* insertion is present at that specific locus but have other copies at variable loci. The presence and absence of *BoMuMITE4* insertional bands in 4 (Last and Best, Fortune, Drakker, Tapidor) and 2 (*NEW, MAR*) *Brassica napus* lines suggested the polymorphism in AACC genome accessions. It also confirmed the contribution of MITE inserted loci from C-genomes rather than A-genomes.

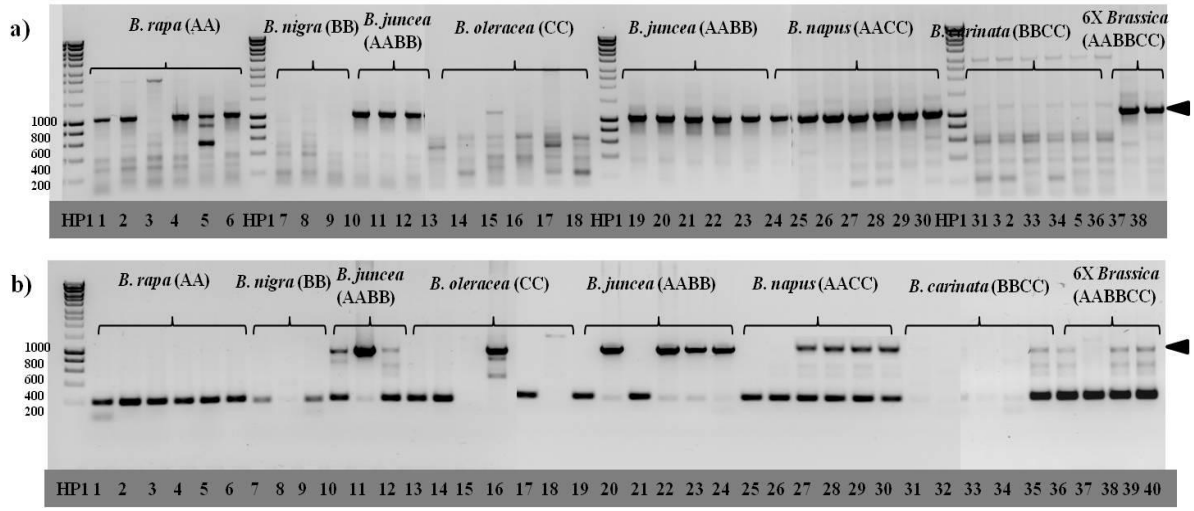


Figure 7.7: Insertional polymorphism of *Brassica* Mutator-like MITEs. a) *BrMuMITE1*; b) *BoMuMITE4*. The black arrowheads (right) indicate upper bands from loci having MITE insertions while lower bands are amplified from loci without MITE insertions. The primers used in (a) amplify some additional polymorphic sites.

7.2.5.2 Fluorescent in-situ hybridization of *BrMuMITE1*

The PCR amplification of *BrMuMITE1* has shown its amplification in *Brassica rapa* (A-genome), while no amplification of MITE inserted loci was detected among *Brassica nigra* and *Brassica oleracea* (except accession 'Early Snowball') lines. The BLASTN searches in GenBank yielded very strong hits to many full length elements from *Brassica rapa* and few hits against *Brassica junceae* and *Brassica napus*. The element represents a high copy numbered family investigated from *Brassica* in present study. To study the localization and distribution of *BrMuMITE1* on chromosomes, the amplified bands from A-genome were labelled with biotin. The probe was used on an allotetraploid *Brassica juncea* having both A and B-genome chromosomes. Metaphase chromosomes were used to localize the *BrMuMITE1* MITEs and their distribution pattern on chromosomes. Signals with varied intensities were observed on nearly all chromosomes. Strong signals were observed in A-specific chromosomes where they are mostly clustered in telomeric regions (Figure 7.8). In contrast, some chromosomes which were deduced to be without signals (B-specific chromosomes) have shown very weak signals dispersed randomly on all chromosomes. This not only confirmed the presence and distribution of *BoMuMITE1* family on *Brassica rapa* chromosomes but also presents the high diversity, abundance and distribution on both A-specific *Brassica* crops and their allotetraploids (AABB & AACC).

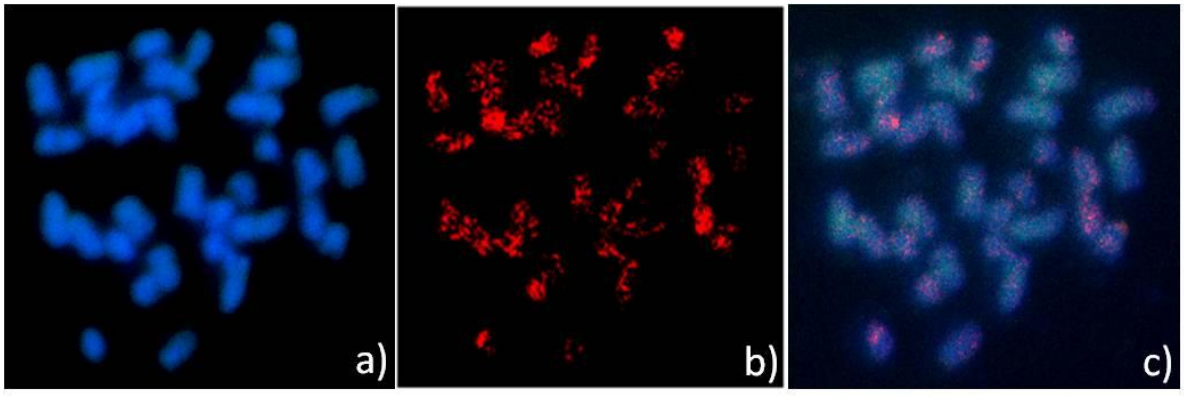


Figure 7.8: Fluorescent *in situ* hybridization showing the widespread genomic distribution of *BrMuMITE1* related sequences on *Brassica juncea* chromosomes. Metaphase chromosomes are stained with DAPI (blue fluorescence in a and c). Hybridization signals are visible as red fluorescent signals (b and purple in c where overlaying blue). Hybridization patterns of the complete *BrMuMITE1* (1 kb with flanking region) showed dispersed distribution along all chromosomes with varying signal intensities: signals observed on A-genome chromosomes were stronger than those on B-genome chromosomes. Magnification x2500.

7.2.6 Structural features of a novel MITE family (*BoXMITE1*) in *Brassica*

Some identified MITEs have shown no strong structural features to characterize and classify them. These un-characterized elements are placed in a novel MITE family. In present study, a MITE-like element with 3 bp TSDs (TTC) and 42 bp imperfect TIRs but no significant homology to any known MITEs family was identified. The element was named *BoXMITE1-1* and represents a low copy number family (*BoXMITE1*) with only 229 estimated copies within whole *Brassica* (A, C-genomes). *BoXMITE1-1*, the first identified element from the family was found inserted in *Brassica oleracea* accession 'EU642504.1' from 86275-86676 bp. Using this as query sequence against *Brassica* Nucleotide Collection database in GenBank, only two complete sequences were retrieved, while searching against *Brassica* Whole-genome shotgun contigs (wgs) database, other 3 full length copies (*BrXMITE1-2*, *BrXMITE1-4*, *BrXMITE1-5*) were also collected, which indicate their localization on chromosome 1, 4 and 7 of *Brassica rapa*. The elements range in sizes from ~308-402 bp with 3 bp TSDs with single bp mismatch. The TIRs of the family members range from 21-42 bp with few bp mismatches. *BoXMITE1-1* is flanked by 42 bp, while *BrXMITE1-4* is flanked by 21 bp TIRs (Table 7.1).

7.2.7 Estimation of genome-wide copy numbers of *Brassica* MITEs

MITEs, generally characterized by high copy numbers have shown diversity in *Brassica*. We believe that *Brassica* genomes would contain further high percentage of MITEs, which are undiscovered yet but proliferating in various *Brassica* crops. To investigate the abundance, distribution and amplification of *Brassica* MITEs, the reference sequences from each family were used as queries in BLASTN searches against *Brassica rapa* and *Brassica oleracea* Nucleotide Collection databases (nr/nt) before February, 2012. The strong hits from the output results were extrapolated to estimate genome-wide copy numbers by the formula. MITEs Copy no. = no. in database x genome size/database size. The genome sizes of *Brassica rapa* and *Brassica oleracea* ranges from 527 and 694 Mbp (Bennett and Leitch, 2011) respectively, but the average sizes of the *Brassica rapa* (535 Mbp) and *Brassica oleracea* (650 Mbp) genomes were considered for calculations. The *Brassica rapa* and *Brassica oleracea* Nucleotide Collection (nr/nt) database size available in GenBank at NCBI before February, 2012 was 51.4 Mbp and 4.7 Mbp respectively, which is 9.5 and 1 % of *Brassica rapa* and *Brassica oleracea* whole genome respectively.

BLAST searches were performed using reference MITE-like sequences as query against *Brassica rapa* and *Brassica oleracea* Nucleotide Collection database at GenBank and sequences with >70% coverage and identity were collected. The sequences with less homology were not included to avoid any false positives. We believe in the accuracy of our calculations but also expect errors in our extrapolation, due to the incomplete genome sequence datasets for *Brassica* species. The total estimated copy numbers from various *Brassica* MITE families ranges from 114 (*BoXMITE1*) to 5964 (*BoMuMITE4*). The estimated copy numbers of Stowaway-like MITE families range in sizes from 230 (*BrSTOW3*) to 3239 (*BoSTOW5*) in whole A and C-genomes collectively. The Tourist-like MITEs in total *Brassica* genomes were estimated from 1224 (*BrTOUR2)* to 3918 (*BrTOUR3*) copies. The average and high copy numbers were estimated in *BrMuMITE2* (420) and *BoMuMITE4* (5964). The second highest copy number were estimated from *BrMuMITE1* with 4892 copies among *Brassica* genomes (Table 7.2). The sharing of MITEs in both *Brassica rapa* and *Brassica oleracea* suggest their origin that predates the divergence of both species.

7.2.8 Phylogenetic analysis of Stowaway, Tourist and unknown MITEs

The Neighbour-Joining method with 1000 bootstrap replicates was used to generate the phylogenetic tree of *Brassica* Stowaway, Tourist and BoXMITE families. The tree is based on genetic distance calculated with Tamura-Nei (1973) genetic model and rooted with *Hordeum vulgare* Tourist-like element *Jura_HV*. Due to very high copy numbers of MITEs, five elements from each respective family were used. In the initial effort, full length elements were aligned and a tree was generated, which failed to resolve elements into their respective groups due to very high AT content resulting in similarity of Stowaway and Tourist-like MITEs. The TIRs of the MITEs were collected, aligned in CLUSTALW and tree in generated in Geneious program. Each family of Stowaway, Tourist and *BoXMITE* MITEs are shown in different colours. The tree revealed that the members of each family clustered together in their respective groups with the exception of few sequences. No clear cut separation of Stowaway or Tourist-like MITEs into two major clades was found but weak clustering of Stowaway or Tourist MITE families together was found (Figure 7.9). The Stowaway MITE families *BrSTOW1, BoSTOW2* and *BoSTOW3* clustered together, while *BoSTOW4* and *BoSTOW5* come close to each other. Both these groups were separated by Tourist-like MITE families *BrTOUR1, BrTOUR2, BrTOUR3*

and *BoTOUR4*, which come close to each others. Three elements *BrTOUR2-1*, *BrTOUR3-1* and *BrTOUR3-5* are found dispersed in *BoSTOW3* clade indicating their misplaced positions due to substitution or replacement of few nucleotide bases in TIRs. Similarly *BrTOUR4-5* is not placed in its specific group due to variations in its TIRs. The sequences from the *BoXMITE* family were clustered together in one group and the family is located between *BoSTOW4* and *BrSTOW5*. Therefore, it was concluded that MITE TIRs can be used to generate weak clustering of family members but there is no specific clustering of Stowaway and Tourist-like elements in two separate lineages. The possible reason might be the high and similar GC content in TIRs of both superfamilies and the short sizes of TIRs, so it is unsuccessful in differentiating the two superfamilies of MITEs.

7.3 Discussion

In the present study, the molecular characterization of 15 novel MITE families was done in *Brassica*. Including the MITE derivatives of Tc1-Mariner, PIF-Harbinger and Mutator-like transposon superfamilies, of the 15 elements, 5 are Stowaway, 4 Tourist-like, 5 Mutator-like and 1 a novel MITE family named *BoXMITE*. Approximately 29112 MITEs-like sequences were estimated from *Brassica rapa* and *Brassica oleracea* genomes (Figure 7.2 & 7.6; Table 7.1 & 7.2) (Several hAT-like non-autonomous families were also investigated from *Brassica*, described in previous chapter). *BraSto*, a well characterized Stowaway MITE family was reported with similar abundance to our family in *Brassica* (Sarilar *et al.*, 2011). The rice genome harbours rather more elements, with ~178,533 MITE related sequences clustering into 338 families (Lu *et al.*, 2012). A parallel study in the *Solanaceae* has also revealed a high level of MITE diversity among the crop species (eg. tomato, potato and tobacco) and 22 families including superfamilies Stowaway, Tourist, hAT, and Mutator-like MITEs. The *MiS1* family occurs in high copy numbers in tobacco (>1000), while low copy numbers were observed in tomato and potato (<60). Like high copy numbers of *Brassica* Mutator-like MITEs, the Mutator-like MITEs (*MiS6–MiS11*) in tomato and potato have shown higher estimated copy numbers (~400–3200). The most abundant family was *MiS22*, which are estimated to have 3516 and 9802 copies in the tomato and potato genomes (Kuang *et al.*, 2009).

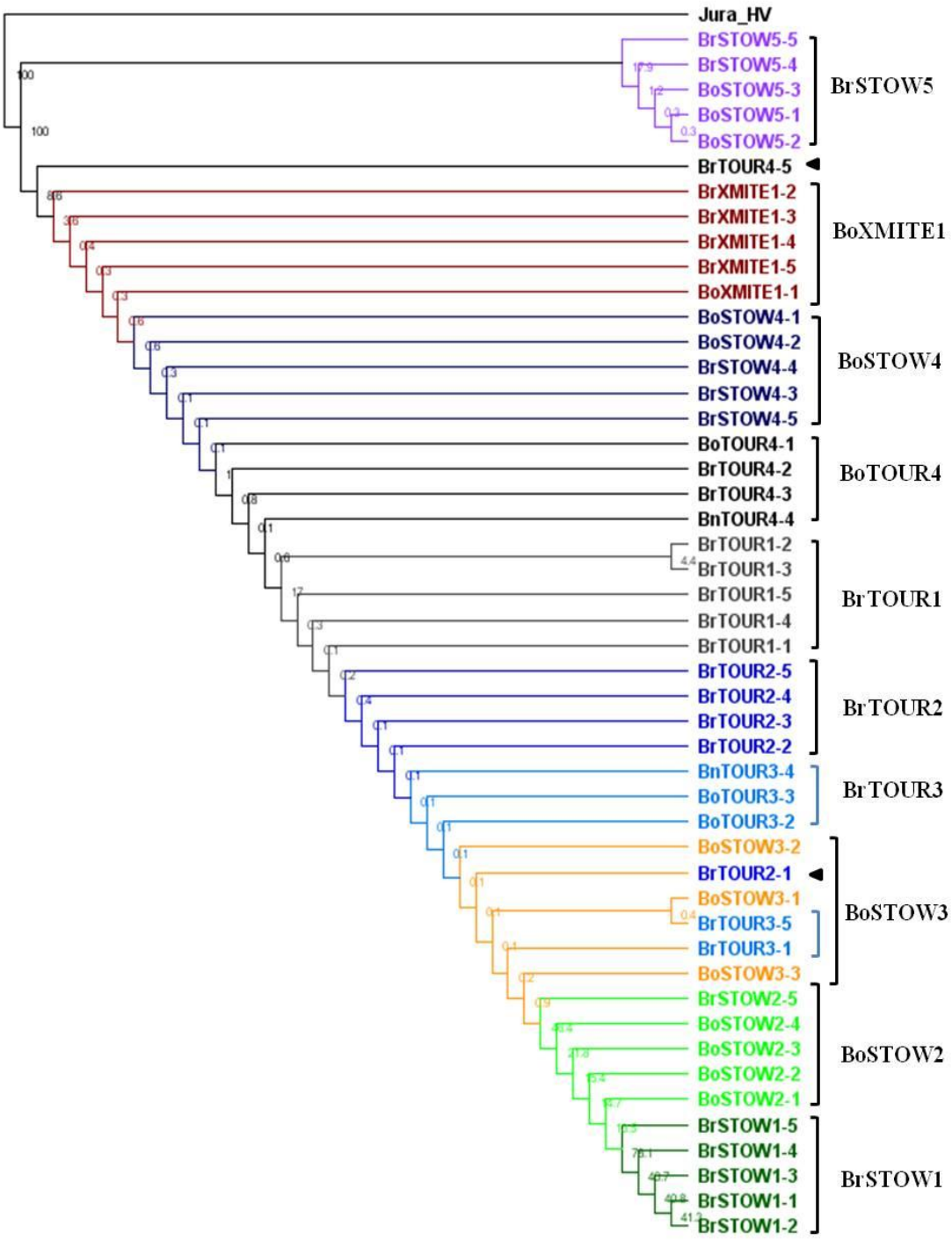


Figure 7.9: Phylogenetic tree of *Brassica Stowaway*, *Tourist* and novel MITEs. Tree is based on genetic distance calculated with Tamura-Nei (1973) and constructed with Neighbour-Joining methods with 1000 bootstraps replicates (substitution values shown). *Hordeum vulgare Tourist-like* element *Jura_HV* was used to root the tree. Each family of *Stowaway*, *Tourist* and novel MITEs are shown in different colours. The arrowheads are indicating misplaced sequences. Stowaway and Tourist superfamilies of MITEs were only poorly resolved on the basis of their small TIRs with high GC content; internal AT-rich regions generated only a weakly supported tree.

7.3.1 Insertional polymorphism of MITEs, a tool to study diversity and evolution

The insertional polymorphism of MITEs is an excellent tool to identify different cultivars and their lineages. MITEs have the ability to transpose into new sites at variable pace in different cultivars or genotypes, building the presence/absence based polymorphism (Lyons *et al.*, 2008) as described for retrotransposons in the RBIP analysis (Flavell *et al.*, 1998). A site is known as empty site (RESite), when the MITE moved from a locus and transposed to a new site (Le *et al.*, 2000). The excision of the MITE from a site causes the presence/absence polymorphism; with few genomes with MITE insertion while the others lacking the insertion. After a MITE transpose and integrates to a new site, the empty donor host site exhibit a foot print having an extra TSD sequence as compared to the locus prior to MITE insertion. This excision is not always normal, sometimes such excisions often generates footprints, including short deletions or the insertions of the unrelated sequences (Kikuchi *et al.*, 2003). Different MITE families have shown different patterns of proliferation. Some MITEs are highly conserved and proliferate only in a genome while others are highly diverse and actively proliferating in various genomes. The high conservations of the MITEs in a genome indicate their recent amplification and burst, while the highly dispersed MITEs families are resulted from the ancient amplification and proliferations (Oki *et al.*, 2008; Zerjal *et al.*, 2012).

The Insertional polymorphism of *Brassica* Stowaway, Tourist and Mutator derived MITEs was observed among 40 cultivars. The amplification of BrSTOW1 in *Brassica rapa* and BoSTOW3 and BoSTOW4 in *Brassica oleracea* suggested the conserved amplification of MITEs in a A and C-genomes respectively and showing RESites in the genomes, where no amplification is observed. Similarly, the amplification of *Brassica* Tourist and Mutator MITEs yielded products with and without insert displaying the Insertional polymorphism. This polymorphism helped us in the identification and differentiation of many cultivars in *Brassica*. The MITE insertion polymorphisms concluded that they are excellent molecular markers to study the diversity and evolutionary phenomena in plant genomes.

7.3.2 *Brassica* MITEs have highly AT rich regions

One of the typical features of MITEs is the presence of highly AT rich sequences (eg. the *AhMITEs* from *Arachis hypogea* exhibit an AT content of 70% (Shirasawa *et al.*, 2012), a

characteristic found in all *Brassica* MITE families. The average AT content within the families ranges from 53% (*BoXMITE1*) to 80% (*BoSTOW4*, *BrMuMITE2*). The Stowaway MITEs showed a range of 62-80% AT content, where *BoSTOW3* family showed minimum while *BoSTOW4* showed maximum ratios. Tourist-like MITEs are also AT rich with an average 63-78%. Mutator-like MITEs has shown very high AT rich regions within TIRs and internal regions. The AT content among the families ranges from 75-80%, with the exception of *BrMuMITE5* (58%) (Table 7.1 & 7.2)

7.3.3 Evolutionary implications of MITEs in plant genomes

Besides the characteristic features of MITEs with conserved TSDs, TIRs and high copy numbers within the genomes, the MITEs have played a role with functional and evolutionary implications. MITEs can either capture the genes or integrate within the functional genes as many MITE-like copies are an integral part of genes. For example the *BoSTOW4* family members are residing in *Brassica oleracea* S12 SLG gene for S locus glycoprotein. *BnSTOW5-6* is homologue of SPH gene, AT4g29040 gene and AT4g29100 gene in *Brassica napus*. *BoTOUR3* members capture *Brassica oleracea* cultivar Reihou FLC2 gene , *Brassica oleracea* cultivar Green Comet truncated FLC2 gene and *Brassica napus* cultivar Westar WRKY transcription factor 18 (WRKY18) allele, *BnTOUR4-4* is present in *Brassica napus* M3.4 protein gene, *BrMuMITE3* family members are located in various genes specifying different functions as *Brassica rapa* BrSRK-8, BrSP11-8, BrSLG-8 genes for S receptor kinase etc. Similarly, other families of the MITEs either capture gene fragments or are proliferating in intergenic regions. The occurrence of MITEs in various genes revealed their ability to alter the genes and playing a prime role in the evolution of new genes and diversity of the organism. This hypothesis was supported by the study of MITEs and their contribution to evolution of gene complexity in members of *Solanaceae*.

The study revealed that several MITEs have the capacity to modify the genes and playing a role in the evolution of new genes and creating new gene structures by various ways. The MITEs identified from *Solanaceae* family like MiS1, MiS2, MiS5 and related elements residing in the members of *Solanaceae* are playing a role in evolutionary properties of the genomes (Kuang *et al.*, 2009). The localization and association of MITEs with plant genes provoked the scientific community to think about the positive role of

MITEs in gene regulation and genomic evolution. Many of the workers have explained the significant role of MITEs in various genomes such as MITEs provides the coding sequences or the poly(A) signals for genes and controlling the host genes in which they are actively proliferating. MITEs are mostly distributed on the chromosomal arms, where they are associated with the functional genes. There is also evidence of transcription of MITE sequences with the plant genes (Oki *et al.*, 2008; Kuang *et al.*, 2009).

7.4 Conclusion

The present work helped in the understanding of various MITE families by their identification, characterization, annotation, distribution and diversity in *Brassica* genome, as well as investigating their flanking genomic sequences, insertional polymorphisms in various accessions and their transposition ability. These findings will contribute to the scientific community in understanding of *Brassica* diversity and assist in the progression of genetics, genomics and the breeding of *Brassica* and its cultivars.

CHAPTER 8
THE LTR RETROTRANSPOSON LANDSCAPE IN *MUSA* GENOMES

Summary

Fifty full length LTR retrotransposons in *Musa* were identified by dot plot analysis and further collected 153 intact copies, 61 truncated and a great number of partial copies and remnants by blast searches from GenBank database. Phylogenetic analysis of 33 autonomous retrotransposons based on RT regions segregated them into 25 families, of which 15 families are Copia, 9 are Gypsy and 1 Pararetrovirus (PRV)-like superfamily. The analysis of 50 elements on the basis of LTR sequences clustered them into 40 families. LARD-like elements were also identified with several copies dispersed among the genomes. The predominant elements are Copia and Gypsy, while Pararetroviruses are very less frequent in *Musa* genomes. The elements belonging to Copia and Gypsy families were of low, middle and high copy number having intact copies, solo LTRs, deletion derivatives and remnants. The phylogenetic classification of elements was performed on the basis of LTRs and conserved RT domains and a distribution analysis of LTR retrotransposons in the genome of *Musa.* This phylogeny analysis revealed the clustering of Copia, Gypsy and PRV superfamilies in different clades.

8.1 Introduction

Among transposable elements, the major proportion in plants is represented by LTR retrotransposons, which reverse transcribe their RNA to generate DNA copy integration to new host sites (Eickbush and Jamburuthugoda, 2008). The previous studies reveal that retroviruses and related retro-transcribing viruses have evolved from LTR retrotransposons. Retroviruses and LTR retrotransposons share similar structural features but presence of an envelope (env) gene in retroviruses distinguish them from other LTR retrotransposons (Lerat and Capy, 1999; Llorens *et al.*, 2009). Recent study have shown a high diversity of retroelements in fungi, where several families of Copia, Gypsy and retroviruses are actively proliferating. The Chromovirideae (chromodomain-bearing) Gypsy elements are common in fungi and duplicating in fungal genomes. It was studied that the transposon expansions are the consequences of both increase in copy numbers of the elements and number of the elements types. The *gag-pol* organization distinguish the

Gypsy and Copia superfamilies, while the evolutionary relationships of *pol* gene revealed that the protease is the fastest evolving domain in comparison to RT and RNase H domains (Muszewska *et al.,* 2011).

The LTR retrotransposons were investigated in many eukaryotic genomes and after developing the SSAP, IRAP, REMAP and RBIP techniques, it becomes more feasible to study the diversity and landscape of LTR retrotransposons in various organisms (Flavell *et al.*, 1998; Schulman *et al.,* 2004). In the recent years these markers are utilized in several plant genomes to study the biodiversity of plants like wheat (Queen *et al.,* 2004), pea (Jing *et al.,* 2010), *Vicia* (Sanz *et al.,* 2007) and several other plants. In the last decade, due to their major role in gene and genome evolution and duplication, the LTR retrotransposons are the centre of focus and the scientific community is collecting data to develop their databases.

Repbase is the most informative database comprising almost all different superfamilies of repetitive sequences. Enormous data about various TE superfamilies including LTR retrotransposons, their superfamilies, families and individual elements from various eukaryotic genomes were identified and deposited and more and more data is regularly updated in Repbase Updates (Jurka *et al.,* 2007). Recently Gypsy database (GyDB) is developed with the target to analyze and classify the diversity of mobile genetics elements. GyDB is regularly updating the analysis of Ty3/gypsy, Ty1/copia, Bel-pao and other LTR retrotransposons including the Pararetroviruses. A large amount of LTR retrotransposon sequences are deposited in GyDB including complete copies as well as individual core domains. Seventy five Copia, Gypsy and Pararetroviruses-like elements (Table 2.6) were collected from GyDB in the present study for comparison of identified elements in *Musa* and *Brassica* and their evolutionary relationship with them (Llorens *et al.,* 2011).

In this chapter, the study aimed to identify mobile elements with long terminal repeats (LTRs) within sequenced BACs from *Musa acuminata* and *Musa balbisiana* to study the complete landscape of LTR retrotransposons in *Musa* genotypes.

8.2 Results

8.2.1 Strategy for characterizing and mining LTR retrotransposons in *Musa*

The *Musa* genome harbours LTR retrotransposons belonging to different superfamilies. They were characterized on the basis of their structural features (TSDs, LTRs), pattern of conserved domains in *gag-pol* regions (PBS, RT, RH, INT, PPT). In this work, available (~46) *Musa* BAC sequences >70 kb from NCBI database were collected before June, 2010 (no further BACs were sequenced before March 2012), which were screened for any detectable LTR retrotransposons by using dot plot analysis, where sequences were plotted against themselves. A sharp diagonal line crosses from one corner to the other showed the complete homology of the sequence, whereas two small lines on both corners of the central line indicate 5′ and 3′ LTRs (Figure 8.1 & 8.2). The LTRs were defined and TSDs were identified by visual inspection. Their 5′-TG….CA-3′ termini are marked and total sizes of the element were counted and enlisted (Table 8.1). The Repbase and Gypsy database were used to characterize the retrotransposons on homology basis to the known elements, very few retroelements were characterized, while others were characterized by visual inspection on the basis of their TSDs, LTRs, pattern of PBS and PPT and protein domain organizations. The LTR retrotransposons were considered in a family, when they showed >85% identity at their nucleotide level in their coding regions or internal domains (Wicker *et al*, 2007; Minervini *et al*., 2009). A novel family was reported when no homology was observed with any known LTR retrotransposon, there were LTRs and internal domains for its transposition, and it showed strong hits to at least another sequence excluding the reference query (Wang and Liu, 2008).

Fifty intact (full length) elements were identified initially by dot plot analysis and considered as reference elements. The identified elements were further used to conduct the BLASTN searches against the *Musa* Nucleotide Collection (nr/nt) database in NCBI. The searches were performed in several steps to identify the intact, truncated, degraded or partial elements, solo LTRs and remnants. An intact element is one that is terminated by TSDs and LTRs, with one or more *gag-pol* genes domains. Solo LTR refers to an LTR with TSD, or LTRs truncated with small deletions displaying >80% query coverage and homology. Truncated elements are defined as elements having deletions at 5′ or 3′ ends of LTRs. Partial sequences are the deletion derivatives showing >40% query coverage, with

or without LTRs and one or more conserved domains (PBS, AP, RT, RH, INT, PPT). The term remnants describes all the small fragments showing <40% query coverage with strong or weak identity to the retrotransposon sequences (see Chapter 2; Figure 2.3).

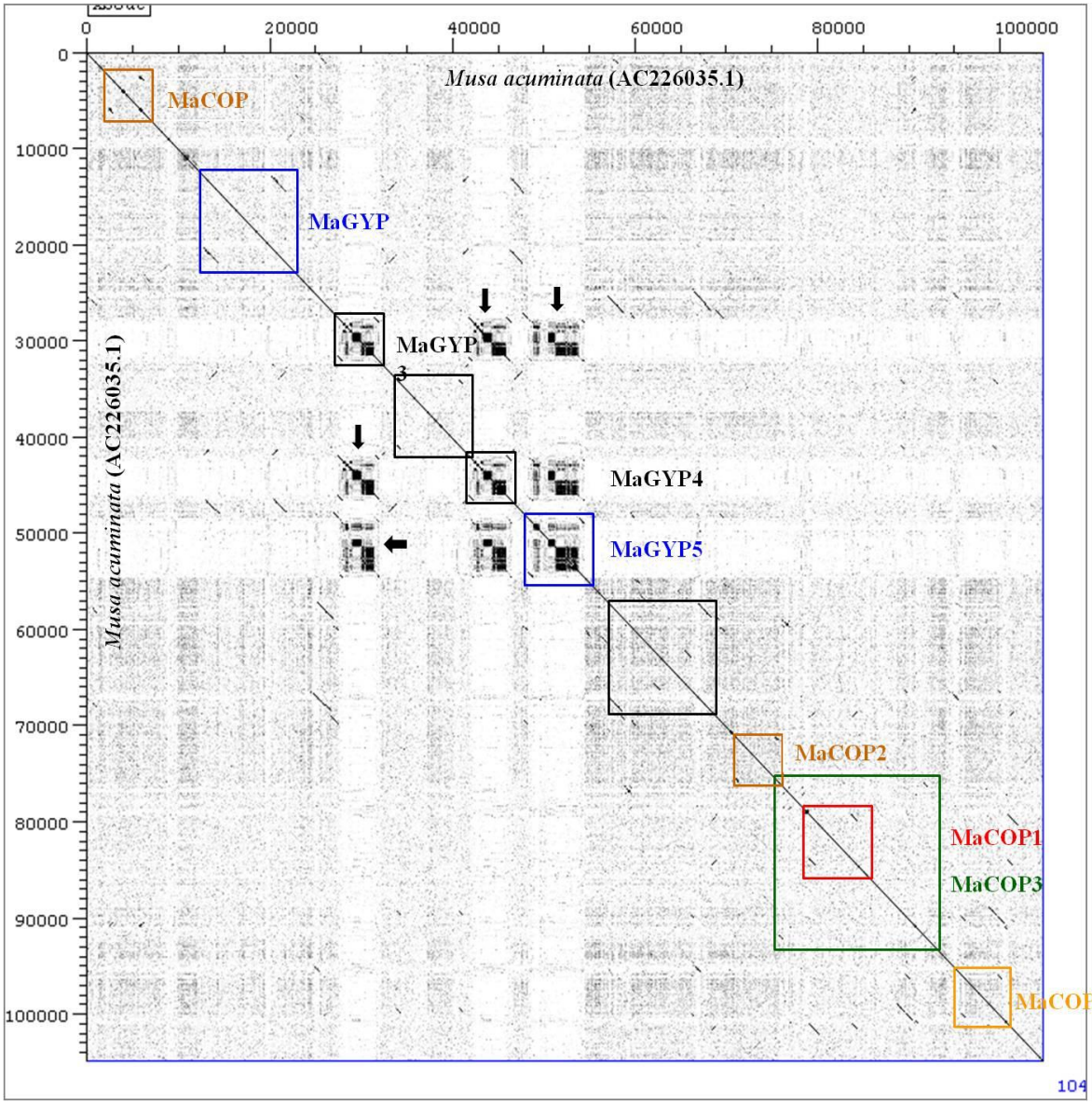


Figure 8.1: Dot plot of *Musa acuminata* BAC AC226035.1 against itself to identify LTR retrotransposons. The central diagonal line running from one corner to other shows the homology of the sequence. The coloured boxes on the diagonal line show the position of LTR retrotransposons insertions with LTRs. Five Copia and four Gypsy elements are inserted with a total size of ~60 kb out of 105 kb BAC size covering 58.5% of total BAC sequence. Uncharacterized elements are not indicated, but account for a further 20-30% of the BAC. The unlabelled black boxes indicate uncharacterized elements lacking any recognizable internal *gag-pol* encoding regions. Amplified elements are shown by mobile or transposed copies indicated by arrows. The nested structure of LTR retrotransposon is also shown. The identified elements are followed by numbers. Elements without a number represent elements, which are not listed in the 50 reference sequences.

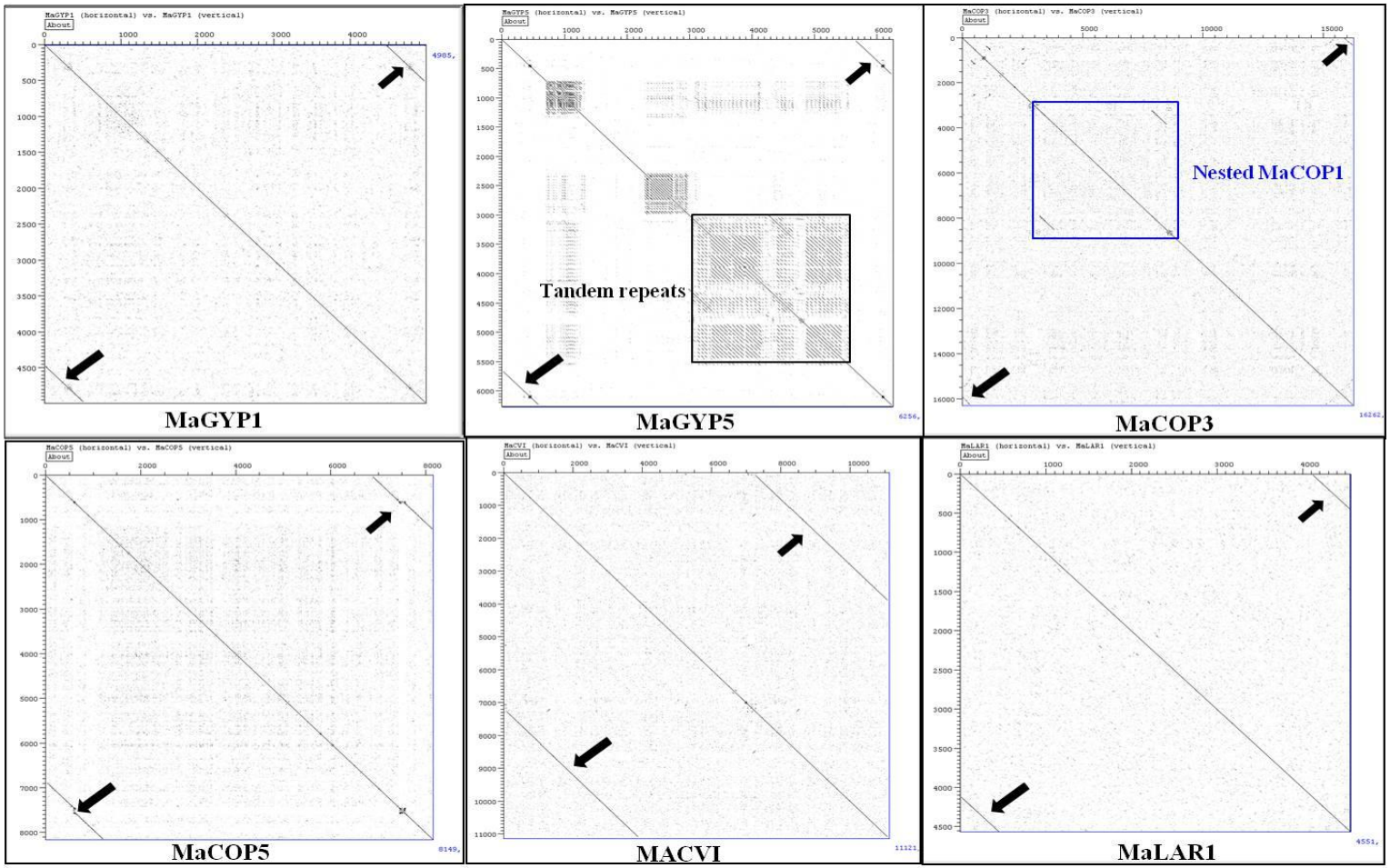


Figure 8.2: Dot plots showing the LTRs in Gypsy, Copia, Pararetroviruses and LARDs. The sequences from elements are plotted against themselves. The central diagonal line shows sequence identity. Parallel lines at the corners indicate the LTRs. The boxed lines across the diagonal line in *MaGYP5* indicate tandem repeats in that specific region. A nested Copia *MaCOP1* is integrated in *MaCOP3*. The largest LTRs are shown by a Pararetrovirus-like element *MACVI*. The LARDs (*MaLAR1*) display perfect LTRs and are considered as non-autonomous LTR retrotransposons due to lack of internal coding domains.

8.2.2 The LTR retrotransposons landscape in *Musa*

Out of the initially identified 50 retrotransposons, 20 elements belong to Gypsy, 19 to Copia, 1 to Caulimoviridae (Pararetroviruses) and 10 to LARD-like elements (Table 8.1). The search was extended by using these reference elements as query in BLASTN searches against *Musa* genomes in GenBank database and all full length, truncated or partial copies and remnants were counted. A total of 16246 elements and fragments belonging to Copia, Gypsy, Caulimoviridae and LARDs were identified, out of which 153 are intact elements, i.e. 58 elements from Gypsy, 48 from Copia, 1 Pararetrovirus and 46 from LARD-like elements. A total of 61 truncated elements, 635 partial elements, 258 solo LTRs and 15140 remnants were counted from *Musa* Nucleotide Collection (nr/nt) database deposited in NCBI (Figure 8.3). The ratio of intact elements to Solo LTRs in *Musa* BAC sequences is ~2:3. Based on retrieved copy numbers from database, total numbers of intact copies were estimated for each Gypsy, Copia and LARDs. The total genome of *Musa acuminata* and *Musa balbisiana* was 610 Mbp and 560 Mbp respectively (Kamate *et al*., 2001). The available genome in *Musa* nucleotide databases for A and B-genome was 4.1 and 2.1 Mbp respectively before December, 2011. Estimated copy numbers in *Musa acuminata* for Gypsy, Copia and LARDs were 4800, 6000 and 4650 respectively, while for *Musa balbisiana* 4400 copies of Gypsy, 4950 copies of Copia and 4125 copies of LARDs were estimated (Figure 8.3).

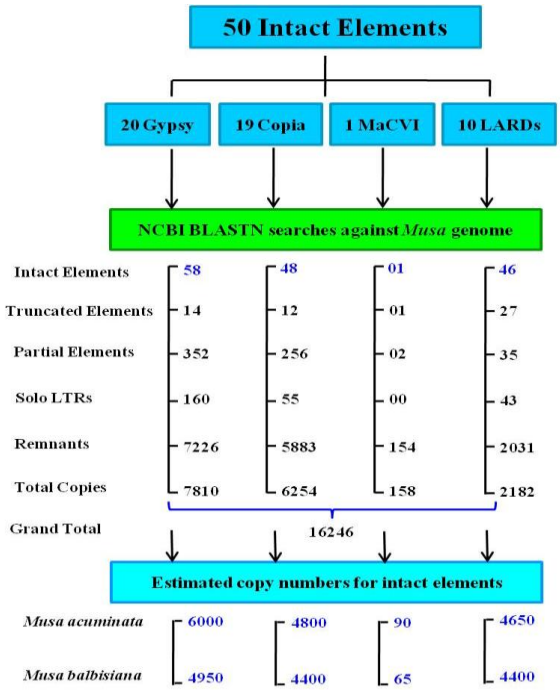


Figure 8.3: Total number of intact, truncated and partial elements with their fragments in *Musa acuminata* and *Musa balbisiana* genomes. Estimated copy numbers were calculated for intact elements. No truncated, partial elements or remnants were used for copy number estimation.

8.2.3 Phylogeny and families of LTR retrotransposons in *Musa* by RT analysis

The phylogenetic relationship of *Musa* LTR retrotransposons identified in present study was analyzed. The RT sequences were present in 33 out of 50 elements. The tree was generated with Neighbour-Joining method with 1000 bootstrap replicates. The *Saccharomyces cerevisiae* Gypsy '*Ty3-1*' was used to root the tree (Figure 8.4). Another Copia '*Ty1*' from the same plant was used to observe its relationship with Copia elements. Three main lineages separated the Copia, Gypsy and Caulimoviridae (Pararetrovirus) elements with 18, 13 and 2 elements respectively. Again the *Musa balbisiana* Gypsy element (*MbGYP20*) come close to a *MaCV1* indicating a relation of this Gypsy with Pararetrovirus-like elements. The detailed structural analysis confirmed that *MbGYP20* possess nested structure, where another retrotransposon and a DNA transposons is inserted. Their might be a possibility of RT gene insertion from Pararetrovirus-like element in it, which cluserd it together in phylogenetic studies. The alignment of Gypsy and Pararetrovirus RT sequences indicate a close relationship of these two superfamilies and distinct from Copia. The tree clearly resolved the elements to the level of a family. The 33 elements segregated into 25 families, of which 9 are from Gypsy, 1 from Pararetrovirus and 15 from Copia elements (Figure 8.4).

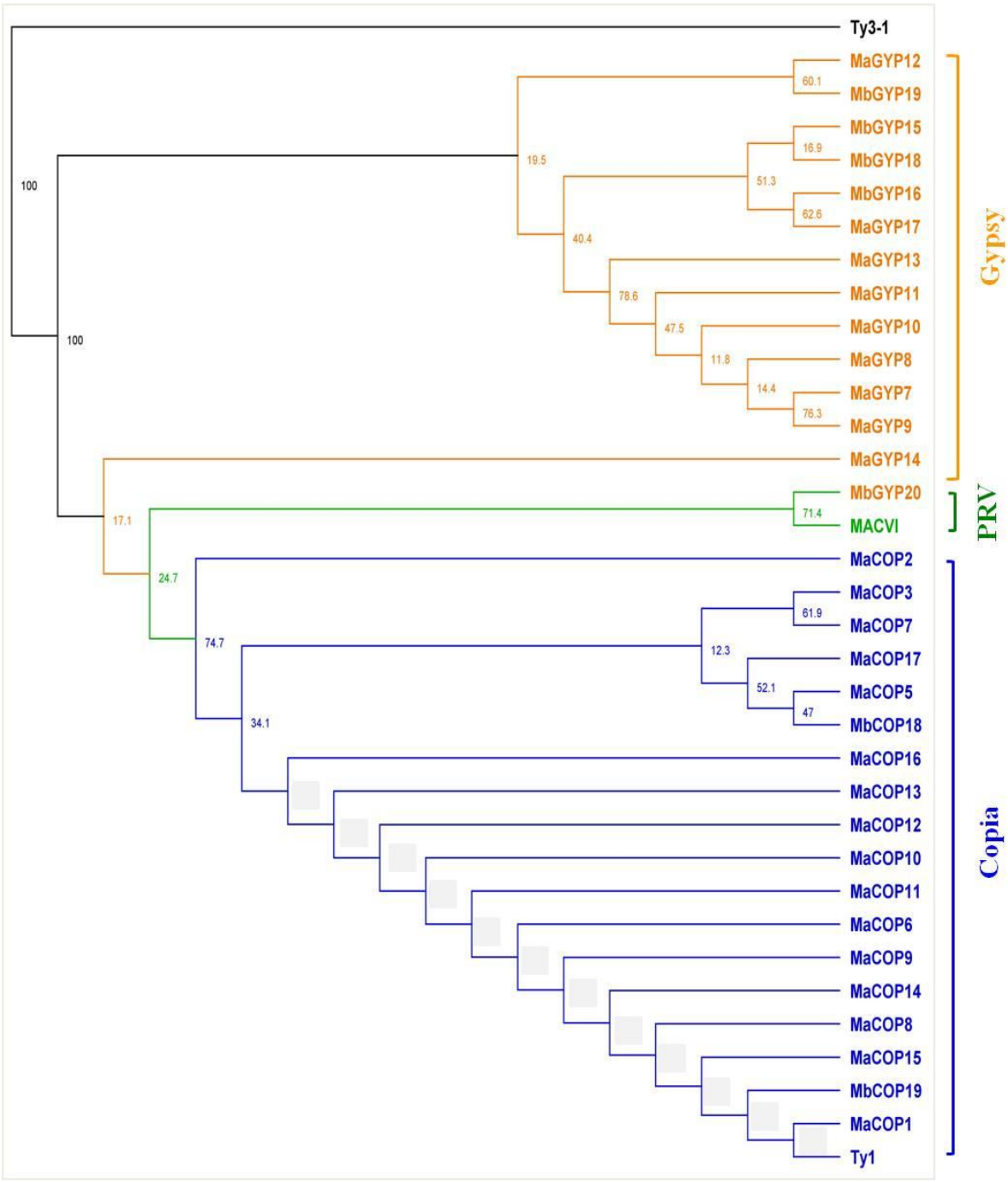


Figure 8.4: Phylogenetic relationships of LTR retrotransposon families from *Musa*. The RT sequences from 33 reference elements were used to construct the trees, which were rooted using the RT sequences of Ty3-1 (Gypsy) element of *Saccharomyces cerevisiae.* Neighbour-Joining tree was constructed with 1000 bootstrap replicates in Geneious Pro. Lineages separate the Gypsy, Copia and Pararetrovirus-like elements. About 25 novel families (9 Gypsy, 15 Copia, 1 Pararetrovirus families) were observed. Ma: *Musa acuminata*. Mb: *Musa blabisiana*. COP: Copia. GYP: Gypsy. MaCVI: *Musa acuminata* chromovirideae. Details of each element are given in the table 2.3.

Table 8.1: List of Copia, Gypsy, LARDs, and TRIMs with their sizes, TSDs, TIRs, positions and orientations in BAC clone sequences. Esterisks after TSD show variable TSDs at 5′-3′. Nuclotide sequences of representative elements are available in Appendices (attached CD). ND: Not determined.

Element name	Super-family	BAC Accession	Species	Size	TSD	LTR	Position	Orient-ation
MaGYP1	Gypsy	AC226032.1	*M. acuminata*	4982	CCCGG	505/505	65772-70752	5′-3′
MaGYP2	Gypsy	AC226033.1	*M. acuminata*	3802	AGATG	543/527	28867-32673	5′-3′
MaGYP3	Gypsy	AC226035.1	*M. acuminata*	4567	ATGAG	458/458	27408-31974	5′-3′
MaGYP4	Gypsy	AC226035.1	*M. acuminata*	4627	TAGGA	458/458	41793-46419	5′-3′
MaGYP5	Gypsy	AC226035.1	*M. acuminata*	6245	ACTTC	586/586	48237-54481	5′-3′
MaGYP6	Gypsy	AC186752.1	*M. acuminata*	3015	TATGT*	655/655	62134-65148	5′-3′
MaGYP7	Gypsy	AC226046.1	*M. acuminata*	5326	AATAT	462/411	116297-121622	3′-5′
MaGYP8	Gypsy	AC226048.1	*M. acuminata*	5907	TGTTT	473/473	1861-7767	5′-3′
MaGYP9	Gypsy	AC226048.1	*M. acuminata*	5318	AGACG	481/506	12597-17915	5′-3′
MaGYP10	Gypsy	AC226048.1	*M. acuminata*	5435	GAGAT	438/438	24477-29911	3′-5′
MaGYP11	Gypsy	AC226048.1	*M. acuminata*	5319	AGACG	481/519	12597-17915	5′-3′
MaGYP12	Gypsy	AC226048.1	*M. acuminata*	5760	CTGAC	671/671	36420-42179	3′-5′
MaGYP13	Gypsy	AC226048.1	*M. acuminata*	5418	AAACT*	1062/1063	123405-128822	3′-5′
MaGYP14	Gypsy	AC186950.2	*M. acuminata*	11605	CCAGT	624/624	9314-20918	3′-5′
MbGYP15	Gypsy	AC226053.1	*M. balbisiana*	4940	GTTAA*	883/884	121237-126176	5′-3′
MbGYP16	Gypsy	AC226051.1	*M. balbisiana*	4014	TAAA	265/264	125881-129840	3′-5′
MaGYP17	Gypsy	AC226196.1	*M. acuminata*	6436	TCCT	792/792	10198-16633	5′-3′
MbGYP18	Gypsy	AP009325.2	*M. balbisiana*	7108	GCACC*	374/383	45693-52799	5′-3′
MbGYP19	Gypsy	AP009334.1	*M. balbisiana*	7368	GGTAT	1105/883	44828-52195	5′-3′
MbGYP20	Gypsy	AP009325.2	*M. balbisiana*	17804	GCCAC	393/351	79527-97303	3′-5′
MaCOP1	Copia	AC226035.1	*M. acuminata*	5290	CTGCA	605/605	79036-84325	5′-3′
MaCOP2	Copia	AC226035.1	*M. acuminata*	4808	TCTCT	358/360	70977-75784	5′-3′
MaCOP3	Copia	AC226035.1	*M. acuminata*	16242	GAGG	338/299	75834-92033	5′-3′
MaCOP4	Copia	AC226038.1	*M. acuminata*	4022	CCATA	261/263	35057-39078	3′-5′
MaCOP5	Copia	AC226038.1	*M. acuminata*	8158	CATAA	1201/1285	63867-72094	3′-5′
MaCOP6	Copia	AC226038.1	*M. acuminata*	7036	GAATC	452/472	100169-107204	3′-5′
MaCOP7	Copia	AC226041.1	*M. acuminata*	5012	ACTAA	149/144	2059-7070	3′-5′
MaCOP8	Copia	AC226044.1	*M. acuminata*	6019	GGATT	499/500	40359-46377	3′-5′
MaCOP9	Copia	AC226047.1	*M. acuminata*	6959	GGTTT	526/530	68302-75260	5′-3′
MaCOP10	Copia	AC226047.1	*M. acuminata*	8767	TGTAT	1597/1597	15958-24725	3′-5′
MaCOP11	Copia	AC226051.1	*M. acuminata*	8478	AAAG	1494/1388	34505-42982	5′-3′
MaCOP12	Copia	AC226051.1	*M. acuminata*	7176	AGCGA*	1132/1238	118482-125657	3′-5′
MaCOP13	Copia	AC226051.1	*M. acuminata*	5938	ND	548/573	119071-125008	3′-5′
MaCOP14	Copia	AC186753.1	*M. acuminata*	6054	GAAAT*	492/548	28872-34925	3′-5′
MbCOP15	Copia	AC226053.1	*M. balbisiana*	4980	ACCTT	449/449	100799-105778	3′-5′
MaCOP16	Copia	AC226040.1	*M. acuminata*	5424	GCAAC	438/406	40145-45568	3′-5′
MaCOP17	Copia	AC226196.1	*M. acuminata*	8084	GATAT*	1024/1000	50839-58921	3′-5′
MbCOP18	Copia	AC226052.1	*M. balbisiana*	9878	TGTC*	1396/1415	184519-194396	3′-5′
MbCOP19	Copia	AC226055.1	*M. balbisiana*	5203	TTCA	592/590	26728-31930	5′-3′
MaCVI	(PRV)	AC226046.1	*M. acuminata*	11077	CTCT	3866/3813	160034-1711104	5′-3′
MaLAR1	LARDs	AY484588.1	*M. acuminata*	4564	GGTT	447/447	48330-52793	5′-3′
MbLAR2	LARDs	AC226055.1	*M. balbisiana*	4428	ATAT	445/445	9329–13756	5′-3′
MbLAR3	LARDs	AP009334.1	*M. balbisiana*	4452	ATGC	383/383	20981-25432	3′-5′
MaLAR4	LARDs	AC186955.1	*M. acuminata*	4318	GTATT*	607/611	47077-51394	3′-5′
MbLAR5	LARDs	FN396604.1	*M. balbisiana*	4449	ATAC	382/382	28462-32910	5′-3′
MbLAR6	LARDs	FN396605.1	*M. balbisiana*	4449	GGAG	382/382	36620-41068	5′-3′
MaLAR7	LARDs	AC186951.1	*M. acuminata*	4571	ATAT	446/446	92933-97503	3′-5′
MaLAR8	LARDs	AC186753.1	*M. acuminata*	4550	GTAG	434/437	15832-20381	5′-3′
MbLAR9	LARDs	AC186754.1	*M. balbisiana*	7712	ATTGT*	626/635	72565-80276	3′-5′
MaLAR10	LARDs	AC226051.1	*M. acuminata*	4005	TTTC*	974/984	129126-133076	5′-3′

8.2.4 Characterization of *Musa* Gypsy retrotransposons

The intact elements were analyzed in detail for their structural features, size, number of TSDs, length of LTRs and their termini, their position and orientation in the BAC sequences. The number of Gypsy and Copia were nearly similar but the later predominated. The Gypsy elements range in sizes from 3015–17804 bp. The average sizes of the Gypsy-like elements are between 4.5 to 5.5 kb. The smallest element *MaGYP6* is a non-autonomous Gypsy coding the *gag* region only. *MaGYP2, MaGYP3, MaGYP4* and *MaGYP5* belongs to same family and their *gag-pol* protein coding domains necessary for transposition are missing indicating them non-autonomous elements. They code some other protein domains, which are not the structural domains for Gypsy superfamily except *MaGYP3* and *MaGYP4*, which encode the aspartic protease (AP) that is a domain present in retrotransposons (Table 8.1 and 8.2). The largest element (17804 bp) *MbGYP20* showed a nested structure, where another element *MaGYP8* is inserted in to it (Figure 8.6). Another 11605 bp long element *MaGYP14* was observed in *Musa acuminata* BAC clone 'AC186950.2'. This element has a ~2 kb simple sequence repeat (SSR) insertion near its 3′ end. All the investigated elements have shown target site duplications (TSDs) on both sides of LTRs. A total of 90% elements are terminated by 5 bp TSDs, while rest (10%) showed 4 bp TSDs. Almost 80% of elements are terminated by perfect TSDs without any bp mismatch, but 20% of them showed a single bp mismatch in their TSDs. The TSDs are mostly AT rich with the exception of *MAGYP1, MaGYP18* and *MaGYP20*, which showed more GC% in their TSDs. The LTRs of the Gypsy elements range in size from 264–1105 bp. The smallest sizes of LTRs were observed in *MaGYP16*, while the largest LTR was 5′LTR of *MbGYP19*, which has an insertion of 265 bp in it. The average sizes of LTRs in Gypsy elements are 450-550 bp (Table 8.1).

The distribution of intact elements in *Musa* BAC sequences were also identified which showed uneven distribution patterns within BAC sequences. The *Musa acuminata* BAC clone 'AC226035.1' showed the highest activity of retrotransposons, which contain 5 Copia and 4 Gypsy elements with a total size of ~60 kb out of 105 kb BAC size covering 58.5% of total BAC sequence (Figure 8.1). Another *Musa acuminata* BAC 'AC226048.1' showed the maximum activity of retrotransposons transposition, where 6 Gypsy elements *MaGYP8–MaGYP13* were detected (Table 8.1). These elements covered a total of ~31 kb (24%) of 134.5 kb BAC sequence. The partial copies or remnants from there elements

further increase the size and percentage of these elements in these BAC sequences. This revealed that some regions of chromosomes are the hotspots for LTR retrotransposons activity and proliferation.

8.2.4.1 Structural features of the Gypsy superfamily

The structural features of all Gypsy elements identified in present study were analysed in detail. *MaGYP1* is 4982 bp in size, flanked by 5 bp TSD 5′-CCCGG-3′ and LTRs of 505 bp. The element has an internal region containing PBS, protein coding domains of *gag-pol* genes (GAG-AP-INT-CHR). The RT and RH domains are not detected or might be deleted from the element during rearrangement of the element. *MaGYP2* is a non-autonomous retrotransposon of 3.8 kb with 5 bp TSDs, flanked by 543 bp 5′LTR and 527 bp 3′LTR. The internal region displays PBS, GAG and Transcriptional regulator (TR) protein but lacking the RT, RH and INT domains. *MaGYP3* and *MaGYP4* belong to the same family and are 4.5 and 4.6 kb in size and are flanked by LTRS of 458 bp. They have a PBS and PPT tracts but lacking the protein domains necessary for their transposition. They have incorporated some other proteins like Transcriptional regulator (TR), Haemthiolate proteins (HP), Tymovirus proteins (TVP), Hepadnavirus proteins (HVP) and a family of proteins with high proportion of positively charged amino acids (APC), where APC is detected only in *MaGYP4*. *MaGYP5* is the close member of this family that is 6.25 kb in size, flanked by LTRs of 586 bp and have an internal region lacking the conserved protein domains present in autonomous LTR retrotransposon for their transposition. It encodes some other proteins called Major Facilitating System (MFS), Transcriptional regulator (TR), Tymovirus proteins (TVP) and a large tandem repeat (Figure 8.6; Table 8.1 & 8.2).

MaGYP6 is the smallest Gypsy and is non-autonomous due to an internally deleted *pol* gene region (Figure 8.6). It is only 3 kb in size, terminated by 5 bp TSDs and flanked by 655 bp LTRs. The internal coding domains are deleted during the reorganization of the element. The PBS was not detected while blasting against *Zea mays* and *Oryza sativa* tRNA databases. Only a GAG region can be identified by screening the sequence against protein database in CDD of NCBI. A 15 bp PPT towards C-terminal is detected which is purely composed of guanine nucleotide bases. *MaGYP7* and *MaGYP9* are 5.3 kb in size and belong to the same family; characterized by 5 bp TSDs, flanked by LTR ranging from 411-519 bp. Small insertions can be observed in 5′LTR of *MaGYP7* and *MaGYP9*.

Another element *MaGYP11* make a sister family with them. Their internal regions display PBS, two ORFs for the *gag-pol* genes, characteristics of canonical LTR retrotransposons and a PPT adjacent to 3´LTR. *MaGYP8* and *MaGYP10* are 5.9 and 5.4 kb large in sizes, flanked by 473 and 438 bp LTRs respectively. They are terminated by perfect 5 bp TSDs and display *gag-pol* poly-proteins (GAG-AP-RT-RH-INT). They also encode a well characterized chromodomain (CHR) towards the downstream of INT, which is a structural domain of chromoviridae clade of Gypsy retrotransposons (Table 8.2).

MaGYP12 displays a genome of 5.76 kb is size, terminated by 5 bp TSDs, flanked by the LTRs of 671 bp. The large size of LTRs in this element differs from the members of closely related families. It presents a PBS downstream to 5´LTR, the protein coding domains of *gag-pol* genes as 5´-AP-RT-RH-INT-3´ and PPT near to the upstream of 3´LTR with an additional Zinc knuckle (ZK) domain (Figure 8.6). *MaGYP13* is 5.4 kb defective Gypsy element that is flanked by 1062 bp LTRs, which are the largest Gypsy LTRs investigated in this study. It generates 5 bp imperfect TSDs. The internal region of the element encoding the *gag-pol* protein domains is deleted and only RT is present. The RT domain showed homology to the RT of Non-LTR elements. LINE-like element inserted into the element was tried to isolate but no clear evidence was observed. There is a possibility that LINE element fragment is incorporated in this element or the RT region have high similarity to the RT of a non-LTR retrotransposons. *MbGYP15* was identified in *Musa balbisiana*, which is 4.9 kb LTR retrotransposon, terminated by imperfect TSDs of 5 bp. It is flanked by 884 bp LTRs, and encodes *gag-pol* genes (Table 8.1 & 8.2).

A family of Gypsy showing homology to *Monkey*-like Gypsy element was also identified from the *Musa acuminata* and *Musa balbisiana*. The elements shown homology to *Monkey* element were *MbGYP16*, *MaGYP17* and *MbGYP18*, which are 4.0, 6.1 and 7.4 kb is size respectively and terminated by 4 bp TSDs. *MbGYP18* showed the structural features of LTR retrotransposons encoding the *gag-pol* protein domains. *MaGYP17* is the complete element encoding CSP-GAG-AP-RT-RH-INT-CHR domains. CSP is an additional protein domain inserted in the *gag* gene, which is not observed in other members of the family. *MbGYP16* encodes GAG-AP-RT-RH, while only AP and RT domains are observed in *MbGYP18*. All the elements have the typical PBS and PPT structures upstream and downstream of 5´LTRs and 3´LTRs respectively. The only difference is the tRNA of *MaGYP17* is complementary to $tRNA_{Met}$, while the tRNA of *MaGYP16* and *MaGYP18* is

complementary to $tRNA_{Lys}$. *MbGYP19* is a 7.36 kb in size, flanked by 1105 bp 5′LTR and 883 bp 3′LTR. A small insertion of 265 bp was detected from the 5′LTR, which makes the two LTRs unequal in sizes. Another AT rich unknown insertion is observed next to the downstream of 5′LTR. The PBS of the element was not detected but a PPT can be observed adjacent to 3′LTR (Figure 8.6; Table 8.2).

8.2.4.2 Nested structures of Gypsy LTR retrotransposons in *Musa*

Two Gypsy-like families of LTR retrotransposons identified in this study have complex and nested structures, where 1 or two other transposons or LARD insertions are incorporated within the element. *MaGYP14* is an 11.6 kb long element, flanked by 624 bp LTRs. It is characterized by the presence of PBS, ORF containing the *gag-pol* genes. The *pol* gene exhibit domains (AP-RT-RH-INT) with PPT motif downstream to *pol* gene. It harbours a transcriptional regulator protein of ~1.7 kb and an unknown insertion of ~2.3 kb made up of simple sequence repeats, highly rich in GC (71%) towards the C-terminal of the element (Figure 8.6). The most complex structure among the Gypsy elements was observed in *MbGYP20*, identified in *Musa balbisiana* accession (AP009325.2). The size of *MbGYP20* is 17.8 kb, which shows a nested structure of 3 insertions and 2 solo LTRs (Figure 8.6). One insertion is 9.6 kb Gypsy-like retrotransposon, in which another unknown insertion of 4.5 kb is inserted in opposite orientation (3′-5′) and duplicating the size of original LTR retrotransposon. This unknown insertion is terminated by ‘TA’ target site duplication and flanked by perfect 162 bp terminal inverted repeat, indicating the characteristic features of Tc1-mariner transposon. But no clear evidence was found due to the lack of any transposase and homology to the known elements. The insertion encodes a protein of unknown function (DUF), a ZK domain and a cauliflower mosaic virus peptidase (CMV). By removing this insertion from the element, where it is inserted, a 5.6 kb large retrotransposon can be isolated. The detail structural and comparative analysis of this element indicated that this is *MaGYP10*-like element, inserted in *MbGYP20*. Two solo LTRs of 1.4 kb are also inserted in *MbGYP20*; one in the last fragment of 4.5 kb Tc1-mariner-like insertion and the other in the upstream of CMV domain of outer most retrotransposon. The outermost LTR retrotransposon (*MbGYP20*) is flanked by 351 bp 5′LTR and 393 bp 3′LTR.

8.2.4.3 The *gag-pol* polyprotein organization in Gypsy elements

The organization of *gag-pol* genes protein domains of 20 individual Gypsy retrotransposons revealed 2 pattern (complete & incomplete) and 14 sub-patterns of domain organization (Table 8.2). The canonical domain structure of Gypsy is 5′GAG-RT-RH-INT-3′, which was observed in 6 elements. A single element *MaGYP6* encodes a *gag* protein only, *MaGYP2* showed *gag* and a transcriptional regulator (TR), *MaGYP13* encodes RT only and *MbGYP18* encodes 5′-AP-RT-3′. The five elements *MaGYP7*, *MaGYP9*, *MaGYP11*, *MaGYP15* and *MaGYP16* lack the INT domain. *MaGYP1* is characterized by displaying a domain pattern 5′-GAG-AP-INT-CHR-3′ but lacking the RT and RH domains. *MaGYP3*, *MaGYP4* and *MaGYP5* showed different protein domain organizations. Three elements *MaGYP8*, *MaGYP10* and *MaGYP17* showed more or less similar domain pattern 5′-GAG-AP-RT-RH-INT-CHR-3′ with one or the other extra domain in them. The nested element *MbGYP20* showed a complex organization of protein domains 5′-GAG-AP-(3′-CMV-RH-DUF-5′)-DUF-CMV-RT-RH-CHR-3′ (Figure 8.6).

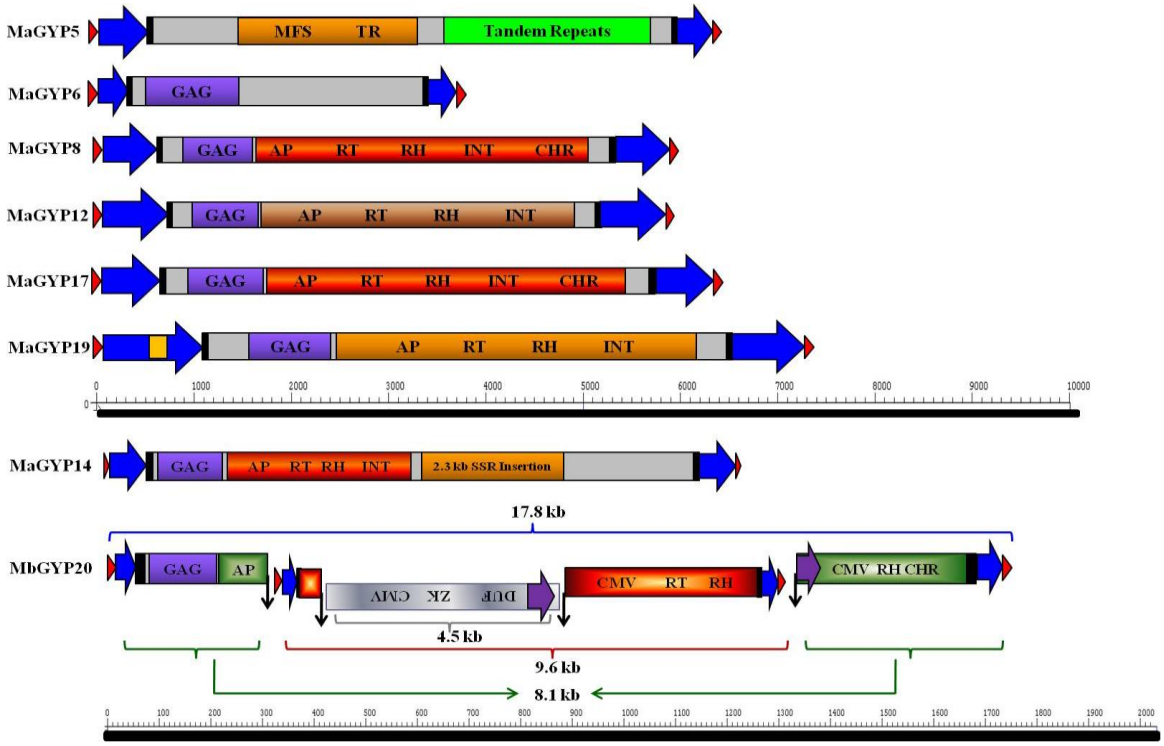


Figure 8.6: Schematic representation of Gypsy retrotransposons in *Musa*. Red arrowheads represent TSDs, while blue arrows indicate TIRs. The *gag* and *pol* regions are drawn with their protein domains. The scale below shows length in bp. Additional insertions or unknown sequences are represented by different colours. A 17.8 kb large nested *MaGYP20* is drawn with other inserted Gypsy element (9.6 kb) having another DNA transposon (4.5 kb) inserted in it. AP: Aspartic protease. RT: reverse transcriptase. INT: integrase. GAG: gag-nucleocapsid. ZK: zinc knuckle. DUF: domain of unknown function. CHR: Chromatin organization modifier. CMV: Cauliflower mosaic virus. UN: unknown. MFS: Major facilitating factor. TR: Transcriptional regulator.

Table 8.2: List of *Musa* retrotransposons with PBS, PPT motifs and *gag-pol* gene protein domains. The abbreviations used for domains are written in abbreviations. UD: Undetermined.

Element name	tRNA type	PBS (5´-3´)	Position	PPT (5´-3´)	Position	Domain organization (5´-3´)
MaGYP1	Met*	TATCAGAGCAGCGATCTT	516-533	ATGAGGAGCTGAAGA	4394-4408	GAG, AP, INT,CHR
MaGYP2	Asn	CGCTAGAAGGAGGGC	560-574	UD	---	GAG,TR
MaGYP3	Asn	CGCTAGAAGGAGGGC	470-484	ACGGACCAGGGAGAA	4012-4026	HP,TR,TVP,HVP
MaGYP4	Asn	CACTAGAAGGAGGGC	472-486	ACGGACCAGGGAGAA	4072-4086	DUF,HP,APC,HVP
MaGYP5	Ala	GGAGCTATGCGTCGGTTC	612-619	AGGAGAAAGCTAACG	5605-5619	MFS,TR,TVP,
MaGYP6	UD	---	---	GGGGGGGGGGGGGGG	2331-2345	GAG
MaGYP7	Pro	TCGAGGCTGACGATTC	497-512	GGAAGGGCAGCGAGA	4869-4883	GAG,AP,RT,RH
MaGYP8	Met	TATCAGAGCAGCGTT	484-499	ATGAGGAGCTGAAGA	5351-5365	GAG,AP,RT,RH,INT,CHR
MaGYP9	Met	TATCAGAGCAGCGTTCTTG	492-511	TGAAGAGGGCGGGTT	4794-4808	GAG,AP,RT,RH
MaGYP10	Met	TATCAGAGCAGCGTT	468-483	TGAAGAGGGCGGGTC	4977-4991	GAG,TIM,AP,RT,RH,INT,CHR
MaGYP11	Leu	TCATGAATTTTGGGAATTTG	555-574	GGAAGGGCAGCGAGA	4792-4806	GAG,AP,RT,RH
MaGYP12	Ala	TGGAGATGACGCTGAGTCG	754-772	AGACTTGAGGACAAG	5049-5063	GAG,ZK,AP,RT,RH,INT
MaGYP13	Leu	AACATACCACTCTGCAGC	1076-1093	TCATTCTTCTATGTT	4334-4348	RT
MaGYP14	Asn	CGCTAGAAGGAGGGCCT	636-652	TTCAGGGGGGGAATA	10962-10976	GAG,AP,RT,RH,INT
MbGYP15	Thr*	CCAACTAAGTTAGGAATTG	893-911	GCATGAAGAAGGAGA	3968-3982	GAG,AP,RT,RH
MbGYP16	Lys	TTCACCATGGCAAAGCATTG	349-368	TGAGTAATTGTTTAT	3729-3744	GAG,AP,RT,RH
MaGYP17	Met	TATCAGAGCCAGGTT	803-817	GACATGAAGAAGAAG	5568-5582	CSP,GAG,AP,RT,RH,INT,CHR
MbGYP18	Lys	TCTCACCATGCGAAGCACCT	431-452	AAGTTGGGGAGAATA	6673-6687	AP,RT
MbGYP19	UD	---	---	CGAGGAAAGAGGGAA	6516-6530	GAG,AP,RT,RH,INT
MbGYP20	Met	TATCAGAGCAGCGTT	362-376	TGAAGAGGACGGGTC	17392-17406	GAG,AP,(CMV,RH,DUF)*DUF,CMV, RH, CHR
MaCOP1	Met	TATCAGAGCGGGGTTTTG	616-633	AAGAAAGACAGGAGA	4589-4603	GAG,AP,INT,RT,RH
MaCOP2	Met	TATCCAGCATGTCAAGTTTC	388-407	AGGAAGAGGCCATAG	4407-4421	GAG,INT,RT,RH
MaCOP3	Arg*	CGACCTTGCATATGATCG	311-328	AAGAGAAAGGAAGAA	15883-15897	(GAG,AP,INT,RT,RH)*, GAG,INT,RT, RH
MaCOP4	Met*	ATCTGATCTAAGAGTTTTG	262-280	GGAAGAACAAGAAAA	3706-3720	GAG,ZK,INT
MaCOP5	Met	TATCAGAGCAAGGTTATC	1296-1313	CAAAAAGGGGGAGAT	6938-6952	GAG,INT,RT,RH
MaCOP6	Met	TATCAGAGCCAAGTTATT	486-503	UD	---	RT
MaCOP7	Thr*	AGGCTTCGTGAGTGAGTCG	229-247	GGGGTTGGAGAGGGA	4779-4793	GAG,INT,RT,RH
MaCOP8	Cys	TGCCATGAAAATGATTTG	561-579	GACCAAGTGGGAGAA	5501-5515	GAG,INT,RT,RH
MaCOP9	Ser	GATGCCTGAATGATTCG	585-601	GGCCAAGTGGGAGAA	6410-6424	RT
MaCOP10	Met	TATCAAAGCCAAGTTGTTCG	1609-1628	AGGTCAAGTGGGAGA	7150-7164	GAG,INT,RT,RH
MaCOP11	Met	TATCAGAGCCAGGTT	1504-1518	UD	---	GAG,INT,RT,RH
MaCOP12	Val*	TATTAAATATGACATACAAA	1207-1226	AGAAAAAAGCTTAAA	5903-5917	GAG,INT,RT,RH
MaCOP13	Val*	TATTAAATATGACATACAAA	618-637	AGAAAAAAGCTTAAA	5314-5328	GAG,INT,RT,RH
MaCOP14	UD	---	---	AAGAAGAAACCAAAA	5703-5417	GAG,INT,RT,RH
MbCOP15	Met	TATCAGAGCCTAGTTTCG	461-478	AGAAGGTGGAGCAAG	4483-4497	GAG,INT,RT,RH
MaCOP16	Val*	ATTCACCATAGAGGCCACAA	443-463	GAACAAGTGGGGGAT	4967-4981	GAG,RT,RH,MT*
MaCOP17	Val*	TATTGAGATAAAGCAAA	1398-1414	AAATCAAATTGAGAG	7043-7057	GAG,INT,RT,RH
MbCOP18	Sup	GTATCAGAGTGAGGCTC	1424-1440	CAAAAAGGAGAAGAT	8464-8478	GAG,INT,RT,RH,PRK
MbCOP19	Lys	GCCCACAAGGGAGGCT	625-640	AAATACAAAATTAAA	4571-4585	GAG,INT,RT,RH
MaCVI	Gly	TGCAAAAGGCCAAGGAATT	3918-3937	GAGCTGGGTAGCGGA	7172-7186	ZK,AP,RT,RH,DUF
MaLAR1	UD	---	---	ATAAGTGGGGGAGAA	4561-4564	UD
MbLAR2	UD	---	---	ATAAGTGGGGGAGAA	3965-3979	UD
MbLAR3	UD	---	---	ATAAGTGGGGGAGAA	4051-4065	UD
MaLAR4	UD	---	---	UD	---	UD
MbLAR5	UD	---	---	ATAAGTGGGGGAGAA	4049-4063	UD
MbLAR6	UD	---	---	ATAAGTGGGGGAGAA	4049-4063	UD
MaLAR7	Asp	GGGACCTAACGGGGCTGCG	505-523	ATAAGTGGGGGAGAA	4107-4121	UD
MaLAR8	Leu*	TGGTATCAGAGTGGGAT	442-458	AATAAGTGAGGGAGA	4088-4102	UD
MbLAR9	UD	---	---	UD	---	UD
MaLAR10	UD	---	---	UD	---	ADM

8.2.4.4 PBS and PPT pattern of Gypsy elements

The 15-18 bp priming binding site (PBS) located immediately downstream to the 5´LTR and a reverse compliment called Polypurine tract (PPT) located adjacent to the 3´LTR were detected by scanning the retrotransposon sequences against *Zea mays* tRNA database using parameter 'Predict PBS by using tRNA database'. A total of 80% and 75% elements showed the presence of 14-18 bp PBS and 15 bp PPT respectively, 10% showed 15 bp PPT only adjacent to 3´LTR. Remaining 10% sequences failed to detect any PBS or PPT by scanning tRNA of *Zea mays*, which were than scanned against *Oryza sativa* tRNA database and their PBS and PPT were successfully achieved. *MaGYP2* lack PPT, while PBS was not detected in *MaGYP6* and *MbGYP19*. Seven different tRNA types were investigated in Gypsy elements. The most frequently use type was $tRNA_{Met,}$ which was present in 30% of the elements investigated. The second important primer type was $tRNA_{Asn,}$ occurred in 20% of the elements. 10% elements showed no evidence of PBS, either deleted or have no homology with the reference elements (Table 8.2).

Table 8.3: List of primers to amplify the RT region of Gypsy, Copia and Pararetrovirus (PRV) elements.

No.	Superfamily	TE Family	Product size	Primer name	Primer Sequence
1	Gypsy	*MaGYP8*	684	MaGYP8F MaGYP8R	CTTCTCGGCAACATGACCA GGTCTACCGCCACTCCTTC
2	Gypsy	*MaGYP10*	897	MaGYP10F MaGYP10R	CATCTGCAACGAACATTCTC CTTTTCATCGGTGCTACTTG
3	Gypsy	*MaGYP11*	550	MaGYP11F MaGYP11R	ACAGGAGTTATTCGGCCAAG TCATGGCTGCCTTAAGTTTG
4	Gypsy	*MaGYP12*	830	MaGYP12F MaGYP12R	CCAATTCCCACATTAGATGC GAGAGCATGAGTCATTGTGC
5	Gypsy	*MaGYP13*	758	MaGYP13F MaGYP13R	GGGAGACCCAAATAAGGAAC CAGGGGCTGATTCACTAGAG
6	Gypsy	*MaGYP17*	835	MaGYP17F MaGYP17R	GCAGCTCAAAAGCACCTTTC CCAATAGCAAAGTCCGAAGC
7	Copia	*MaCOP5*	744	MaCOP5F MaCOP5R	CTTAGTCGCAGTACTCATAG TGGAAGCTTGTTCCTAGACC
8	Copia	*MaCOP7*	748	MaCOP7F MaCOP7R	GTTTGGACGGGTGAAAGCTA CGGACCATTCCTCATCAAAG
9	Copia	*MaCOP8*	964	MaCOP8F MaCOP8R	CTTTCACAATGGGAGCAACA GTTGAACCACAAGTTCCTCA
10	Copia	*MaCOP9*	720	MaCOP9F MaCOP9R	GCGACTCAAAGGACAATATC GAGCATAGACTTCCAACTAC
11	Copia	*MaCOP10*	752	MaCOP10F MaCOP10R	CCCATGTCTTATCGGAATGA CCTCCGGAGAGATATGTGAG
12	PRV	*MACV1*	425	MACV1F MACV1R	CAACTACAAGAGGCTGAACG CTATTTCCTTGACTGCTATC

8.2.4.5 RTAP markers to study diversity of Gypsy elements in *Musa* genomes

The distribution and abundance of Gypsy LTR retrotransposons in *Musa* genome were performed by reverse transcriptase amplification polymorphism (RTAP) by PCR among 48 *Musa* accessions. The primers were designed from conserved regions of RT around D-DD triad. Of the forty eight *Musa* genomes (Table 2.3), 6 were *Musa acuminata* (AA), 6 were *Musa balbisiana* (BB), 3 were hybrids (AB), 8 were triploid *Musa acuminata* (AAA), 19 were allotriploids (AAB) and 6 (ABB). The primer pair MaGYP8F and MaGYP8R (Table 8.3) was used to amplify 684 bp RT regions of *MaGYP8* family. The products were amplified from *Musa acuminata* (AA) (Calcutta 4, Sannachenkadali, Pisanglilin, Kadali, Matti, Cherukadali), *Musa balbisiana* (BB) (PKW1, PKW2, Javan, Klutuk, Tani, Batu), AB genome (Njalipovan, Adukkan, Padalamukili), AAA genome (Manoranjitham, Grandnain, Grow-michel, Greenred, Red, Monsmari, Robusta, Dwarf Cavendish), AAB genome (Motta povan, Karimkadali, Perumadali, Kunoor ettan, Palyamcodan, Mysoreettan, Krisnavazhai, Poovan, Doothsagar, Charapadati, Kumbillakannan, Velipadati, Vellapalayamcodan, Ettapadati, Padati, Chinali, Nendran, Poomkalli, Kamaramasengi) and ABB genome (Kosta bontha, Peyan, Kanchikela, Boothibale, Monthan, Karpooravali). This showed that the element is ancient, was present in a common ancestor predating the separation of A and B-genome *Musa* (Figure 8.7a).

The abundance of *MaGYP10* family in *Musa* genomes was investigated by primers MaGYP10F and MaGYP10R to amplify 897 bp RT regions. All the 48 accessions yielded the product indicating its high distribution and diversity. A 550 bp RT region from *MaGYP11* family was amplified by MaGYP11F and MaGYP11R from all the 48 accessions suggesting its mobility in all genomes. The RT based amplification polymorphism of *MaGYP12* family revealed that, it is amplified from all the 47 accessions with the exception of *Musa acuminata* (Calcutta 4), where no amplification suggests its absence in the genome. The 835 bp RT regions from *MaGYP17* family were amplified from all the 48 *Musa* accessions by primer pair MaGYP17F and MaGYP17R. Very strong bands from all the genomes tested suggest its high amplification and proliferation in *Musa*. The amplification of Gypsy from the *Musa* genomes revealed that they are highly abundant, distributed in all genomes and proliferating actively in all genomes regardless of A or B-genomes specificity (Figure 8.7b-e).

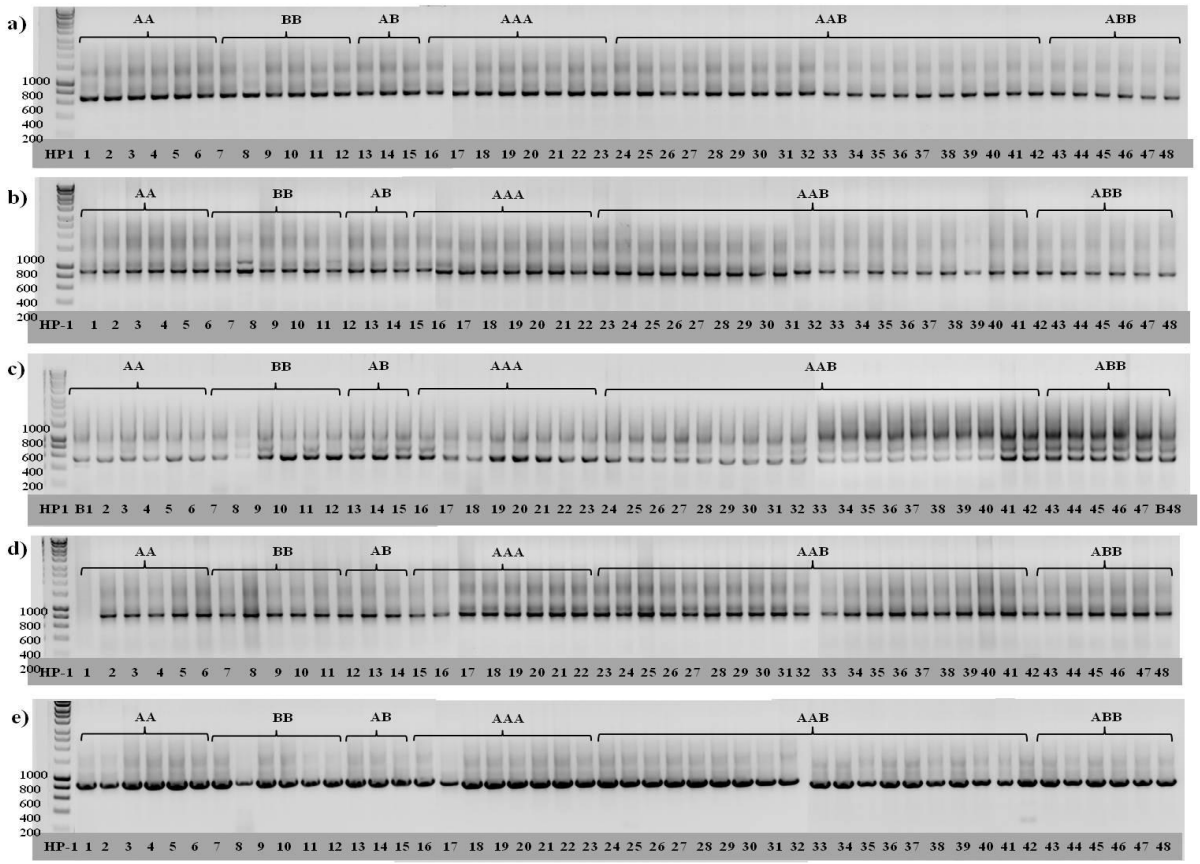


Figure 8.7: Detection of Gypsy RT polymorphisms across 48 cultivars in *Musa*: PCR amplification of a) *MaGYP8*; b) *MaGYP10*; c) *MaGYP11*; d) *MaGYP12*; e) *MaGYP17*. All PCR figures show reversed images of size-separated ethidium bromide-stained DNA on agarose gels after electrophoresis; ladders (left) show fragments sizes in base pairs; numbers at the base indicate accessions defined in Table 2.3.

8.2.5 Structural features of Copia superfamily

Ninteen Copia LTR retrotransposons were identified from the screened *Musa* BACs sequences by comparison in dot plot analysis. The molecular structures were studied in detail. *MaCOP1* and *MbCOP19* belongs to the Copia lineage of LTR retrotransposons and are grouped in sister families having homology in their conserved domains. They are 5.3 and 5.2 kb in size, flanked by LTRs of 605 and 592 bp respectively. Both encode the conserved protein domains of 5′-INT-RT-RH-3′, features indicating an autonomous LTR retrotransposons encoding the necessary proteins for the transposition process. *MaCOP1* encodes an AP domain, which is not detected from *MbCOP19* suggesting its deletion during the rearrangement of element during the evolutionary phases. They also have PBS and PPT necessary for the transposition of RT (Table 8.2). *MaCOP2* is 4.8 kb large element having the characteristic features of Copia-like retrotransposons. It generates 5 bp TSDs upon insertion and is flanked by 358 bp 5′LTR and 360 bp 3′LTR. It exhibits a PBS,

gag-pol genes to encode 'GAG-INT-RT-RH' and PPT upstream of 3′LTR. Another defective element *MaCOP4* was identified from *Musa acuminata* (AC226038.1), which is 4.0 kb in size, flanked by 263 bp LTRs. It encodes only INT domain from the *pol* gene while other domains are deleted. A ZK motif is encoded in the element towards upstream of INT domain in *gag* gene (Figure 8.8).

MaCOP5 and *MaCOP17* are similar in size and are ~8.1 kb in size. They are flanked by large LTRs of 5′-1285/1201-3′ and 5′-1000/1324-3′ respectively. The difference in the number of nucleotides in LTRs is due to the insertions or deletions during the evolutionary stages. They show evidence of PBS and PPT in upstream and downstream of 5′LTR and 3′LTR respectively. They have shown typical Copia *gag-pol* gene polyproteins (Table 8.1 & 8.2). *MaCOP6* and *MaCOP9* though ~7.0 kb in sizes do not code the *pol* polyproteins except RT, which suggests that these defective Copia have lost their conserved domains during the evolutionary stages. *MaCOP7* is a 5.0 kb large element terminated with 5 bp TSDs. It is flanked by 5′-144/149-3′ bp LTRs, the shortest LTRs investigated in present study. It is characterized by the integration of PBS, *gag-pol* gene encoding the GAG-INT-RT-RH domains and a PPT motif adjacent to 3′LTR. *MaCOP8* and *MaCOP14* are 6.0 kb in sizes, are flanked by LTRs of 5′-499/500-3′ and 5′-492/548-3′ bp respectively. *MaCOP8* have the PBS and PPT, while no PBS was identified from *MaCOP14* but a PPT is present near the upstream of 3′LTR.

MaCOP10 and *MaCOP11* are 8.7 and 8.4 kb long in sizes, flanked by the longest LTRs from Copia retrotransposons investigated so far in this work. *MaCOP10* is terminated by 1597 bp LTRs while *MaCOP11* is flanked by 5′-1494/1388-3′ LTRs. Their domain organization and structural features indicate that they have PBS, pol genes coding for INT, RT and RH domains. *MaCOP12* and *MaCOP13* are 7.1 and 5.9 kb respectively. They are the members of the closely related families, but display highly variable LTRs. *MaCOP12* and *MaCOP13* are flanked by 5′-1238/1132-3′ and 5′-573/548-3′ bp respectively. Their genomic organization showed the typical Copia *gag-pol* structure (GAG-INT-RT-RH). Both elements have homologous or exactly similar PBS and PPT sequences indicating their origin from a common ancestor. *MbCOP15* and *MaCOP16* are ~5.0 and 5.4 kb Copia investigated in *Musa balbisiana* (AC226053.1) and *Musa acuminata* (AC226040.1) BAC sequences respectively. The former is flanked by 449 bp LTRs while lateral has 5′-406/438-3′ bp LTRs. Their internal region displays the ORF encoding the *gag-pol* products

(GAG-INT-RT-RH) and other features of Copia elements such as PBS and PPT. *MbCOP18* is a 9.8 kb Copia retrotransposon investigated in *Musa balbisiana* accession 'AC226052.1'. It generates a 4 bp TSD with a single bp mismatch and terminated by long 5′-1415/1396-3′ bp LTRs (Figure 8.8; Table 8.1 & 8.2).

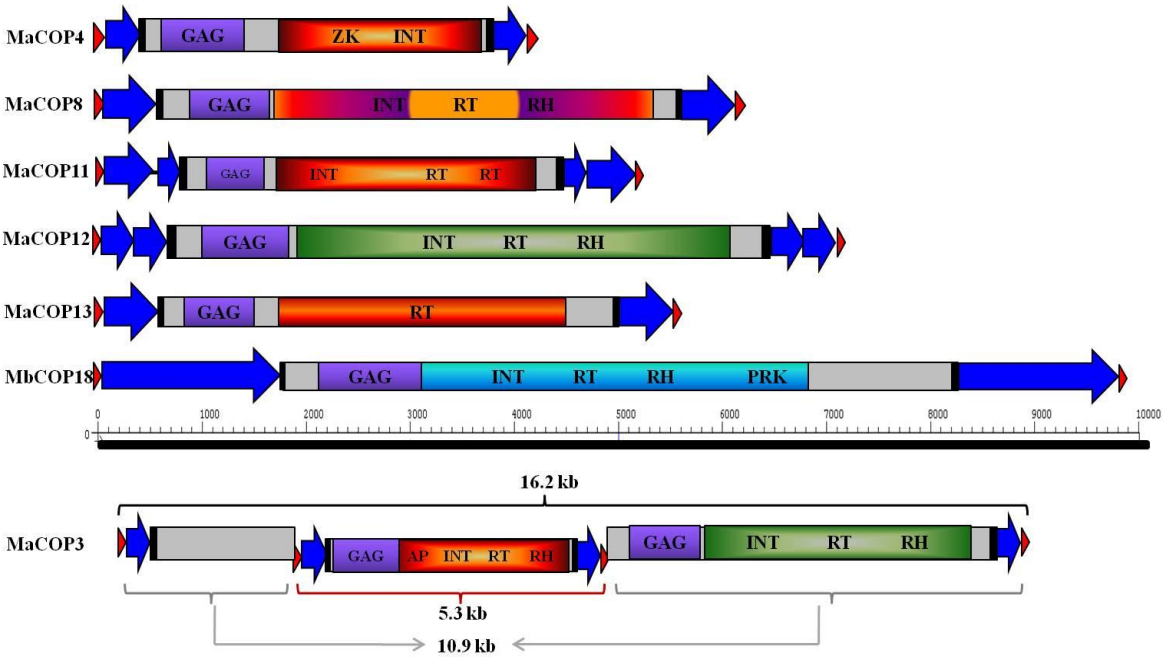


Figure 8.8: General structures of different Copia elements in *Musa*. The red arrowheads at ends represent the TSDs. TIRs are represented by blue arrows. The *gag* and *pol* regions are drawn with their protein domains. The scale below shows lengths in bp. A 16.2 kb MaCOP3 has a 5.2 kb Copia element inserted in it. GAG. *gag*-nuclocapsid. AP: Aspartic protease. RT: reverse transcriptase. INT: integrase. ZK: zinc knuckle. DUF: domain of unknown function. AIR1: Arginine methyltransferase-interacting protein. PRK: Hypothetical protein. UN: unknown.

8.2.5.1 Nested structures of Copia LTR retrotransposons in *Musa*

A 16.2 kb sequence from *Musa acuminata* BAC sequence (AC226035.1) from position 75834-92033 bp was identified comprising a nested structure of LTR retrotransposons. A 5.3 kb *MaCOP1* element is inserted in *MaCOP3* starting from 3203-8492. The outer element is 10.9 kb in size, where another 5.3 kb insertion increased the size of retrotransposon to 16.2 kb. Both *MaCOP1* and *MaCOP3* are in 5′-3′ orientation and are members of two different families. The outer element *MaCOP3* is flanked by 5′-338/299-3′ bp while the inserted element *MaCOP1* is terminated by 605 bp LTRs indicating that both are Copia elements belonging to two different families. *MaCOP3* generated TSDs of 4 bp and is characterized by the presence of *gag-pol* genes coding domains for GAG-INT-RT-RH and exhibiting the PBS next to 5′LTR and PPT prior to 3′LTRs. The GC

proportion of outer and inserted Copia retrotransposon is is 43.8% and 41.3% respectively. The internal *MaCOP1* displayed the perfect *gag-pol* poly-proteins structures (Figure 8.8).

8.2.5.2 The *gag-pol* domain organization in intact Copia elements

The protein domains encoded by *gag-pol* gene revealed that seven different sub-patterns of the two main patterns (canonical and defective) were observed from 19 intact Copia-like elements identified in present study. The canonical pattern of protein domains for Copia retrotransposons is 5′-GAG-INT-RT-RH-3′, which was observed in 90% of the elements with one less or additional domains. Several different arrangements of *pol* genes were detected from these elements. *MaCOP6* and *MaCOP11* encode only a RT domain and no other domain was detected from them. The domain organization in *MaCOP1* was 5′GAG-AP-INT-RT-RH-3′ while *MaCOP17* showed a slight different pattern 5′-GAG-RT-RH-MT-3′ without encoding the INT domain, where MT is a Mannosyl transferase protein. A nested LTR retrotransposon *MaCOP3* showed a complex pattern 5′-GAG-AP-INT-RT-RH/GAG-INT-RT-RH-3′, where two sets of proteins domains are detected encoded by the *gag-pol* genes of 2 different Copia retrotransposons. Rest of 13 elements showed the same protein organization 5′-INT-RT-RH-3′, which are the characteristic features of Copia-like retrotransposons (Table 8.2).

8.2.5.3 PBS and PPT organization of Copia elements

The PBS and PPT structures of copia-like retrotransposon sequences were detected by scanning them against *Zea mays* and *Oryza sativa* tRNA database using parameter 'Predict PBS by using tRNA database' in LTR_FINDER. A total of 95% elements showed the presence of 14-18 bp PBS next to the 5′LTR, while *MaCOP14* only failed to identify its PBS sequence. Eight different types of tRNA types were observed in all the Copia-like elements investigated. Like the Gypsy elements, the most frequent tRNA type in Copia elements was $tRNA_{Met,}$ detected in 40% of the elements; the second important type was $tRNA_{Val,}$ observed in 20% elements. All the other 6 types of tRNA contributed 5% of the tRNA type. PPT adjacent to the 3′LTR was detected in 90% of all Copia elements. *MaCOP6* and *MaCOP14* failed to show any evidence of PPT in their genomic sequence. The *MaCOP6* is a defective element, who has deleted the *gag-pol* protein domains except RT during the rearrangement of their genome. All the other elements have 15 bp PPT sequence towards the downstream of 3′LTR (Table 8.2).

8.2.5.4 Diversity and distribution of Copia in *Musa*

The mobility and proliferation of various Copia families were investigated in 48 *Musa* accessions (Table 2.3) by PCR analysis. The data revealed that the Copia are actively proliferating in *Musa* genomes and are highly abundant. Of 10 primer sets, the results of 4 pairs are discussed here. Out of 292 potential products, 287 products were achieved resulting 98.8% yield. The primer pair MaCOP5F and MaCOP5R (Table 8.3) was designed to amplify a 744 bp RT region, which was yielded by *Musa* accessions including the *Musa acuminata* (AA) (Calcutta 4, Sannachenkadali, Pisanglilin, Kadali, Matti, Cherukadali), *Musa balbisiana* (BB) (PKW1, PKW2, Javan, Klutuk, Tani, Batu), AB genome (Njalipovan, Adukkan, Padalamukili), AAA genome (Manoranjitham, Grandnain, Grow-michel, Greenred, Red, Monsmari, Robusta, Dwarf Cavendish), AAB genome (Motta povan, Karimkadali, Perumadali, Kunoor ettan, Palyamcodan, Mysoreettan, Krisnavazhai, Poovan, Doothsagar, Charapadati, Kumbillakannan, Velipadati, Vellapalayamcodan, Ettapadati, Padati, Chinali, Nendran, Poomkalli, Kamaramasengi) and ABB genome (Kosta bontha, Peyan, Kanchikela, Boothibale, Monthan, Karpooravali). This amplification of *MaCOP5* is all genomes suggest its high diversity, mobility and high proliferation rate within *Musa* accessions (Figure 8.9a).

Several other families of Copia were amplified from *Musa* genomes with different primer pairs. The amplification polymorphism of *MaCOP7* yielded the expected bands, with additional bands of varied lengths. No amplification in *Musa balbisiana* accession ‘Batu’, *Musa acuminata* triploid (Grandnain; AAA) and an allotriploid accession (Perumadali; AAB) indicate their absence in the genomes. A 964 bp *MaCOP8* RT amplicons were amplified from all *Musa* genomes, but no amplification in *Musa acuminata* accession ‘Calcutta 4’ suggests its absence from the genome. The 752 bp products from *MaCOP10* were amplified from all 48 *Musa* accessions with ~700 additional bands in A-genome and its triploids (Figure 8.9b-d).

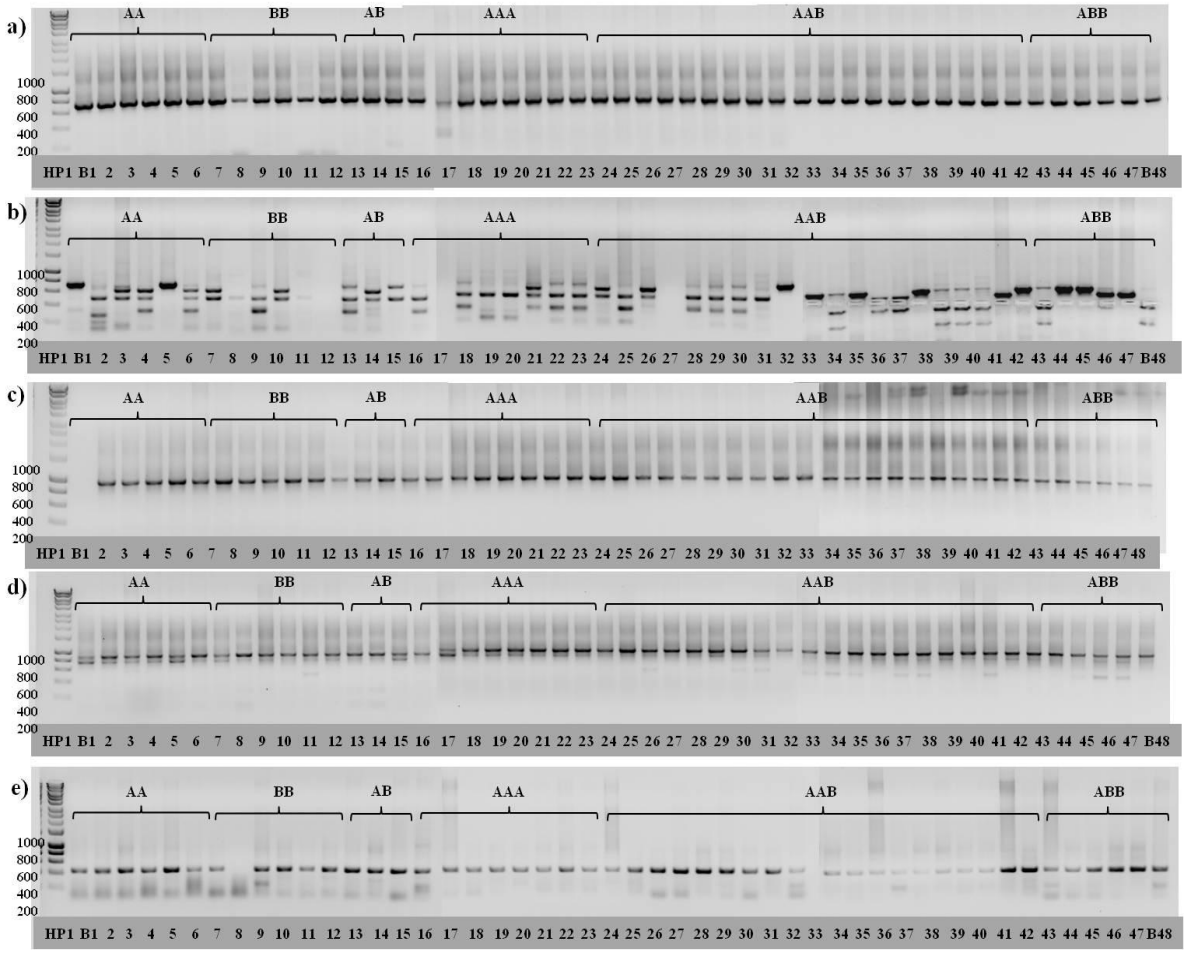


Figure 8.9: PCR showing fragments with Copia and PRV RT regions. DNA samples were obtained with primers hybridizing to conserved RT regions of various Copia and PRV families a) *MaCOP5*; b) *MaCOP7*; c) *MaCOP8*; d) *MaCOP10*; e) *MACVI*.

8.2.6 Structural features and diversity of a Pararetrovirus-like element in *Musa*

During this study, a 11.1 kb long element was investigated from *Musa acuminata* BAC 'AC226046.1' from position 160034-1711104 bp. It was flanked by 3.8 kb of LTRs, the largest of all the elements investigated in *Musa* genomes. These LTRs were larger than LTRs studied from any member of Gypsy or Copia-like retrotransposons. The comparative analysis with the known LTR retrotransposons identified a group of Pararetrovirus-like elements. The element was named as *MACVI* (*Musa acuminata* chromovirus). It was characterized by having 3.8 kb LTRs, and an internal region containing the PBS, *pol* genes encoding the AP, RT and RH domains and a PPT adjacent to 3´LTR. Two other proteins were also encoded in its internal region. A ZK domain was detected in the start and a protein domain of unknown function (DUF) towards the end of *pol* gene. One intact and a truncated copy were identified from the *Musa* BAC clone sequences available in NCBI

database. The element showed a low GC ratio (42.3) as compared to AT (47.7%), which is constant and exactly similar in LTRs and internal regions. The PBS of *MACVI* is also different from all other members from Copia and Gypsy elements, with $tRNA_{Gly,}$, which is not observed in any other element investigated in this study. A 15 bp PPT was also found near the upstream of 3´LTR, with different sequence structure from other investigated elements.

The diversity and abundance of *MACVI*-like elements was examined by PCR analysis. The primer pair MACVIF and MACVIR was designed from conserved RT regions to amplify a 425 bp RT product. The analysis revealed that the product was amplified from *Musa acuminata* (AA) (Calcutta 4, Sannachenkadali, Pisanglilin, Kadali, Matti, Cherukadali), *Musa balbisiana* (BB) (PKW1, Javan, Klutuk, Tani, Batu), AB genome (Njalipovan, Adukkan, Padalamukili), AAA genome (Manoranjitham, Grandnain, Grow-michel, Greenred, Red, Monsmari, Robusta, Dwarf Cavendish), AAB genome (Motta povan, Karimkadali, Perumadali, Kunoor ettan, Palyamcodan, Mysoreettan, Krisnavazhai, Poovan, Doothsagar, Charapadati, Kumbillakannan, Velipadati, Vellapalayamcodan, Ettapadati, Padati, Chinali, Nendran, Poomkalli, Kamaramasengi) and ABB genome (Kosta bontha, Peyan, Kanchikela, Boothibale, Monthan, Karpooravali). Only one *Musa balbisiana* accession 'PKW2' showed no amplification, otherwise amplification in all other 47 genomes suggests the high activity and proliferation in the *Musa* genome (Figure 8.9e).

8.2.7 Characterization and structural features of LARD-like elements

Despite of several autonomous LTR retrotransposons, a group of elements with 4-5 bp TSDs, flanking LTRs and internal non-coding regions was identified which displayed the PBS and PPT motifs downstream and upstream to 5´LTR and 3´LTR respectively. Due to structural resemblance with LARDs, they were considered as members of LARDs. The elements have shown no homology with already known TEs but a high homology to each other, so they were placed in a major family named *Hazara*. *MaLAR1* was the first identified member from *Musa acuminata* accession 'AY484588.1'. The element is 4564 bp large in size, flanked by 4 bp TSD and 447 bp LTRs. No identifiable PBS was detected while PPT motif is traced adjacent to 3´LTR. *MbLAR2* is another homologue of *MaLAR1*, with the difference that it is proliferating in *Musa balbisiana* genomes. It is 4428 bp in size

including 445 bp LTRs at both ends with a PPT motif similar to *MaLAR1* indicating the members of same family. Two other elements with similar structural features indicating the members of the same family are *MaLAR7* and *MaLAR8*, which are 4.5 kb in size including 446 and 437 bp LTRs respectively. *MbLAR3* share a family with *MbLAR5* and *MbLAR6* with a size of 4.4 kb, displaying LTRs of 382-383 bp. *MaLAR4* is 4.3 kb including the large LTRs (5′-607/611-3′) and flanked by 5 bp imperfect TSDs. *MbLAR9* was identified from *Musa balbisiana* BAC 'AC186754.1' from 72565-80276 bp. It is 7.7 kb element flanked by 5′-626/635-3′ with no detectable PBS and PPT motifs. It exhibits an unknown insertion and a non-autonomous hAT element with two additional Solo LTRs (Figure 8.10). *MaLAR10* is the smallest LARD-like element studied here with a size of 4 kb, flanked by large LTRs (5′-974/984-3′) and 4 bp TSDs (Figure 8.10; Table 8.1).

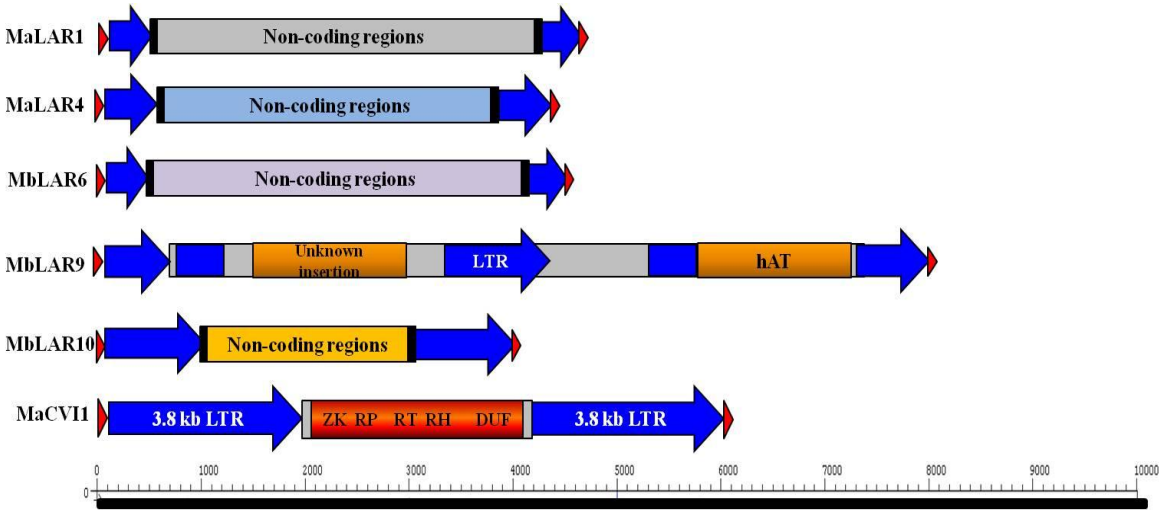


Figure 8.10: Schematic representation of LARDs and the Pararetrovirus-like element. Red arrowheads represent the TSDs, while blue arrows indicate TIRs. The internal non-coding region is represented by different colours. The scale below shows lengths in bp. AP: Aspartic protease. RT: reverse transcriptase. INT: integrase. GAG: gag-nucleocapsid. ZK: zinc knuckle. DUF: domain of unknown function. AIR1: Arginine methyltransferase-interacting protein. CHR: Chromatin organization modifier. UN: unknown.

8.2.7.1 Domain patterns and organization in LARD-like elements

All the LARD members from Hazara family were investigated for their *gag-pol* genes and the pattern of PBS and PPT. They have shown well characterized hallmarks for a LTR retrotransposons like the presence of perfect TSDs, highly homologous LTRs, PPT and in few cases PBS. No *gag-pol* genes coding the proteins were detected from any of the members. In contrast, the dotplot analysis of the *Musa* BAC clones showed the presence of these elements and their transposed copies indicating their recent movement in the

genome. On the basis of their transposed copies and high copy number, they can be considered as non-autonomous elements, defective elements or LARDs-like elements. Out of 10 individual elements, only 2 elements (20%) *MaLAR7* and *MaLAR8* display the PBS and PPT in their 5′LTR and 3′LTRs respectively. The tRNA type in *MaLAR7* was $tRNA_{Asp,}$ while in *MaLAR8* was $tRNA_{Leu.}$ No evidence of PBS was observed in other 8 elements. PPT was detected in 7 elements (70%), while *MaLAR4*, *MaLAR9* and *MaLAR10* failed to show the PPT sequence by scanning against *Zea mays* or *Oryza sativa* database (Table 8.2).

8.3 Discussion

The comparative sequence analyses have shown very fast variations in the plant genomes. One of the major sources of such rapid changes are the repetitive DNA sequences present in many genomes (Bennetzen, 2000). The genome of *Musa* is also rich in LTR retrotransposons belonging to Copia, Gypsy and Pararetrovirus-like elements. As the genomic sequence is progressing and updated, there is probably a need to discover and characterize the transposable elements. The LTR retrotransposons in the plants are terminated by few hundred base pairs to several kilobases, generating 4-6 bp TSDs and are generally terminated by dinucleotides 5′-TG….CA-3′(Kumar and Bennetzen, 1999).

To our knowledge, this study is the first detailed survey of Copia, Gypsy and LARDs-like elements in *Musa* genomes over long stretches of DNA sequence: previous analyses have focussed on selected repeats revealing that nearly 30% genome composition of repetitive sequences in *Musa* (Hribova *et al.*, 2010). The approach of comparative analysis of BAC sequences by dot plot was novel and used to identify the LTR retrotransposons in the sequenced genome of *Musa*. This strategy helped in the identification of most of the elements present in *Musa* BAC sequences. In the initial effort, 50 intact elements belonging to three main lineages of Copia, Gypsy and Pararetrovirus-like elements were identified, including evidence for mobility and hence a history of the elements. Further BLAST analysis using these full length elements retrieved a total of 153 intact elements from 6 Mbp of *Musa* BAC sequences screened. The intact copies covered 15-18% of the genome surveyed, which is further strengthening the investigations revealing high repetitive proportions found in the *Musa* genome analysis using short reads from 454 sequencing (Hribova *et al.*, 2010) and BAC-end sequencing (Cheung and Town, 2007).

About 61 truncated copies, 635 partial copies, 258 solo LTRs and 16246 small fragments (remnants) were also identified; precise alignment of truncated or partial copies is not possible due to deletions and the numbers, but their contribution to the *Musa* genomes was counted. These deleted elements, and insertions in LTR retrotransposons are common in plants like maize (Jin and Bennetzen; Ramakrishna *et al*, 2002), wheat (Wicker *et al*. 2001 2003), barley (Rostoks *et al*, 2002), Rice (Ma *et al*, 2004) and *Arabidopsis* (Devos *et al*. 2002). It was noted that most of the deletions or insertions in the intact elements were bounded by few bp terminal duplications. Such terminal duplications were observed around the deletions within retroelements from *Arabidopsis* (Devos *et al*, 2002). The percentage of partial or deleted copies, truncated elements and remnants are very high analysed in our study as compared to the full length copies. The full length elements range in size from 4 kb to 17.8 kb and the LTRs flanking them range in sizes from 149 bp to 3.8 kb. These findings are in accordance with the investigations of LTR retrotransposons in *Medicago truncatula*, where the full length elements range in size from 4-18.7 kb with more or less similar LTRs flanking the elements (Wang *et al*., 2008).

A Pararetrovirus-like element residing in *Musa* genome was investigated, which displayed the structural features common to caulimoviruses present in many plant genomes including *Musa* and potato. In potato three families of Pararetrovirus-like sequences were isolated from potato genome and their distributions on chromosomes were studied by fluorescent in situ hybridization (Hansen *et al.*, 2005). The phylogenetic analysis of the various LTR retrotransposons in *Musa* gives insight into diversity of these elements within the genome. Out of the 50 reference elements analyzed, 20 were Gypsy, 19 Copia, 1 Pararetrovirus-like element and the other 10 were the LARD elements. The RT alignment and phylogenetic analysis revealed that Pararetrovirus-like element (*MaCV1*) make a sister clade with Gypsy elements suggesting early evolutionary separation. In *Brassica*, the virus-like elements grouped with Gypsy lineage indicating their common ancestral origin but followed two different evolutionary pathways (Alix and Heslop-Harrison, 2004). The domain organization of the elements also varied, consistent with earlier studies: Copia-like elements were 5′-AP-INT-RT-RH-3′, Gypsy-like elements 5′-AP-RT-RH-INT-3′, and Pararetrovirus-like elements showed 5′-ORF-AP-RT-RH-3′ (Hansen and Heslop-Harrison, 2004).

The LARD elements were out grouped in different clade without clustering with any of the known groups of elements. In *Medicago truncatula*, 11 LARD families of elements lacking protein domains in their *gag-pol* genes were characterized (Wang and Liu, 2008). It was observed in present study that very few elements were species specific, either in *Musa acuminata* or *Musa balbisiana* and the majority were present in both and (considering that nearly twice as much BAC sequence data is available for *Musa acuminata* compared to *Musa balbisiana*) similar proportions of LTR retrotransposons were identified in both species. The PBS and PPT pattern of the various classes of elements were also investigated. The results showed that most of the elements contained both sequences in them, while in few either PBS or PPT is missing or might be deleted. The $tRNA_{Met}$ was the most frequently used type in both superfamilies. This is in accordance to the findings of tRNA type in *Medicago truncatula*, where $tRNA_{Met}$ is the most frequently used tRNA type occurred in 60-80% of the retrotransposons families investigated in *Medicago truncatula* (Wang and Liu, 2008). Some of the retrotransposons have acquired an extra sequence that does not have any role in their transposition or life cycle of the retrotransposons (Havecker *et al.*, 2004).

8.4 Conclusion

The LTR retrotransposons in *Musa* genome were identified and described by computational analysis and PCR amplification. Fourty novel families were described in detail including their structural features, protein domain organizations, pattern of PBS and PPT, classification, evolutionary dynamics and impact on their host genome by their transduplication. The total number of copies and their percentage in *Musa* genome revealed that a high proportion of *Musa* is made up of these repetitive elements. This work provided the detail analysis of LTR retrotransposons landscape in the *Musa* genomes by concluding their role in *Musa* genome duplication, diversification and evolution. Major portions of retrotransposons belonging to Copia, Gypsy and LARD superfamilies were described and annotated. This will be helpful to other workers to understand the LTR retrotransposons landscape of *Musa* and related genera and their evolutionary dynamics.

CHAPTER 9
MOLECULAR CHARACTERIZATION OF DNA TRANSPOSONS AND NOVEL MOBILE INSERTIONS IN *MUSA*

Summary

Musa is a monocotyledonous plant and this work aimed to characterize the diversity of superfamilies of TEs present in its genome. Autonomous and non-autonomous TEs belonging to different superfamilies were identified by comparing transposon-rich BAC clones of *Musa acuminata* (AA) with homeologous genomic sequence regions in *Musa balbisiana* (BB). Class II DNA transposons were abundant comprising non-autonomous members from hAT, Mariner, MITEs and few novel families of elements. By comparative genomics and PCR/gel-based assays, active autonomous copies and fossil remnants, or deleted derivatives of active members were detected. Using comparisons over >100 kb genomic regions, the present approach identified any sequence which had been inserted or deleted. Most of DNA transposons were non-autonomous and ranged in size from 82 bp to a few kilobases (kb). The transposons display hallmarks of superfamilies but some mobile insertions-deletion pairs were detected without terminal inverted repeats (TIRs), which were not reported earlier in *Musa* and exhibit structural features not observed in other known TEs such as varied TSDs and lack of any TIRs. The mobility, diversity and abundance of TEs in 96 diverse *Musa* germplasm were analysed by PCR amplification using developed TIP markers. The abundance and localization of these elements on chromosomes was studied by flourescent *in situ* hybridization (FISH) and found some TEs to be A or B-genome specific. Overall, the analysis provides insight into the nature and mechanisms of changes in abundance and diversity of TEs as an important genomic component in *Musa* genomes.

9.1 Introduction

Transposable elements (TEs) are ubiquitous components of eukaryotic genomes with an ancient history of coexistence and proliferation in their host. They represent a high fraction of total genome size in many eukaryotic species. Their impact on genome is highly noteworthy, as they perform important roles in evolution by generating genetic variability, duplications, mutations, restructuring genomes and acting as sources of new genes (Flavell

et al., 1994; Kidwell and Lisch, 2001). Their activity is regarded as a major driving factor for gene and gene evolution in various organisms (Feschotte and Pritham, 2007). These TEs are residing in the genomes as (1) autonomous elements with an active RT or a transposase required for mobilization and integration into new sites, (2) non-autonomous elements which lack the enzymes for their mobilization, depends on the autonomous partners to utilize their enzymes for proliferation and reintegration to a new site and (3) 'relics', fossil remnants or deletion derivatives, which are normally immobile sequences. Based on their structural features and hallmarks, the Class II DNA transposons are categorized into 12 major transposon superfamilies, of which, Tc1-Mariner, hAT, Mutator, PIF-Harbinger, CACTA, P and Helitron are common in plants (Wicker *et al.*, 2007).

The hAT is the most diverse family of DNA transposons, predominant in many plant species. They are characterized by 8 bp TSDs, terminal inverted repeats of 9-27 bp and an internal ORF encoding a transposase displaying six amino acids conserved blocks across the animal-fungi-plant transposases. A total of 147 hAT-related sequences in plants, animals, and fungi were studied suggesting the diversity of hATs among various eukaryotes (Rubin *et al.*, 2001). The hATs were studied in several plants like maize (Shimatani *et al.*, 2009; Fujino and Sekiguchi, 2011), sugar beet (Menzel *et al.*, 2012), *Petunia hybrida*, *Phaseolus*, *Brassica napus* (De Keukeleire *et al.*, 2004) and *Arabidopsis* (Bundock and Hooykaas, 2005).

Miniature inverted-repeat transposable elements (MITEs) are the most high copy number elements proliferating in several plant species. They have shown to represent the derivatives of most DNA transposons superfamilies including Tc1-Mariner-like (Stowaway), Harbinger-like (Tourist), hATs derived MITEs and Mutator derived MITEs (Benjak *et al.*, 2009). The MITEs display TSDs and TIRs similar to DNA transposons, but lack any transposase due to the deletion of internal region encoding ORF. Their mobilization is dependent on the transposase of their precursor, who recognizes their TIRs and help in their mobilization. MITEs are recognized by very high amplified copies in contrast to their precursors, which are typically less abundant (Yang *et al.*, 2006). Autonomous transposons are easy to identify due to their well-known structural features. On contrary, the non-autonomous and small degraded elements or fossil remnants of autonomous transposons are harder to identify and classify due to poorly characterized structural features (Wicker *et al.*, 2007).

The purpose of this study was to investigate and characterize the small non-autonomous TEs, deleted derivatives, fossil remnants and novel insertions. The methodology here was different from others, where a known transposon is screened in any organisms or an identified element is classified by comparing with known elements or on basis of clear structural features. The method used in this study involve the dot plot comparison of homeologous BAC genomic sequences to identify any insertions in one or other BAC and to characterize them. This method not only helped us to identify the new autonomous elements but also helped us to detect small non-autonomous elements or their deletion derivatives, which are very hard to identify by other methods.

9.2 Results

9.2.1 Transposon identification by comparison of homoeologous BAC sequences

In this study, two homeologous BACs, *Musa acuminata* 'MA4_82I11' and *Musa balbisiana* 'MBP_81C12' were compared to study the most conserved and varied regions (TE insertions) (Figure 9.1). The lengths of BACs were 102232 bp and 142973 bp respectively. The homologous region of two BACs is ~102 kb (101.9 kb in MA4_82I11 and 101.7 kb in MBP_81C12, indicating a very small divergence). This supports the equal genome size of the species and suggests that the proportion of TEs insertions or deletions in two BACs was quite similar. Fourteen unequally distributed gaps (>80 bp) were identified from two BACs ranging in size from 82 to 4192 bp. All these gaps showed mobile elements, some of which were easily characterized, while others were not classified into their respective families and considered as novel insertions of unknown superfamilies. Interestingly, the investigated gaps were not random: gap numbers 1, 2, 5, 6, 9, and 10 were in *Musa acuminata* BAC 'MA4_82I11' and 3, 4, 7, 8, 11, 12, 13 and 14 were in *Musa balbisiana* BAC 'MBP_81C12' sequences. The major size difference in BAC 'MA4_82I11' was due to the presence of a 4192 bp DNA transposon (Figure 9.1 & 9.2).

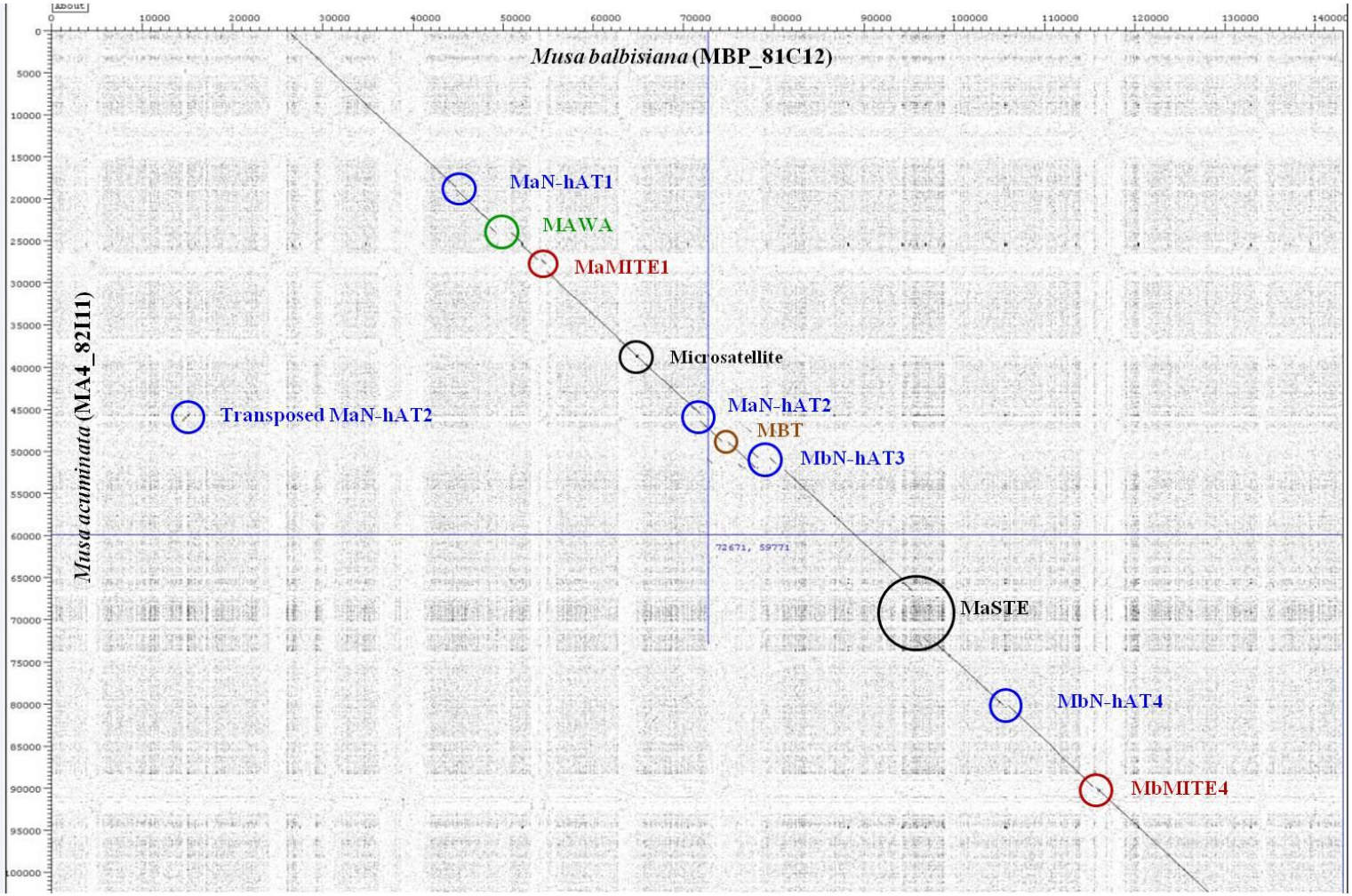


Figure 9.1: Dot plot of homoeologous BAC clones *Musa balbisiana* MBP_81C12 (horizontal) against *Musa acuminata* MA4_82I11 (vertical). The comparison of the BACs showed large homologous region (continuous diagonal line) broken by multiple gap-insertion pairs. The gaps showed transposon insertions present in one BAC and absent in other. Different TEs are encircled and named. Several small insertions <500 bp are not highlighted here.

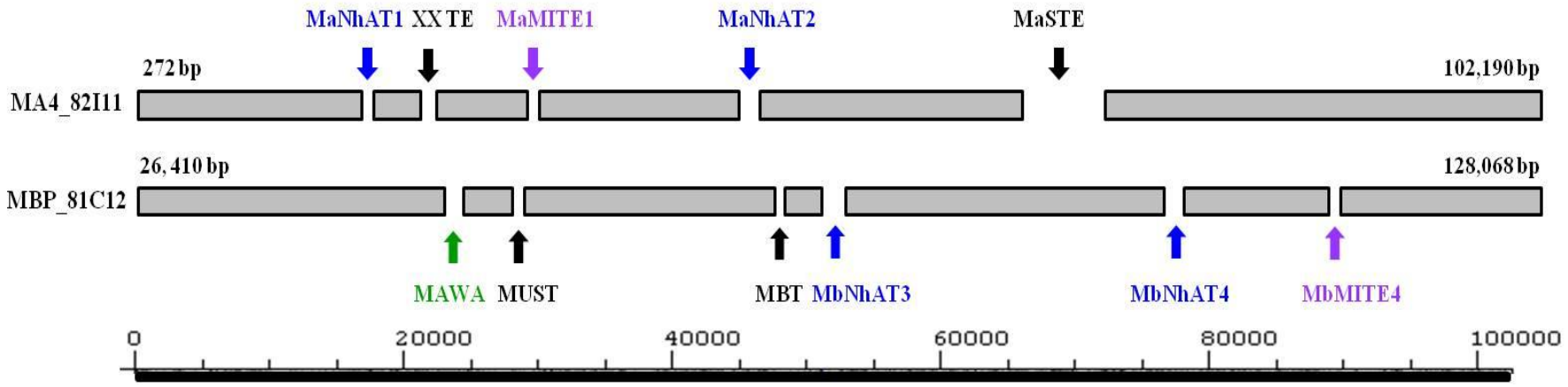


Figure 9.2: Alignment of two homoeologous BACs. The 101.7 kb homoeologous region from *Musa acuminata* (MA4_82I11; from 270-102190 bp) and *Musa balbisiana* (MBP_81C12; from 26400-128068) shows insertion sites of various TEs (named arrows) in the two BAC sequences. XXTE indicates an uncharacterized insertion.

Table 9.1: List of various autonomous and non-autonomous DNA transposons, MITEs and novel insertion from unknown families identified from *Musa* BAC sequences. The names of each element, their sizes, TSDs, TIRs and name of superfamilies are listed. Mut: Mutator. Sequences of elements are available in Appendices (attached CD).

Element name	Super-family	BAC Accessions	Species	Size	TSDs size	TSD sequences	TIRs size	Position in BACs	Autonomous/Non-autonomous
MaN-hAT1	hAT	AC186955.1	*M. acuminata*	273	8	TCCCTGAG	30	18563-18835	Non-autonomous
MaN-hAT2	hAT	AC186955.1	*M. acuminata*	874	8	GTGcTAaC	15	45436-46309	Non-autonomous
MbN-hAT2	hAT	FN396604.1	*M. balbisiana*	926	8	GTTCTATT	15	34017-34874	Non-autonomous
MbN-hAT3	hAT	AC186754.1	*M. balbisiana*	1292	8	GTTGCAAC	15	78250-79541	Non-autonomous
MaN-hAT3	hAT	AC226036.1	*M. acuminata*	1279	8	TTCATGAC	15	17131-18409	Non-autonomous
MbN-hAT4	hAT	AC186754.1	*M. balbisiana*	524	8	TTCAAATG	9	105427-105950	Non-autonomous
MaN-hAT4	hAT	AC226036.1	*M. acuminata*	577	8	GTAGGTTC	9	3837-4413	Non-autonomous
MahAT1	hAT	AC226051.1	*M. acuminata*	5204	8	TTttTTTt	13	73500-78703	Autonomous
MahAT2	hAT	AC226047.1	*M. acuminata*	3336	8	GAAGGAAG	19	34728-38063	Autonomous
MaMITE1	Mut/MITE	AC186955.1	*M. acuminata*	781	5	ATGCG	374/299	27748-28528	Non-autonomous
MaMITE2	Mut/MITE	AC226196.1	*M. acuminata*	664	9	CTACTATAG	291/323	38641-39304	Non-autonomous
MaMITE3	Mut/MITE	AC226047.1	*M. acuminata*	1042	9	TTTGATTTC	333/336	39732-40770	Non-autonomous
MbMITE4	Mut/MITE	AC186754.1	*M. balbisiana*	2067	9	TAATTACAT	427/441	113622-115684	Non-autonomous
MaSTE	Unknown	AC186955.1	*M. acuminata*	4192	9	CATAATGTA	14	67848-72039	Non-autonomous
MbHAL	Unknown	AC186754.1	*M. balbisiana*	632	3	ATA	10	97124-97755	Non-autonomous
MAWA	Unknown	AC186754.1	*M. balbisiana*	1676	2	AG	Nil	49161-50836	Non-autonomous
MUST	Unknown	AC186754.1	*M. balbisiana*	367	2	AA	Nil	53,800-54,166	Non-autonomous
MBT	Unknown	AC186754.1	*M. balbisiana*	621	3	ATG	Nil	74,240-74,860	Non-autonomous
TATA	Unknown	AC186955.1	*M. acuminata*	504	19	TATATATATAT ATATATAT	Nil	65047-65550	Non-autonomous
MAX	Unknown	AC186955.1	*M. acuminata*	261	2	TA	Nil	67426-67686	Non-autonomous
TINY	Unknown	AC186754.1	*M. balbisiana*	82	5	TGGTC	Nil	100030-100111	Non-autonomous

9.2.2 The hAT transposons diversity in *Musa*

Four non-autonomous hAT elements (*MaN-hAT1, MaN-hAT2, MbN-hAT3, MbN-hAT4*) were identified by comparing *Musa acuminata* 'MA4_82I11' against *Musa balbisiana* 'MBP_81C12' BAC sequences. These non-autonomous hATs range in sizes from 273 bp to 1292 bp without having an active transposase necessary for their mobilization and integration to a new site (Figure 9.3a). For *in silico* identification of their autonomous copies, *Musa acuminata* and *Musa balbisiana* sequence data available in GenBank was screened and autonomous copies of *MaN-hAT1* and *MaN-hAT2* were detected, which are 5204 bp (*MahAT1*) and 3336 bp (*MahAT2*) respectively (Figure 9.3b). All the six hATs are studied in detail and are characterized on the basic of structural features (TSDs and TIRs). The detailed structural and phylogenetic analysis showed the clustering into two major clades further splitting hATs into four families.

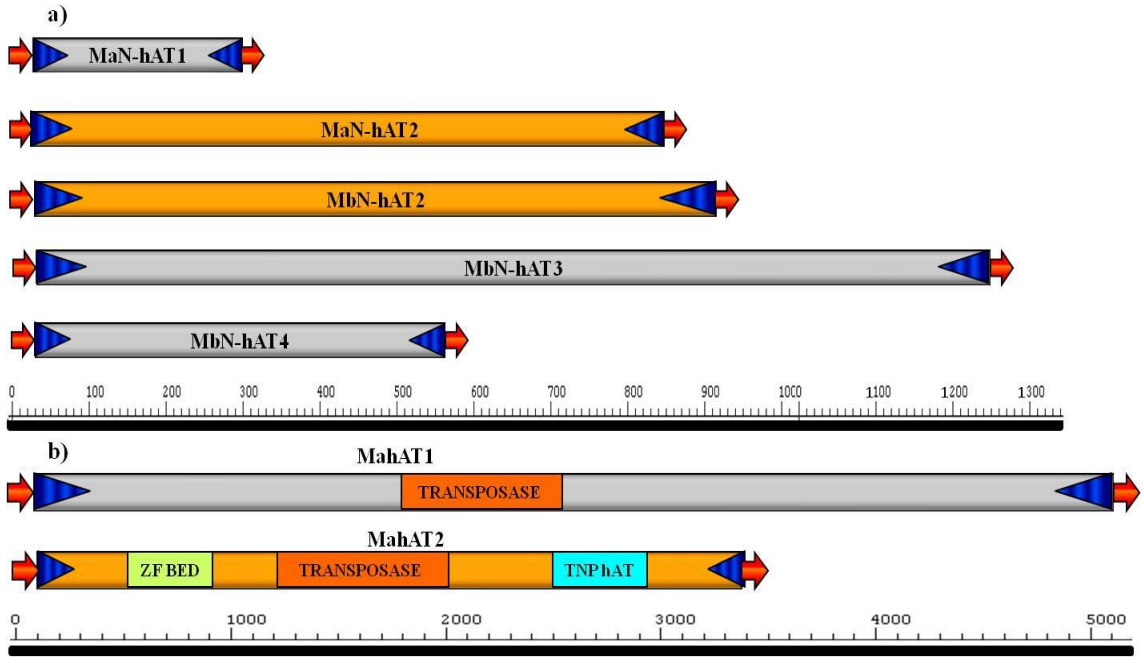


Figure 9.3: Schematic representation of a) non-autonomous hATs b) autonomous hATs in *Musa*. Red arrows indicate 8 bp TSDs, blue triangles represent TIRs. The autonomous hATs showed transposase, Bed Zinc finger (ZF BED) and hAT family dimerization domain (TNP hAT); scale in bp.

9.2.2.1 Structural features and characterization of non-autonomous hATs

A 273 bp non-autonomous hAT was identified from *Musa acuminata* BAC 'MA4_82I11' (Figure 9.1 & 9.4a). It is named as *MaN-hAT1*; '*Mu*' indicates *Musa*, '*N*' represent non-autonomous and '*1'* indicates the number of hAT identified in respective *Musa* BACs.

Initially, it was identified in *Musa acuminata* but few copies were also detected in *Musa balbisiana* genomes. The *MaN-hAT1* was characterized by generating 8 bp TSDs (5′-TCCCTGAG-3′) and 22-30 bp TIRs with a 4 bp mismatch (Table 9.1). The 5′ termini of TIRs are highly conserved in all the copies and are 5′-CAAGG-3′ (Figure 9.6). The GC content of TIRs is very high (63%) as compared to AT (37%). There are 2 copies of 5′-CGGC-3′ and 8 copies of 5′-CCGG-3′ direct tetranucleotide repeats in the sub-terminal region suggesting the binding sites for transposase.

Musa acuminata 'MA4_82I11' harbours an 874 bp insertion starting from 45436-46309 bp (Figure 9.4b), with characteristics of hAT transposons and was named *MaN-hAT2*. The insertion is flanked by 8 bp imperfect TSD 5′- GTGcTAaC-3′, and has a 15 bp TIR (Table 9.1). In general, the insertion is equally rich in GC and AT (50% each). The first half of the insertion is adenine rich with 12 copies of 2-6 bp poly(A) sequences. It has 15 small segments of 5′-CGAG-3′, 9 segmnts of 5′-GAAG-3′, 7 segments of 5′-CAAC-3′, and 5 segments of 5′-GGGC-3′ tetra-nucleotide sequences repeating at intervals. The termini of TIRs are well conserved with 5′-CAGTG-3′ in all the copies. Another analogue of *MaN-hAT2* was found inserted in *Musa balbisiana* clone 'MBP-26I6' (FN396604.1) from 34017–34874 bp. Due to homology to *MaN-hAT2*, it is named as *MbN-hAT2* as it was detected from *Musa balbisiana*. *MbN-hAT2* is 926 bp, well characterized by 8 bp TSD 5′-GTTCTATT-3′ and 23 bp TIRs. It is high in GC content with a high GC:AT ratio (57:43%). The ~150 bp terminal regions from both sides of *MaN-hAT2* and *MbN-hAT2* are highly similar indicating their common family (Figure 9.3a; Table 9.1).

While comparing the *Musa acuminata* 'MA4_82I11' against *Musa balbisiana* 'MBP_81C12' BAC sequences, a 1292 bp insertion was identified from *Musa balbisiana* from position 78250-79541 bp (Figure 9.5a). This was named *MbN-hAT3* and has well characterized hallmarks (TSDs and TIRs) of hAT superfamily. *MbN-hAT3* is flanked by 8 bp TSDs as 5′-GTTGCAAC-3′ and 15 bp TIR (5′-CAAGGTctGCaTACC-3′). The GC content of insertion was low (39.5%) as compared to AT (60.5%). The alignment of homologues of *MbN-hAT3* revealed that the termini (5′-CAAGG-3′) of TIRs are highly conserved like other hATs (Figure 9.6). Another insertion was found located in *Musa balbisiana* 'MBP_81C12' BAC sequence and is named *MbN-hAT4*. *MbN-hAT4* insertion displays the structural feature of 8 bp TSDs (5′-TTCAAATG-3′) and 9 bp TIRs (5′-CAAGGTtTG-3′) (Table 9.1). The termini of TIRs are highly conserved and are 5′-

CAAGG-3′ (Figure 9.6). There are 4 copies of microsatellites or simple sequence repeats (SSRs) with TA nucleotide repeats. The first copy is 66 bp long $(TA)_{33}$, second is 22 bp long $(TA)_{11}$, third is imperfectly 20 bp $(TA)_{10}$ long and fourth is 22 bp $(TA)_{11}$ long.

9.2.2.2 Structural features and characterization of autonomous hATs

To identify the autonomous partners of non-autonomous hATs, the sequences were used as query in blast searches. By using *MaN-hAT1* as query, ~5.2 kb autonomous hAT element was identified designated as *MahAT1*. The structural features of *MahAT1* showed either truncated or degraded element, with no perfect TSDs at 3′ terminal end. On the basis of TIRs at 3′ end and imperfect TSDs flanking them, the size was counted and found to be 5204 bp. It is flanked by 8 bp imperfect TSDs (TTttAAAt) and 10-12 bp TIRs. The internal region displays a coding region for transposase but no other domains were found (Figure 9.3b: Table 9.1). It is a defective element due to the presence of several stop codons in the coding regions. *MahAT1* is highly AT rich (67%) with high AT content in 1 kb terminal regions. Several small A and T rich sequences are dispersed in the element. The blast searches yielded no strong hits indicating that the element is less active and rarely present in *Musa* genomes.

Using *MaN-hAT2* as query in blast searches, the autonomous partner of the element was identified from *Musa acuminata* clone 'MA4-86B3' (AC226047.1) from 34728-38063 bp and designated as *MahAT2*. It has very prominent hallmarks of hATs with 8 bp TSDs and 19 bp TIRs (5′-CAGTGATTTAAAAAGCGCT-3′) (Figure 9.3b; Table 9.1). The upstream ~400 bp starting from TIRs are GC rich while ~400 bp towards downstream are highly AT rich. The central part of transposon is AT rich with many small poly adenine and thymine sequences. The transposase (DUF659) of *MahAT2* is ~555 bp long, with additional Bed Zinc finger (ZF BED) and hAT family dimerization (TNP hAT) domains. These dimerization domain forms very stable dimmers in vitro. The Bed Zinc finger domain is located on N-terminus, while dimerization domain is located on C-terminus of transposase. BLASTN searches using 3336 bp *MahAT2* sequence against the Nucleotide Collection (nr/nt) retrieved 184 hits; with 4 complete copies and remaining represent the partial copies and deletion derivatives. Using transposase as query sequence, 136 hits returned from *Musa acuminata*, *Musa balbisiana*, *Vitis vinifera*, *Glycine max*, and *Lotus japonica* indicating its diversity in other plant genomes.

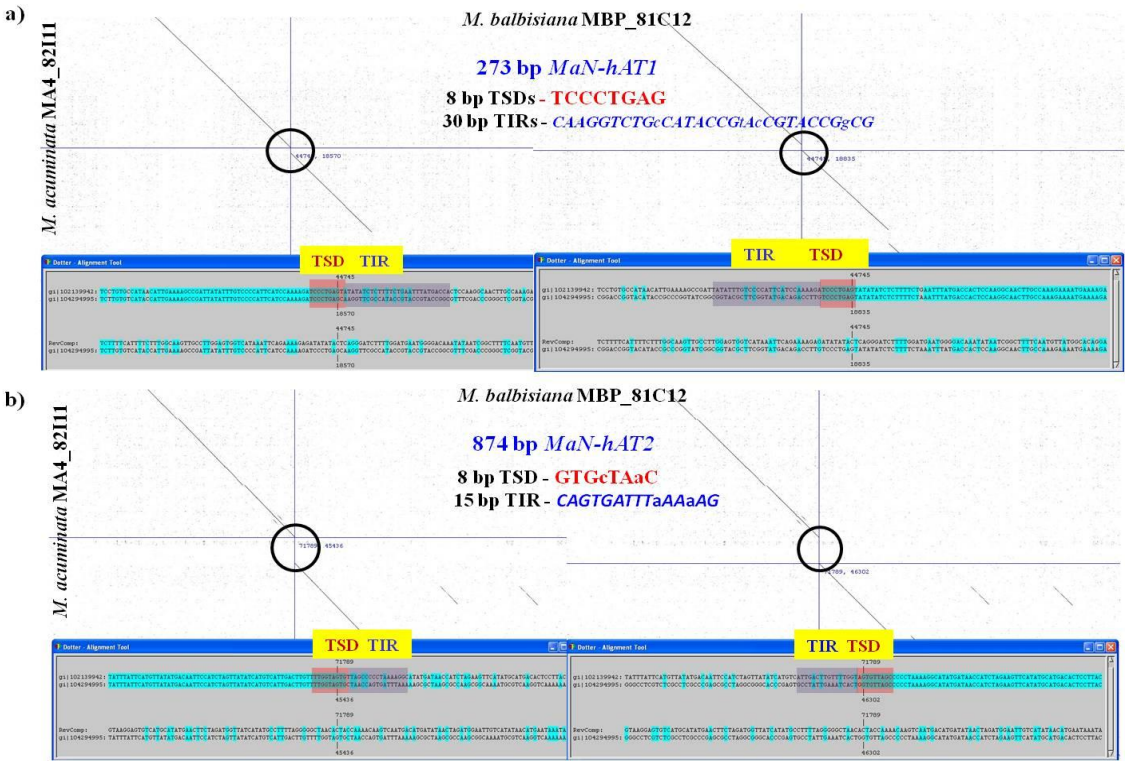


Figure 9.4: Dot plot comparison of homoeologous BAC clones *Musa balbisiana* MBP_81C12 (horizontal) against *Musa acuminata* MA4_82I11 (vertical) showing *MaN-hAT1* and *MaN-hAT2* insertion sites in *Musa acuminata*. The size, TSDs and TIRs are also shown in text and in the insets showing both ends of the insertion (large crosshair in left and right dot plots).

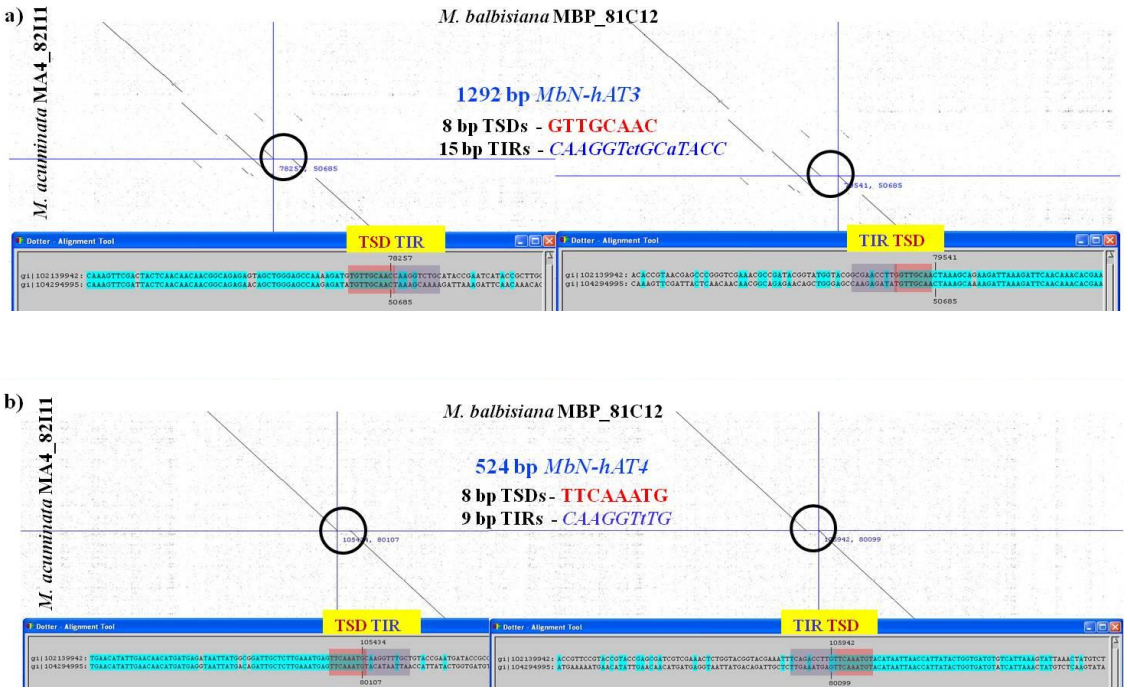


Figure 9.5: Dot plot comparison of homoeologous BAC clones *Musa balbisiana* MBP_81C12 (horizontal) against *Musa acuminata* MA4_82I11 (vertical) showing a) *MbN-hAT3* and b) *MbN-hAT4* insertion sites in *Musa balbisiana*. The size, TSDs and TIRs are also shown (insets).

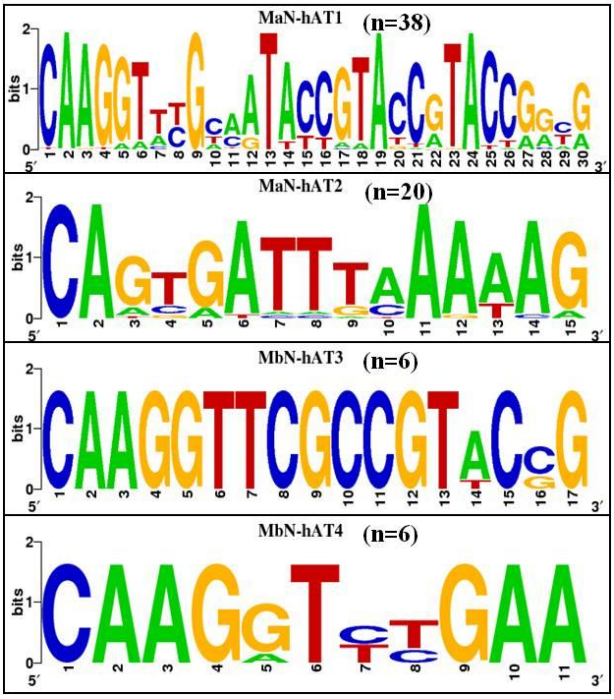


Figure 9.6: Sequence logos of *Musa* hATs TIRs. The *hAT* CAAGG motif is highly conserved and observed in all TIRs of *MaN-hAT1*, *MbN-hAT3* and *MbN-hAT4*. The height of nucleotides shows the information content of nucleotides within the TIRs of the four groups of elements.

9.2.2.3 Insertional polymorphisms of non-autonomous hATs in *Musa*

By using degenerative primer pair MaNhAT1F 5′-ACCCACCTGGCTCTTGTGTC-3′ and MaNhAT1R 3′-AGCGAATGTGTTTTGACCAC-5′, *MaN-hAT1* was amplified in different *Musa* genomes. The primers were 66 bp upstream and 220 bp downstream respectively, of the *MaN-hAT1* insertion site. A total of 96 *Musa* genomes (Table 2.2 & 2.3) were used to study the insertion polymorphism of *MaN-hAT1*. The product size of 560 bp including 273 bp *MaN-hAT1* insertion and flanking regions were successfully amplified in several *Musa acuminata* (AA, AAA, AA cv) and 2 *Musa balbisiana* genomes. Few *Musa* triploids (AAB, ABB) also showed the amplification of *MaN-hAT1*, indicating that the insertion is contributed by A-genome. The genomes amplifying the *MaN-hAT1* insertions were AA (Banksii 623, Long Tavoy pied, Zebrina, Tomolo, Calcutta 4, Kadali, Matti, Cherukadali), BB (Javan), AB cv (Safet Velchi, Kunnan, Njalipovan, Adukkan), AAA (Mbwazirume, Intokatoke, Yangambi KM5, Manoranjitham, Grow-michel, Monsmari, Robusta, Dwarf cavendish), AAB (Figue Pomme Géante, Karimkadali, Perumadali, Poovan, Doothsagar, Vellapalayamcodan, Poomkalli), ABB (Simili Radjah, Namwa Khom, Kosta bontha, Monthan, Karpooravali), and ABBB (Yawa 2). The lower bands (~300 bp) representing

the pre-insertion sites were amplified from most of the genomes except *Musa acuminata* (AA), which amplified the higher bands only. Many of them showed both higher (560 bp) and lower bands (~300 bp) indicating the heterozygous nature, while others showed either higher or lower bands (Figure 9.7a & b).

By using degenerative primers MaNhAT2F and MaNhAT2R (Table 9.2), 872 bp long *MaN-hAT2* was amplified in few accessions out of 48 *Musa* genomes (Table 2.2). The expected product size of insertion with flanking sequences was 1287 bp. Variable products ranging from ~650-1300 bp in different *Musa* genomes (AA, BB, AB, AAB and ABB) were amplified. Some of the genomes amplified complete copy of *MaN-hAT2*, while other amplified the partial or degraded fragments as confirmed by sequencing of these fragments. This advocates that some accessions possess full size elements, while others harbour many degraded or partial copies of this element. This insertion was present in *Musa acuminata*, *Musa balbisiana* and other *Musa* polyploids (AB, AAA, AAB, ABB) suggesting its diverse nature and amplification (Figure 9.7c).

MbN-hAT3 was amplified from many *Musa* accessions by the primer set MbNhAT3F 5′-CTCAACAACAACGGCAGAGA-3′ and MbNhAT3R 5′-GCTTTGCCCATGGTATTCTC-3′. The expected products including insertion and flanking sequences were 1441 bp. The results indicated its amplification in *Musa balbisiana* (BB) and *Musa* polyploids having 'B' allele in them (AAB, ABB, ABBB). No amplification in any of *Musa acuminata* (AA, AAA) genomes indicates B-genome specificity of this hAT. The insertion polymorphisms was observed in BB (P. Klutuk Wulung, P. Batu, Tani, Lal Velchi, Batu), AB cv (Safet Velchi, Kunnan, Njalipovan, Adukkan, Padalamukili), AAB (Lady Finger, Foconah, Prata Ana, Figue Pomme Geante, Popoulou, P. Raja Bulu, P. Rajah, P. Ceylan, Motta povan, Perumadali, Palyamcodan, Krisnavazhai, Poovan, Doothsagar, Charapadati, Kumbillakannan, Vellapalayamcodan, Ettapadati, Padati, Chinali, Nendran, Poomkalli), ABB (Orishele, Pelipita, Dole, Saba, Monthan, Simili Radjah, Ice Cream, Namwa Khom, Peyan, Kanchikela, Boothibale, Monthan, Karpooravali) and ABBB (Yawa 2) accessions amplifying the *MbN-hAT3*. The lower bands showing the flanking sequences were amplified in majority, except few genomes (Figure 9.8a & b).

By using primer pair MbNhAT4F + MbNhAT4R, a product size of 860 bp was achieved with *MbN-hAT4* insertion. The insertion polymorphism pattern showed their amplification

in *Musa balbisiana* (BB), and *Musa* polyploids (AAB, ABB and ABBB) having 'B' allele indicating the B-genome specificity of *MbN-hAT4*. The genomes that showed *MbN-hAT4* amplifications are AA (Paliama), BB (PKW, P. batu, Tani, Lal Velchi, Javan), AB cv (Safet Velchi, Kunnan, Njalipovan, Adukkan, Padalamukili), AAB (Orishele, Figue Pomme Géante, Popoulou, P. Raja Bulu, P. Rajah, P. Ceylan, Doothsagar, Kumbillakannan, Velipadati, Ettapadati, Padati), ABB (Pelipita, Dole, Saba, Monthan, Simili Radjah, Ice Cream, Namwa Khom, Kosta bontha, Karpooravali, Pisang Awak), and ABBB (Yawa 2). The lower bands (~300 bp) with flanking regions were observed in many *Musa* genomes amplifying the flanking sequences (Figure 9.8c & d).

Table 9.2: List of primers for amplification of various DNA transposons in *Musa*. The transposon sizes, product sizes, primer names and sequences are given.

Sr.No.	TE family	TE Size	Product size	Primer name	Primer Sequence
1	*MaN-hAT1*	273	560	MaNhAT1F	ACCCACCTGGCTCTTGTGTC
				MaNhAT1R	AGCGAATGTGTTTTGACCAC
2	*MaN-hAT2*	874	1287	MaNhAT2F	TTGATCATACCTAGGTGGATG
				MaNhAT2R	AACAACATGCCATGGTATCAG
3	*MbN-hAT2*	874	986	MaNhAT2F	GAGGAAGTCAAATGCAGAAATG
				MaNhAT2R	GATACTTTTGATGGAGAATTTG
4	*MbN-hAT3*	1292	1441	MbNhAT3F	CTCAACAACAACGGCAGAGA
				MbNhAT3R	GAGAATACCATGGGCAAAGC
5	*MbN-hAT4*	524	860	MbNhAT4F	GAACCAAGCTTACATTGAGAGT
				MbNhAT4R	GAGACACAAATCAATCACCTAT
6	*MBN-hAT4*	524	800	MbNhAT4F	ATTGAGGAAGCACAAGAACATA
				MbNhAT4R	CACCTATGCAACAAAGAAAATC
7	*MahAT1*	358	5204	MahAT1F	ATGAGATGCGAGTTCCATTG
				MahAT1R	CATGGAGTCCAATATAAGTG
8	*MahAT2*	456	3336	MahAT2F	ATGAGATGCGAGTTCCATTG
				MahAT2R	CATGGAGTCCAATATAAGTG
9	*MITE*	781	1052	MaMITE1F	AACGGGACGAGTCTTGAGAA
				MaMITE1R	TAAATGTCTCCGCTTAGGCC
10	*MBT*	621	864	MBTF	GATCAAATGGGGAAGCAACC
				MBTR	ACTTCTCCCGTGTGTGTCGT
11	*AGNABI*	1676	1844	MAWAF	AGGAGCCACAAGGAAGATTG
				MAWAR	GCCAATTGTAGCTCAAAATC
12	*MUST*	384	548	MUSTF	GGGAGCACGGAATTTGCCC
				MUSTR	CAAGACGGACACCGAGGAC
13	*MaSTE*	4192	1138	MaSTEaF	CGCATGATGTTTTTGATGTA
				MaSTEaR	GAGGTACAACTCAACAAAAG
14	*MaSTE*	4192	1122	MaSTEbF	GGTTTTGATTGATTGAAGAC
				MaSTEbR	CAAGAATGAGTGACAAGTCG

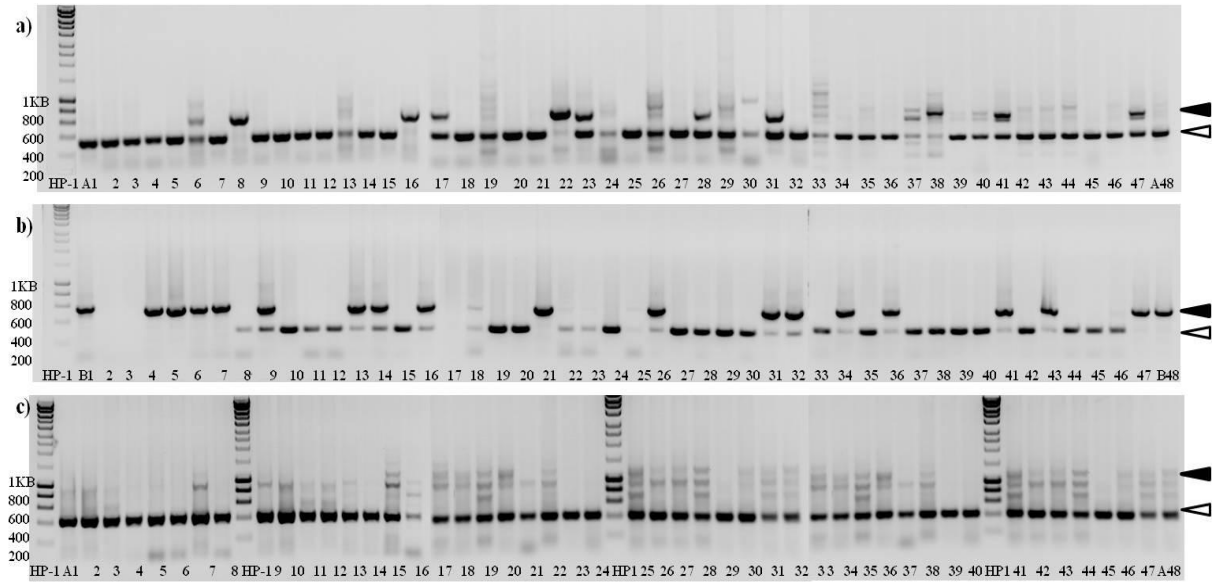


Figure 9.7: Transposon insertional polymorphisms of *Musa* hATs. a-b) *MaN-hAT1* insertion sites in *Musa* accessions. Long bands (560 bp) represented by filled arrowheads (right) indicate amplified *MaN-hAT1* insertions; short bands amplify the flanking sequences only (open arrowheads). c) *MaN-hAT2* insertion sites in *Musa*. Filled arrowheads pointing to ~1287 bp *MaN-hAT2* and open arrowheads show products without the *MaN-hAT2* insert. All PCR figures show reversed images of size-separated ethidium bromide-stained DNA on agarose gels after electrophoresis; ladders show fragments sizes in base pairs; numbers at the base indicate accessions of the species indicated in Tables 2.2 and 2.3.

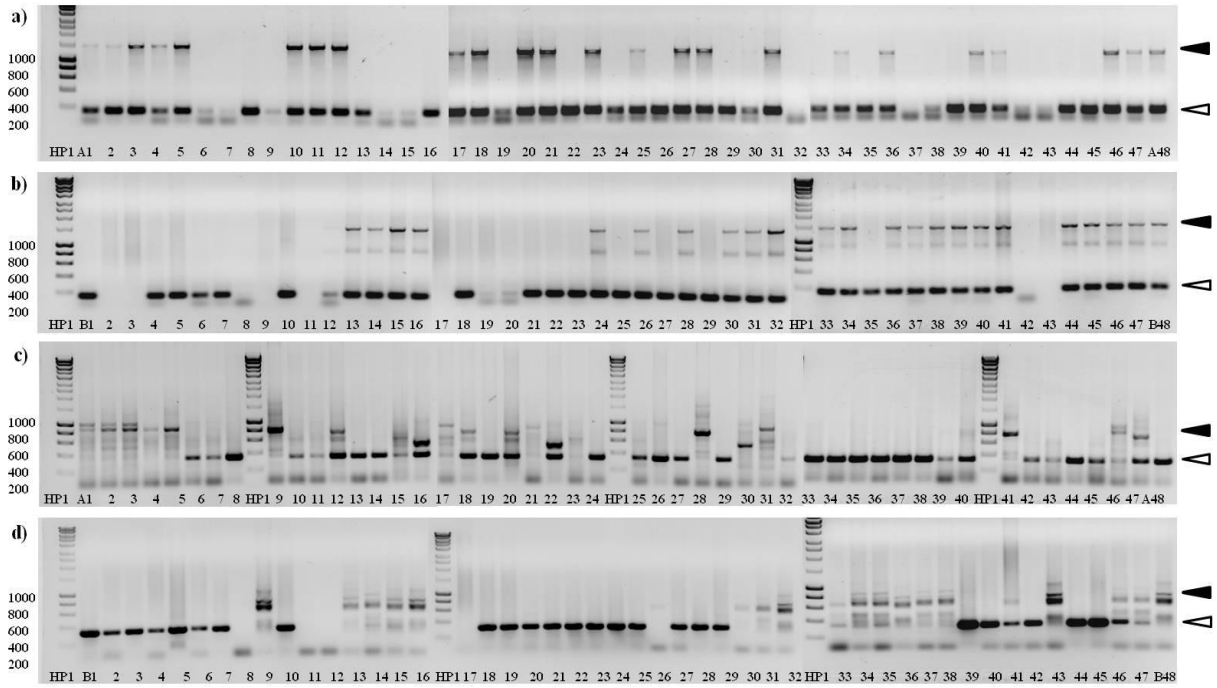


Figure 9.8: Transposon insertional polymorphisms of *Musa* hATs. a-b) *MbN-hAT3* insertion sites in various *Musa* accessions*:* Long bands (1441bp) showed the amplified element (filled arrowheats) and short bands (open arrowheads) show amplification of the pre-insertion sites only. c-d) *MbN-hAT4* amplification with degenerate primer pair MbNhAT4F and MbNhAT4R. Long bands (860-bp) show the amplified *MbN-hAT4* element and short bands amplify the flanking sequences only.

9.2.2.4 Fluorescent *in situ* Hybridization (FISH) of hAT sites on *Musa* chromosomes

The chromosomal distribution of hATs in the *Musa* species was analyzed by FISH. Using *MaN-hAT1* as probe in *Musa* diploids and triploids, very strong signals were observed clustered in central regions on all A-genome chromosomes. The distribution pattern of *MaN-hAT1* on the chromosomes of *Musa* triploid (ABB) revealed that strong signals were observed on 11 chromosomes, out of 33, indicating they were contributed by A-genome. The distribution of *MaN-hAT1* signals on these 11 chromosomes was not uniform, as some chromosomes showed strong and clustered signals while others showed few but dispersed signals. In Cavendish (AAA), all the 33 chromosomes showed very strong signals of *MaN-hAT1* on the central regions (9.10a). In diploid *Musa acuminata malaccensis* (AA; 2n=22), the strong signals of *MaN-hAT1* were observed in the central regions of all 22 chromosomes revealing its abundance on all chromosomes. In contrast, the *MbN-hAT3* only hybridized to few chromosomes showing its patchy distribution (Figure 9.9c-f). Similarly, *MbN-hAT4* was used as a probe to see its distribution pattern on different chromosomes in *Musa acuminata* (AA) and *Musa balbisiana* (BB). Very strong and dispersed signals were observed on all chromosomes in *Musa balbisiana* (BB) indicating its contribution from B-genomes. Two of the 22 chromosomes showed very strong clustered signals painting nearly half of the chromosomes. The telomeric signals were very strong and clear (Figure 9.9a-j).

9.2.2.5 Phylogeny of hATs in *Musa*

The evolutionary relationships of *Musa* non-autonomous hATs were studied. Seventy complete copies identified by dot plot and blast searches were aligned and the ~200 bp region from 5′ terminal end including TIRs was used to generate the tree by Neighbour-Joining algorithm with 1000 bootstrap repetitions. The 200 bp region from autonomous hAT '*MahAT2*' was used to root the tree. The tree clearly separated into two main lineages bringing *MaN-hAT2* family in one and *MaN-hAT1, MbN-hAT3* and *MbN-hAT4* families in the other lineage (Figure 9.10). This clearly indicates that the elements fall into two major groups, one having 20 members from *MaN-hAT2* family, the other having 50 members. Although the second lineage is shared by 3 families, family specific groups are clearly separated. The five members from *MbN-hAT3* family clustered together, while six members from *MbN-hAT4* shared a group with some *MaN-hAT1* elements dispersed in it.

The detail structural analysis of the elements revealed that *MaN-hAT1*, *MbN-hAT3* and *MbN-hAT4* families share the similar TIRs, but their internal regions are variable, due to which they are assigned into their respective families on the basis of homology in the entire lengths. This suggests that the TIRs segregated the 70 elements into 2 groups, which can be further classified into four families on the basis of homology in their entire lengths (Figure 9.10).

9.2.2.6 *Musa* hAT transposon diversity and copy number estimation

To identify the distribution and copy numbers, the search was extended by using initially identified hATs as query in blast searches. The 273 bp *MaN-hAT1* yielded 34 complete copies with many defective or truncated copies. Four copies of *MaN-hAT1* were added after sequencing from PCR amplicons from *Musa* genomes to analyse their phylogenetic relationships. *MaN-hAT2* retrieved 19 complete sequences from A, B-genomes. Using 1292 bp *MbN-hAT3* insertion as a query in BLASTN, 2 complete copies from *Musa balbisiana* and a truncated copy from *Musa acuminata* was found. All the other hits were partial sequences or truncated copies, which were removed from the analysis. The 524 bp *MbN-hAT4* sequence generated 101 blast hits, with 4 complete copies. Two more sequences were added after sequencing from triploid *Musa*. Based on these numbers, total copy numbers for each family were estimated for whole genome of *Musa acuminata* and *Musa balbisiana.*

The estimated copy numbers for *MaN-hAT1* family was 4200 and 2475 in *Musa acuminata* and *Musa balbisiana* respectively, while 2100 and 1110 copies were counted for A and B-genome from *MaN-hAT2* family. The *MbN-hAT3* is the low copy number family with 165 copies in *Musa acuminata* and 550 copies from *Musa balbisiana. MbN-hAT4* is B-genome specific, where 1100 copies were estimated while 150 copies were estimated from A-genome *Musa*. In general, a total of 6565 non-autonomous hATs are proliferating in *Musa acuminata*, while 5235 copies are residing in *Musa balbisiana*. Similarly the copy numbers for their autonomous counterparts were also estimated. Approximately 300 and 260 copies of *MahAT1* were estimated for *Musa acuminata* and *Musa balbisiana* respectively, while 450 and 275 *MahAT2* copies are estimated for A and B-genomes respectively. This suggests that the non-autonomous copies are 10 fold higher than their autonomous partners.

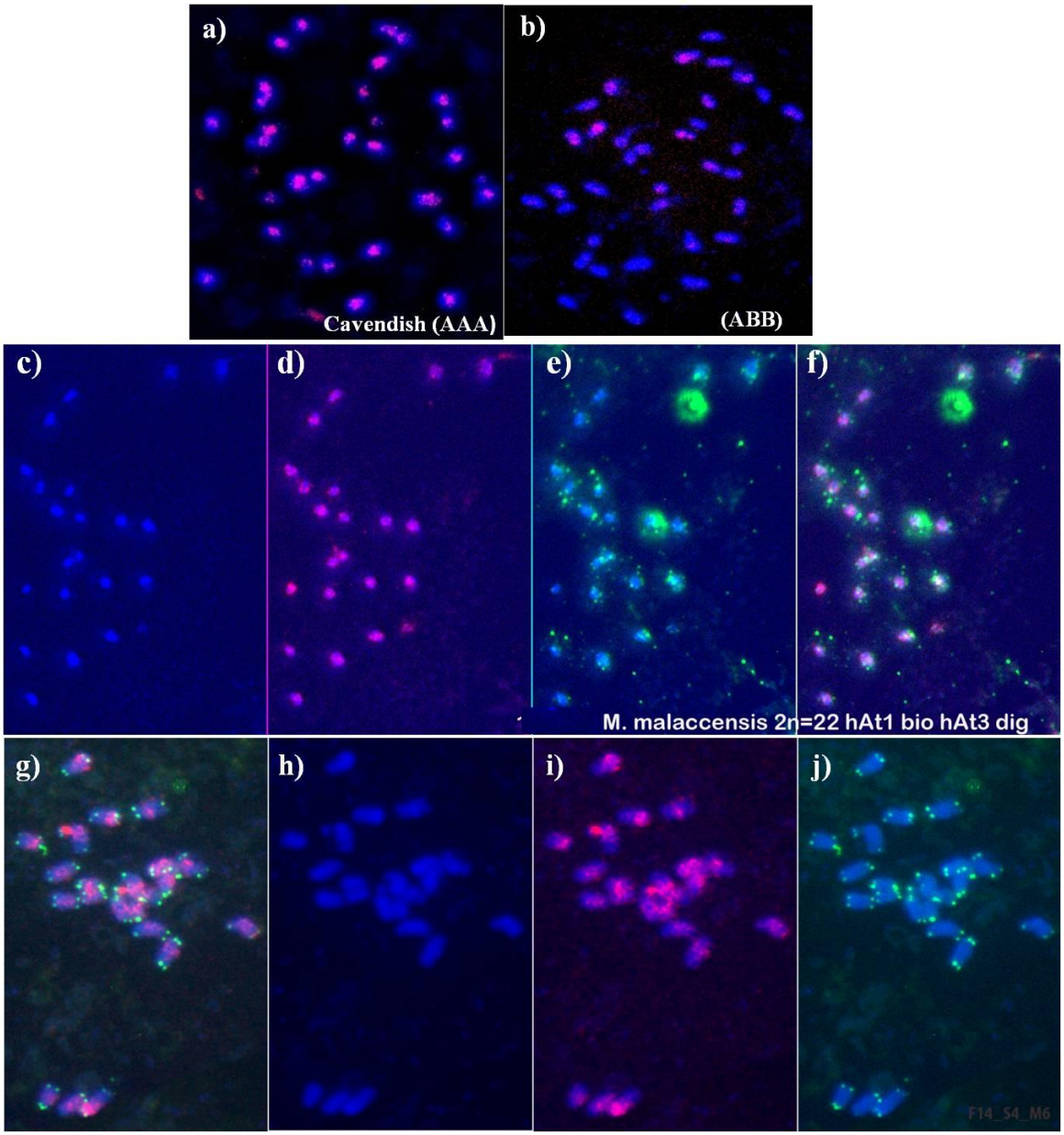


Figure 9.9: Fluorescent *in situ* hybridization showing the distribution of hATs on DAPI stained (blue) *Musa* chromosomes. a-b) A-genome specific *MaN-hAT1* in two *Musa* triploid (2n=3x=33) accessions indicated in the figure located on A-genome chromosomes. c-f) Chromosomes of *Musa acuminata* (ssp. *malaccensis*; AA; 2n=2x=22) showing the locations of *MaN-hAT1* (red probe) and *MbN-hAT3* (green probe) elements. The *MaN-hAT1* (red) elements are distributed on all the 22 chromosomes, while the *MbN-hAT3* (green) is present on few chromosomes indicating its patchy distribution and lower abundance. g-j) chromosomes of a wild banana (BB, ITC0545, 2n=2x=22) with dispersed but not entirely uniform location of a B-genome specific *MbN-hAT4* (red) and telomeric probe (green) on DAPI-stained chromosomes (blue). Magnification x2000.

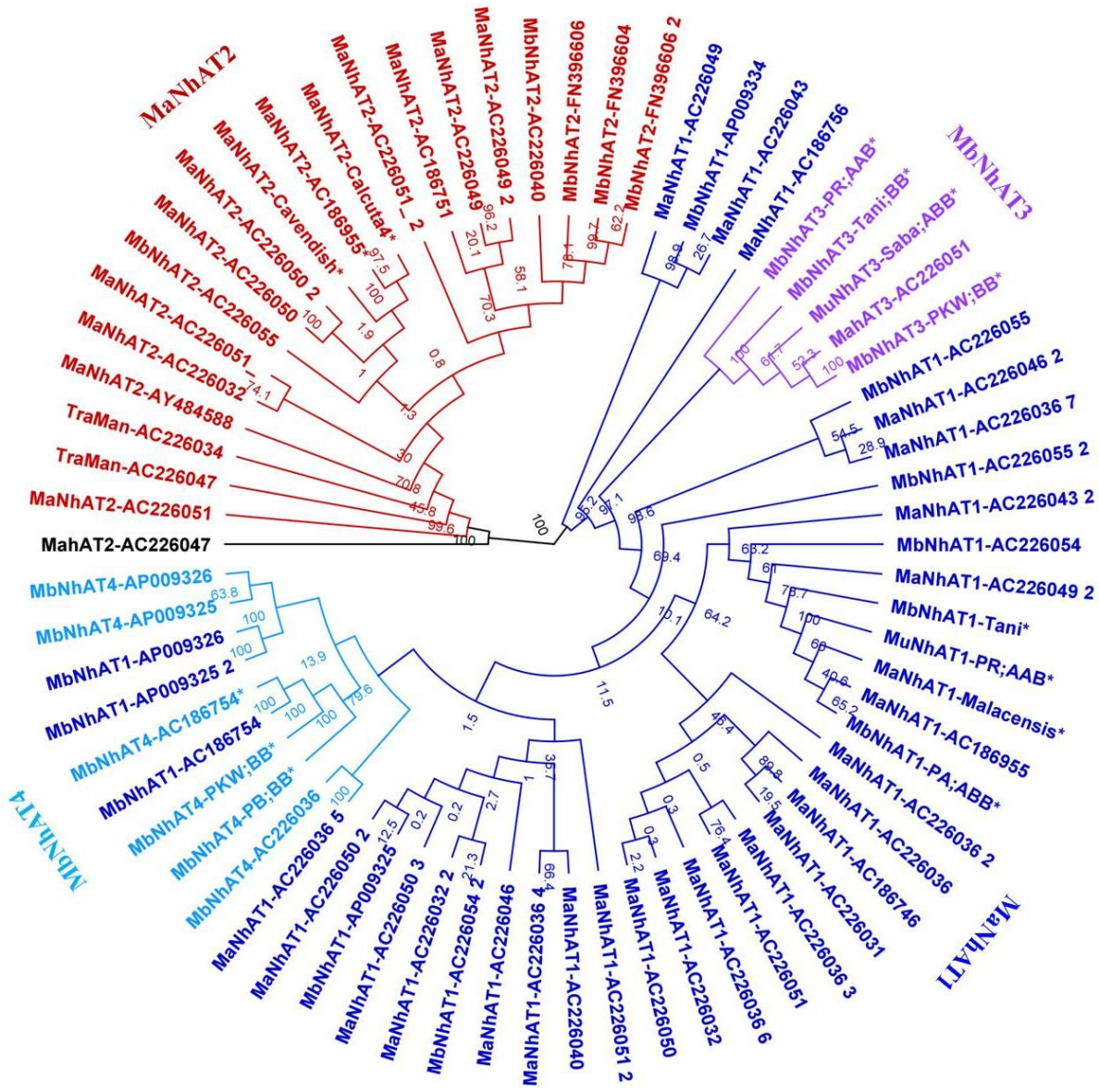


Figure 9.10: Neighbour-Joining tree showing relationship of hAT families identified in *Musa*. The phylogenetic tree of *Musa* hATs based on the 200 bp DNA sequence from 5′ terminal end was constructed by the Neighbour-Joining method with 1000 bootstrap repetitions using the Geneious Pro program. The tree is rooted with the autonomous hAT *MahAT2*. The bootstrap support is shown near the nodes. The names of the elements are followed by the BACs in which they were identified. The hATs cluster into two strongly supported lineages; the *MaN-hAT2* elements grouped in the first lineage; and *MaN-hAT1*, *MbN-hAT3* and *MbN-hAT4* clustered in the second lineage where the *MbN-hAT3* clade is well supported.

9.2.3 Identification and characterization of MITEs in *Musa*

The comparison of two homeologous BACs, *Musa acuminata* (MA4_82I11) and *Musa balbisiana* (MBP_81C12) led to the discovery of a MITE-like element named as *MaMITE1* (*Musa acuminata* MITE 1) and characterized by having 781 bp size including 5 bp TSDs and long TIRs (374/299 bp). The insertion was present in *Musa acuminata* from 27748-28528 bp (Figure 9.11 & 9.12). Although generating 5 bp TSDs, but long TIRs suggested it a MITE originated from Mutator-like elements, but no homology to a known Mutator element was found. BLASTN searches using *MaMITE1* as query sequence against *Musa* Nucleotide Collection (nr/nt) database returned 454 hits from *Musa acuminata,* 138 in *Musa balbisiana*, 48 against *Musa ornata* and 13 in other *Musa* hybrids (AAB). Approximately 80 copies were complete copies, while remaining were either incomplete or appear to be low degenerate with low percentage of query coverage. The majority of hits were against Ty3/Gypsy and *Musa acuminata* Monkey LTR retrotransposon suggesting its nested structure. Another MITE designated *MaMITE2* was found in *Musa acuminata* BAC accession (AC226196.1), 664 bp long including 9 bp TSDs and long TIRs (5′-291/332-3′). The 3′ terminal TIR had uneven activity and increased its size as compared to its 5′ TIR. The MITE is highly AT rich (65%), suggesting the typical MITEs features.

MaMITE3 was detected from *Musa acuminata* (AC226047.1) with 9 bp flanking TSDs and 5′-291/332-3′ bp TIRs (Figure 9.13), with no internal region encoding any protein domain. A 2064 bp long element (*MbMITE4*) was identified from *Musa balbisiana* (MBP_81C12) from 113622-115684 bp terminated by 9 bp TSDs (5′-TAATTACAT-3′) and long TIRs (5′-427/441-3′ bp). The starting and ending TIRs are highly AT rich, which are 70.5% and 64.6% respectively. In general the element GC content is 42.7%, with central portion having more GC% as compared to the TIRs. '*MbMITE4*' is a non-autonomous hAT element lacking any active transposase but capture a protein domain from pectinesterase superfamily, which is ~240 aa long and cover most of the internal sequence of element. Blast searches, using 2064 bp as query sequence yielded 134 hits, out of which only two are the complete copies (Figure 9.13; Table 9.1).

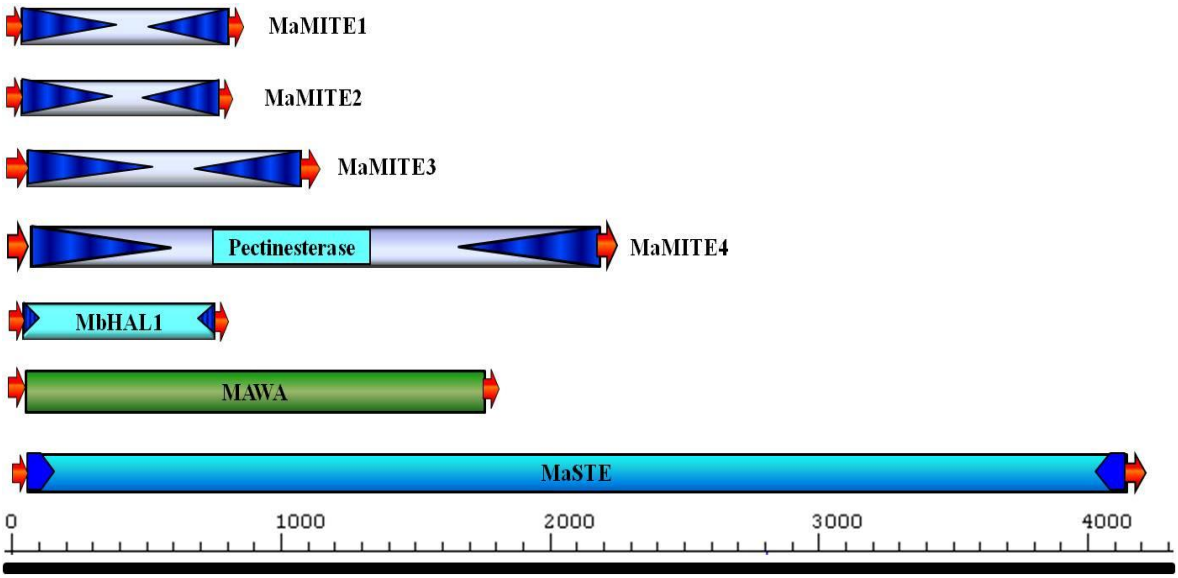


Figure 9.11: Structure of different MITEs and novel transposons in *Musa*. Red arrows indicate varied sized TSDs, blue triangles represent TIRs. The non-autonomous transposons showed no transposase in their structures. Two bp TSDs but no TIRs in the MAWA element are detected.

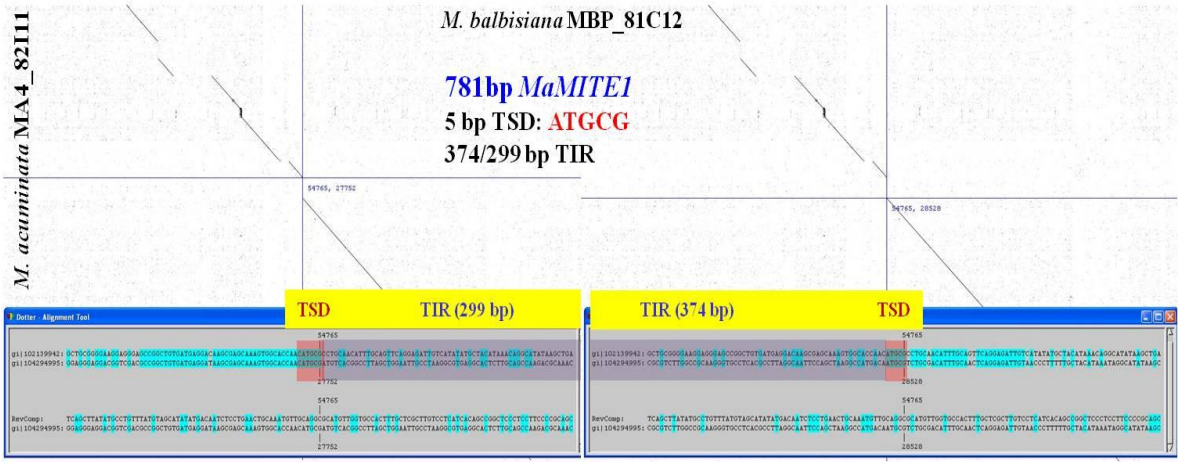


Figure 9.12: Dot plot showing the *MaMITE1* insertion site in *Musa acuminata BAC* MA4_82I11. The TSDs and TIRs in aligned sequences are highlighted with red (inset). A long TIR following the TSDs is highlighted with gray colour.

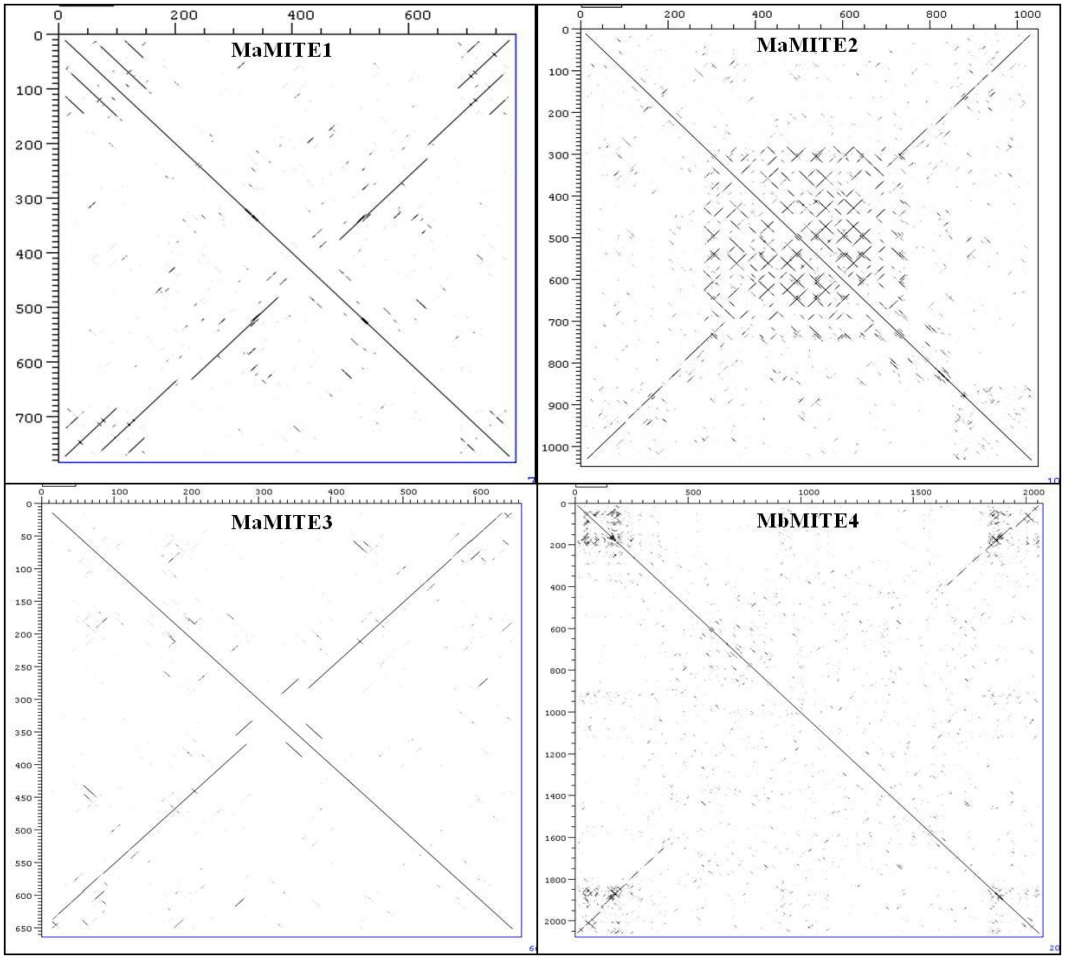


Figure 9.13: Dot plots of four MITEs identified from *Musa* plotted against themselves. The central top-left to bottom-right diagonal line represents the self-homology. The lower-left to upper-right reverse diagonals reveal the presence of >300bp long TIRs. *MaMITE1* shows tandem repeats at its ends, while *MaMITE2* has many repetitive sequences near its centre. Scales in bp.

9.2.3.1 *MaMITE1* diversity in *Musa* genome by TIP based molecular markers

Degenerate primers pair MaMITE1F 5′-AACGGGACGAGTCTTGAGAA-3′ and MaMITE1R 5′-TAAATGTCTCCGCTTAGGCC-3′ were designed from the most conserved flanking sequences of the insertion. *MaMITE1* was PCR amplified in different *Musa* genome to see its mobilization, amplification and diversification. The expected ~1052 bp fragment was amplified from some *Musa* accessions; which were mostly *Musa acuminata* diploids or triploids accessions as *Musa acuminata* (Calcutta 4, Banksii 623, Khae, Long Tavoy pied, Tomolo, Pisanglilin, Kadali, Matti, Cherukadali), AAA (Grow-michel, Monsmari) and AAB (Kumbillakannan, Vellapalayamcodan, Poomkalli). All *Musa acuminata* accessions which amplified *MaMITE1* showed only upper band indicating homozygous nature. No amplification of *MaMITE1* in *Musa balbisiana* genomes suggests the absence of this MITE at this locus (Figure 9.14a & b).

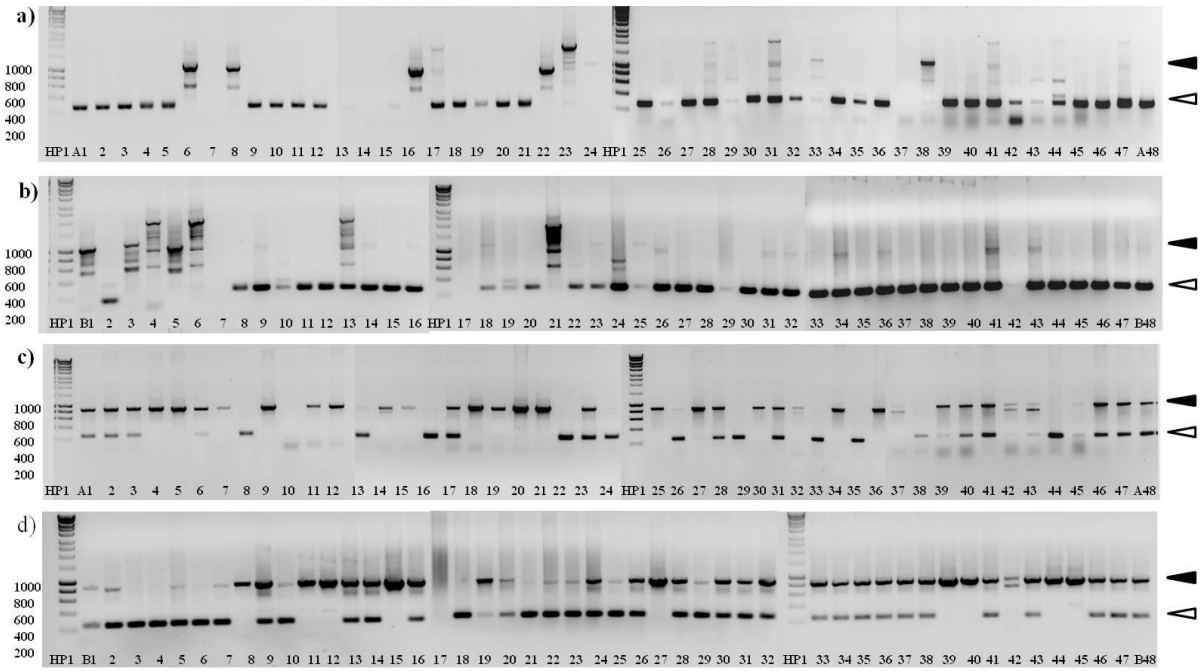


Figure 9.14: Insertional polymorphisms of MITEs and novel transposon insertions in *Musa*. a-b) *MaMITE1* in various *Musa* accessions. Filled arrowheads point to 1052 bp MITE amplification bands and lower bands (open arrowhead) amplify flanking sequences only across empty sites. c-d) The 621 bp MBT insertion sites in *Musa* accessions. Long bands (864 bp) showed the amplified MBT element and short bands amplified flanking regions (pre-insertion sites).

9.2.4 Structural features of novel transposons from unknown superfamilies

A 4192 bp insertion was detected in *Musa acuminata* BAC 'MA4_82I11' terminating with 5 bp TSDs (CATAA) and 14 bp 5′-TGTAACAcCCTTGA-3′ TIRs. Due to unusual 5 bp TSDs, 14 bp TIRs and only single hit to *Musa acuminata* 'Calcutta 4', it was named *MaSTE* (*Musa acuminata* single transposable element). No putative transposase in the insert suggested either deletion of active transposase or the degraded element of either hAT or Harbinger-like elements but no clear evidence showed its relation to any known superfamily. Due to its large size, the primers were designed from the first and last portion of the insert to amplify a product of 1138 bp and 1122 bp with primer pair MaSTE1F and MaSTE1R and primer pair MaSTE2F and MaSTE2R respectively. Interestingly, the first and last parts of insert were only amplified from *Musa acuminata* (Calcutta 4) genome, which was the only significant hit in blast searches. This strongly revealed its proliferation in *Musa acuminata* or its recent introduction to the genome (Figure 9.15a & b). The amplicons were sequenced and aligned with the reference sequence which showed a complete homology to *MaSTE*. Another 632 bp insertion was detected displaying 3 bp TSDs (ATA), 10 bp TIRs 5′-ATaATTATTG-3′, AT rich (72.8%) and having small repeats of

poly A/T starting from TIR on the upstream orientation. The structural features showed similarities to Harbingers, so it is tentatively considered to be a non-autonomous Harbinger-like element and named *MbHAL* (Figure 9.11; Table 9.1).

9.2.5 Mobile insertions/deletions without TIRs

The comparison of *Musa acuminata* (MA4_82I11) and *Musa balbisiana* (MBP_81C12) BAC sequences directed the discovery of a several novel insertions/deletions without any recognizable TIRs but other structural features: the flanking TSDs at gap-insertion pairs and amplification polymorphisms strongly suggested their mobile nature. The elements were studied by computational and molecular analysis to characterize them and see their diversity in various *Musa* genomes. An insertion (1676 bp) flanked by 2 bp TSD 'AG' and lacking any visible TIRs was detected and named '*MAWA*'. Blasting *MAWA* element against Repbase database of transposons showed hits to a Gypsy Monkey_MA element indicating its nesting position in Gypsy retrotransposon. *MAWA* was amplified in *Musa* genomes with degenerative primer pair MAWAF and MAWAR. The primers were designed from the conserved flanking sequences common to *Musa acuminata* and *Musa balbisiana*. *MAWA* elements showed a medium level of amplification in many *Musa* genomes including *Musa balbisiana* (BB) and the *Musa* triploids (AAB, ABB) suggesting the contribution of B-genome (Figure 9.15c & d). A small insert of 367 bp long was identified implanted in *Musa balbisiana* (MBP_81C12) accession from 53800-54166 bp. Due to its small size and no homology to any known superfamily; it is named *MUST* (*Musa* small transposon). *MUST* is flanked by 'AA' TSDs and lack any detectable TIRs. The flanking regions of the insertion are GC rich (52-58%), while insert itself is AT rich (59.4%). Blast searches showed very surprising results with single hits from *Musa balbisiana* clone BAC MBP_81C12, from where it was originally identified (Figure 9.11; Table 9.1).

A small insertion was identified with 3 bp TSD (ATG), but no TIRs or any protein domain with high AT rich (66%) regions. The element is named MBT (*Musa* balbisiana transposon). The MBT transposon was amplified by using a pair of degenerative primers MBTF 5′-GATCAAATGGGGAAGCAACC-3′ and MBTR 5′-ACTTCTCCCGTGTGTGTCGT-3′ designed from the conserved flanking sequences. The short bands (~240 bp amplifying the flanking regions) and long bands (~864 bp amplifying the transposon insertion) were

amplified from various genomes. PCR analysis showed insertion polymorphisms of MBT in almost all tested *Musa* genomes suggesting its distribution and abundance in *Musa* genomes. Some of the genomes amplified only long or short band, while others amplified both bands (heterozygous). The genomes amplifying the insertion were the various accessions from *Musa acuminata*, *Musa balbisiana* and *Musa* triploids (AAA, AAB, AAB) (Figure 9.14c & d).

A 504 bp insertion was found in *Musa acuminata* (MA4_82I11) BAC from 65047-65550 bp. The insertion was flanked by 19 bp TA microsatellite and is named as TATA element due to generating TATA-like TSDs. A 261 bp TE was identified in *Musa acuminata* from 67426-67686 with 2 bp (TA) TSD. The insertion lacks any characteristic feature and was named MAX (*Musa acuminata* unknown). The insertion was AT rich with one third of GC content. The Mariner-like elements or Stowaway MITEs generates TA TSDs, but due to lack of any TIRs and very low copy numbers, the element cannot be sorted to its respective superfamily. The smallest insertion investigated in this study was a 82 bp insert in *Musa balbisiana* (MBP_81C12) from 100,030-100,111. It was named as 'TINY' due to its smallest size. It is flanked by 5 bp TSDs (TGGTC) and lacks any detectable TIRs. The GC content of the insert was 42.7%, which corresponds to the low GC content of all other transposons (Table 9.1).

9.3 Discussion

With exploitation of bioinformatics and computational analysis, various DNA transposons belonging to different superfamilies were identified which had been inserted or deleted during and after the divergence of *Musa* species. Most of the elements were characterized and classified into their previously known families, but several transposon insertions were novel exhibiting different hallmarks (TSDs or TIRs) or generating a TSD without any TIRs and internal coding regions. Among DNA transposons, the non-autonomous hATs or hAT-like MITEs as named by several authors are predominant, followed by other MITEs. The investigated hATs generate 8 bp TSDs, 9-30 bp TIRs and conserved sub-terminal regions; all characteristic hATs features. These structural features were studied in hAT elements from plants, animals, fungi and flies (Rubin *et al.*, 2001; Huang *et al.*, 2009; de Freitas Ortiz *et al.*, 2010). The investigated elements have shown some dispersed copies of tetra and penta-nucleotide sequences in their sub-terminal regions, which are thought to be the

remnants of binding sites for transposase. These kind of small sequences were also observed in maize transposon nDaiZ, which contains 11 copies of such repetitive tetra-nucleotides (ACCC and GGGT) (Huang *et al.*, 2009). Due to the presence of remnants for binding sites, there is an indication that they were the part of large autonomous transposons, which during their life cycle deleted their active transposase and other motifs necessary for transposition (Kidwell and Lisch, 2001).

9.3.1 Non-autonomous hATs are dominant in *Musa* genomes

Four non-autonomous hAT families were identified by comparing two *Musa* BAC sequences, while autonomous partner of *MaN-hAT1* and *MaN-hAT2* were detected from other *Musa* BAC clone sequences. The non-autonomous *MaN-hAT1* yielded >10 fold of copies (34 copies) as compared to its autonomous partner '*MahAT1'* (3 copies) in blast searches. Similarly, *MaN-hAT2* resulted in 5 folds (19 copies) homologues in comparison to its autonomous element *MaN-hAT2* (4 copies). Based on these retrieved intact copies; the number of copies for each element was estimated. A total of ~11800 non-autonomous hATs were estimated residing in *Musa acuminata* and *Musa balbisiana*, while their autonomous copies were 1285 only, indicating that non-autonomous hATs are >10 fold more common in *Musa* genomes. These results confirmed the investigations of Rubin *et al.*, (2001) which also reported 1:10 ratio of non-autonomous to autonomous hATs collected from various animal, fungi and plant genomes.

9.3.2 The hATs are an ancient and abundant superfamily of transposons in *Musa*

The hATs are the most prevalent superfamily in many plants including *Musa*. Around 1000 hATs related complete or partial copies were found by BLASTN searches of *Musa* genomes. Out of these, only 70 copies were intact, while all others were partial copies or deletion derivatives, which indicate that the degraded copies of hATs are still persisting in *Musa* genomes. The higher proportion of defective or partial copies makes it the most ancient superfamily of transposons. The non-autonomous hATs investigated in present study still retain evidence for their mobility, which is indicated by the presence of their TIRs, TSDs and conserved sub-terminal regions. The hATs from different families were aligned and high heterogeneity was found in their internal regions, except TIRs, which showed homology within few families. The ancient nature of hAT superfamily in

investigated in several eukaryotic genomes. Due to their ancient nature, a high variability and a high proportion of defective and partial elements can be expected (Rubin *et al*., 2001).

Our results showed the presence of several copies of non-autonomous hATs, which are still active in *Musa* genomes. We suggest that the proliferating of non-autonomous hATs might be the result of utilizing the enzymatic machinery (transposase) of related autonomous hATs. Many degraded copies or deletion derivatives of these hATs are found in *Musa* genome. The previous investigations suggested that degradation mechanisms are the last phases of genomic colonization by TEs, where such copies are inactivated and slowly eliminated or split into fragments (Le Rouzic *et al.*, 2007).

9.4 Conclusion

In *Musa*, it was notable that not all superfamilies of elements were present: although 7 superfamilies are reported as being ubiquitous in plants (Wicker *et al.*, 2003), no CACTA nor Mutator elements were found in *Musa*. Some MITEs had 9 bp TSDs and long TIRs signatures which could suggest derivation from Mutator-like elements, but no homology was found with any known Mutator element, and it seems likely that the ancestor of *Musa* included both CACTA and Mutator elements, indicating that they have been swept from the genome. Such events are unusual but reported: for example, the lack of Copia-like elements in mammalian genomes, despite present in plants, fungi and most animal taxa (Flavell *et al.*, 1998). Our results showed the abundance of hATs, MITEs and several novel mobile elements in *Musa* with conserved structural features, not found in known superfamilies. The mobility and high activity of these elements has contributed to the genomic duplication and variability of *Musa* species.

CHAPTER 10
CONCLUSIONS

10.1 Conclusions structure

These general conclusions will give a short overview of the progress towards the aims presented in the Introduction based on the seven chapters detailing specific results of this work. While some aspects of broader significance have been presented in these chapters, some future needs for work and the prospects will be discussed briefly here.

10.1.1 Mobile elements in diploid and polyploid *Musa* and *Brassica* crops

The exploitation of homoeologous BAC sequences from pairs of species within the two genera studied here, *Musa* and *Brassica*, proved to be efficient in identifying mobile elements of both Class I (retrotransposons and non-LTR retrotransposons) and Class II (DNA transposons). The approach taken based on a bioinformatic, dot plot and comparative analysis was not based upon known sequences or homologies, but driven by data. In the pairs of BACs compared here, all the regions corresponding to a gap in one species with sequence in another species were investigated in detail. As shown in the data chapters, most of these sequences belonged to orders and superfamilies of transposable elements that have been identified in other species. However, some of the elements showed characteristic TIRs, TSDs and gene structures that make them candidates for definition as new superfamilies, but it will be necessary to search for these structures in more species to define the universal nature, abundance, and conserved characteristics. Notably, a very high proportion (more than 80%) of all the gaps/insertion pairs were transposable element related, suggesting that spontaneous genome insertion/deletions are very rare events (within the size limits defined here, which would for example include the length of most introns).

10.1.2 Small mobile element structures

Small, non-autonomous transposable elements have few conserved components, there may be no genes/open reading frames, and TSDs and TIRs may be short, variable and imperfect. The BAC targeting approach here was able to identify many such elements in

both *Brassica* and *Musa* showing that indeed many different families are active in recent evolutionary time. In other reports, particular motifs have been found to be abundant, but it has not previously been possible to identify the full population of short mobile element structures. There are many software approaches and methodologies to identify transposable elements, but they are not efficient to identify non-autonomous TEs especially with variable structural features (varied TSDs, no TIRs and short sizes). In contrast, the comparison of homoeologous sequences displays a complete profile of any insertion/deletion due to the activity of these mobile elements.

10.1.3 Autonomous transposable element families

The analysis here showed that most orders and superfamilies of transposable element were both present and active within the *Brassica* and the *Musa* genomes. Notably, Mutator, CACTA and Harbinger elements were missing from *Musa*, but highly abundant in *Brassica.* Although there are many proteins characteristic of non-LTR LINE elements in the genomic sequence, no LINE elements were found in *Musa* BACs. SINEs were found to be very active with many polymorphisms in *Brassica*, but not in *Musa*. The evolutionary history and activity of each superfamily was different and showed contrasting characteristics in the different species with the two genera. Some elements were specific to one genome (A, B in *Musa*; A, C in *Brassica*). These suggest that the controls (whether through RNAi, methylation, and genome positioning) on the activity of elements will be different, and would provide an interesting topic for further study. In the polyploids, activity of genome specific elements, and movement between genomes leading to homogenization, will be important for future study, with implications over timescales from the years relevant to plant breeding and alien chromosome introgression, to the tens of millions of years that defines the impact of whole genome duplications on angiosperm speciation.

10.1.4 Transposon marker development and exploitation

The primers designed here proved to be both well conserved from flanking sequences and robust over the range of 40 and 96 accessions tested in *Brassica* and *Musa* respectively showing that these PCR based markers are valuable for detecting polymorphisms. The known mobility was valuable in targeting elements of interest. Both those that are genome-

specific, and polymorphic within genomes, will be useful for different studies in a plant breeding context: for identifying chromosome origin, and for varietal identification or pedigree analysis. These molecular markers can be used to identify the unknown accessions or cultivars, as in the present study a commercial *Brassica* variety named NATCO was used without knowing its genomic composition and the developed markers clearly sorted its position in *Brassica juncea* (later confirmed by cytogenetic analysis).

10.2 Dot plot: a highly effective method for transposable element identification

The dot plot approach used in the present study was highly informative in the identification of all types of TEs including the large LTR retrotransposons and very small SINEs or MITE elements in the compared BAC genomic sequences (Figure 10.1-10.3). Dot plots of the sequences were used to identify and detect LTR retrotransposons, LARDs, TRIMs or the Mutator-like MITEs. The parallel lines across the diagonal lines represent the long terminal repeat of retrotransposons, while the inverted lines indicate the terminal inverted repeats of Mutator-like MITEs. Their starting and ending points of LTRs and TIRs can be defined by sliding the window from the start to the end of the LINEs. Two BAC sequences with high homology in sequences were used to see the insertion or deletion of any mobile TE by gap-insertion pairs. The DNA transposon insertions can be easily identified by TSDs in their terminal ends and TIRs internal to the TSDs. The deletions can be identified by having footprints of TSDs. The effectiveness of this method can be measured by the results, where >90% investigated gap-insertion pairs were TEs of one or other superfamilies (Figure 10.1-10.3). Similarly the identification of LTR retrotransposons by various softwares such as 'LTR FINDER' and 'LTR_STRUC' were compared with dot plot identification of LTR retrotransposons and dot plot was found as most efficient method to detect all LTR retrotransposons, while the other programs failed to detect all elements. Some mobile insertions with features not common to the known superfamilies were identified by dot plot method, not detected by other softwares. The PCR amplification polymorphism of these insertions indicated their mobile nature. This method is straightforward, if more time consuming than alternatives, but highly efficient, precise and informative in identifying larger and smaller TEs. The figures 10.1-10.3 have shown that nearly all different superfamilies of TEs (Copia, Gypsy, LARDs, TRIMs, LINEs, SINEs, CACTA, Harbinger, Mutator, MITEs and unknown TEs) can be identified and characterized by comparison of homeologous and homologous BAC sequences.

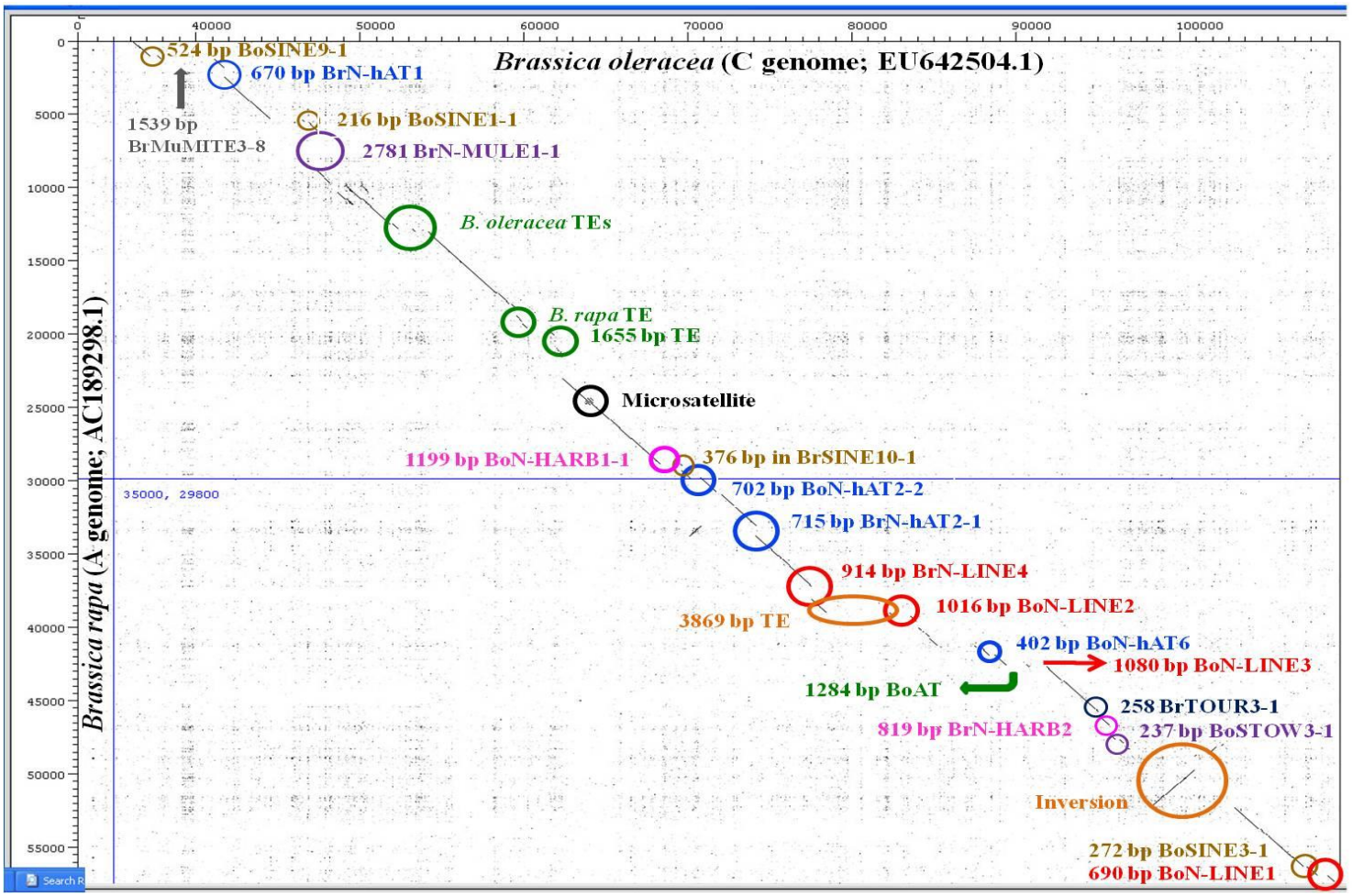


Figure 10.1: Dot plot of homoeologous BAC clones *Brassica oleracea* (EU642504.1) (horizontal) against *Brassica rapa* (AC189298.1) (vertical). The comparison of the BACs showed a large homologous region (continuous diagonal line) broken by multiple gap-insertion pairs. The gaps showed transposon insertions from various LINEs, SINEs (retrotransposons), hATs, Mutator, Harbinger (DNA transposons), MITEs and novel insertions shown by various colours. The details of the elements are given in respective chapters. Many small insertions <500 bp are not highlighted here.

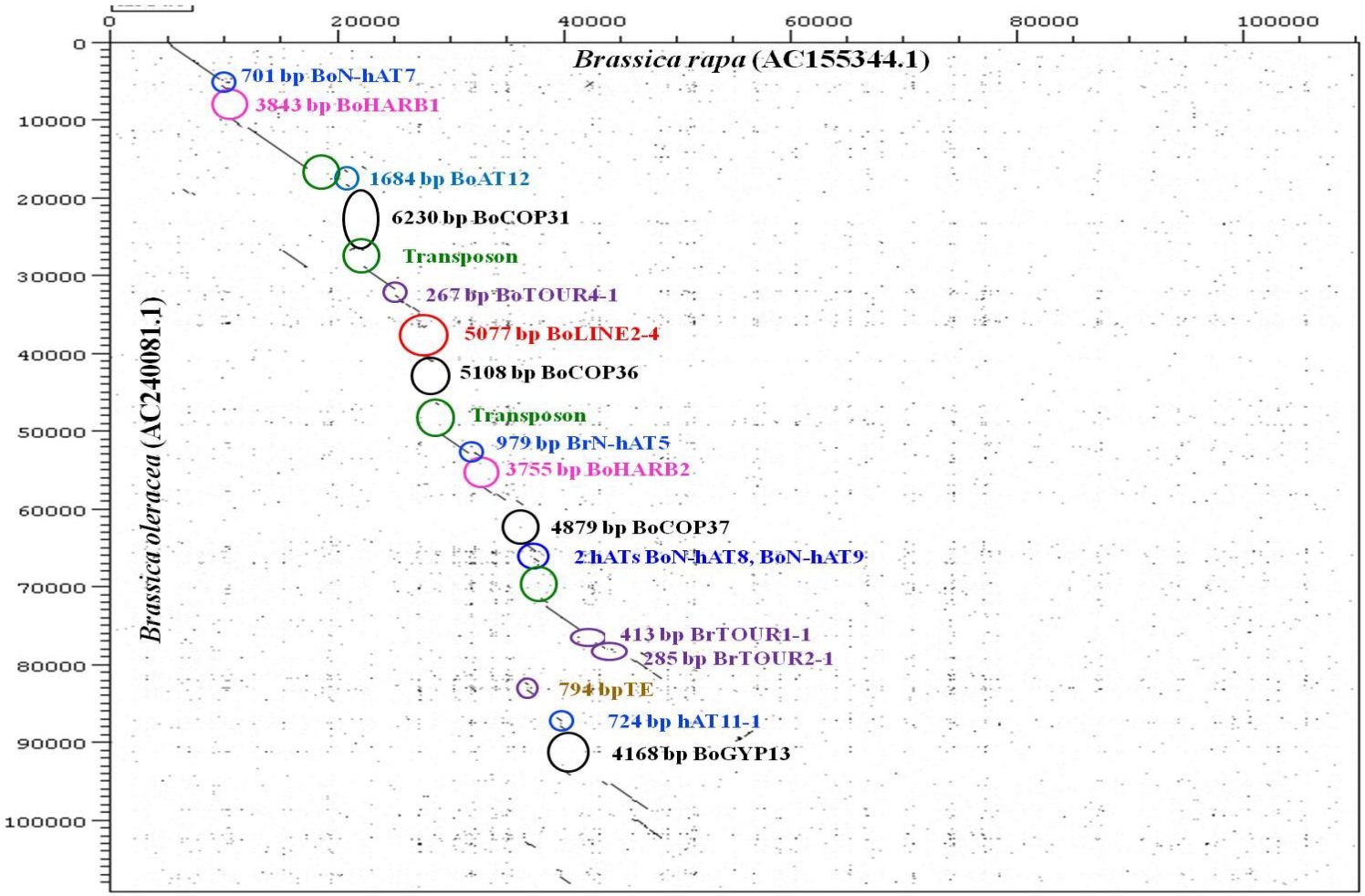


Figure 10.2: Dot plot of homoeologous BAC clones *Brassica rapa* (AC155344.1) (horizontal) against *Brassica oleracea* (AC240081.1) (vertical). The gaps showed transposon insertions from various Copia, Gypsy, LINEs, SINEs (retrotransposons), hATs, Harbinger (DNA transposons), MITEs and novel insertions. The uncharacterized elements are represented by green colours. The details of the elements are given in respective chapters. Many small insertions <500 bp are not highlighted here.

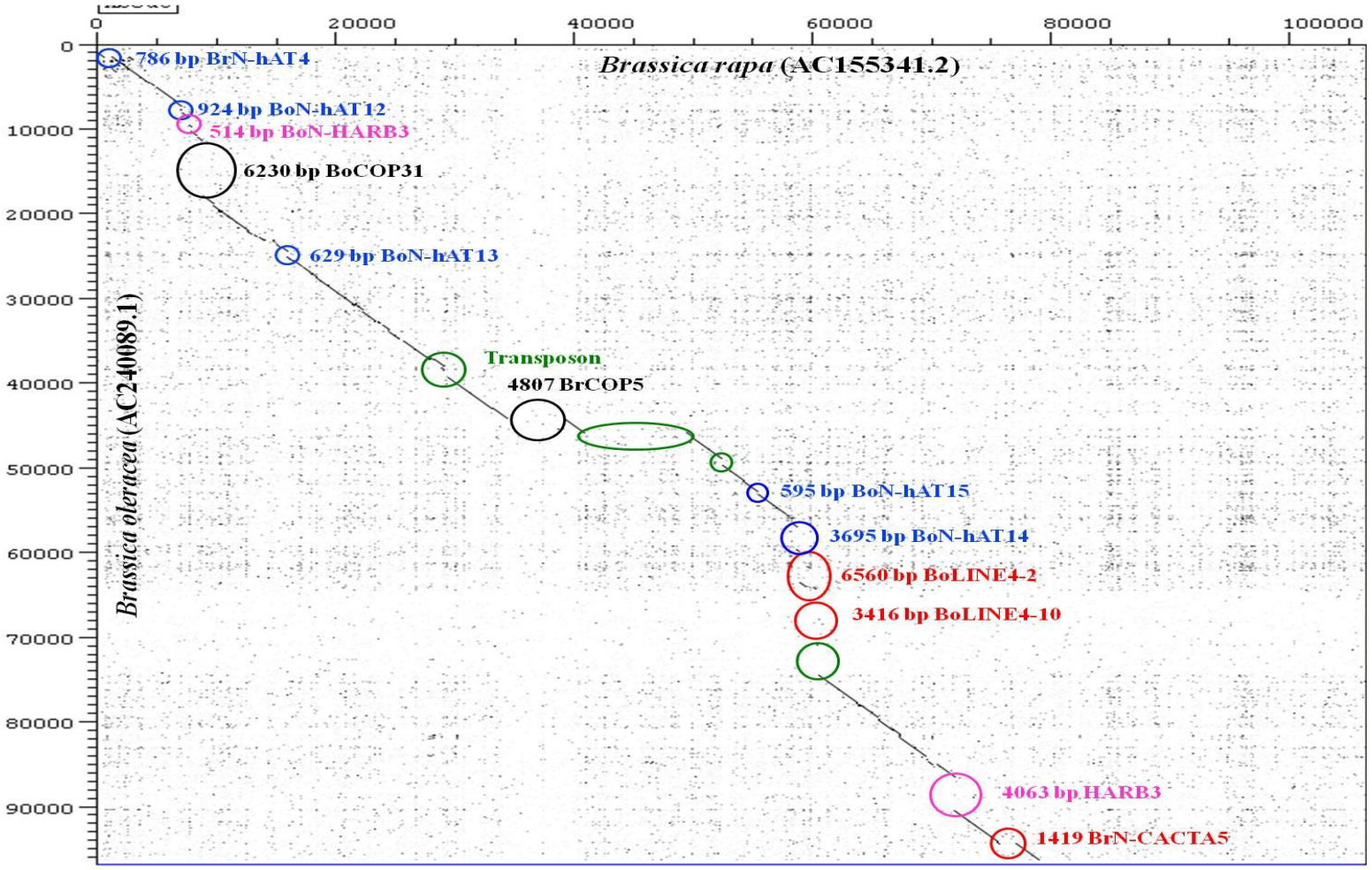


Figure 10.3: Dot plot of homoeologous BAC clones *Brassica rapa* (AC155341.2) (horizontal) against *Brassica oleracea* (AC240089.1) (vertical). The gaps showed transposon insertions from various LINEs, SINEs (retrotransposons), hATs, CACTA, Harbinger (DNA transposons), MITEs and novel insertions. The details of the elements are given in respective chapters. Many small insertions <500 bp are not highlighted here.

10.3 Total copy numbers of TE superfamilies in *Brassica* genome

The genome of *Brassica rapa* and *Brassica oleracea* is rich in various types of TEs. Almost all different types of Class I retrotransposons and Class II DNA transposons investigated in other plant genomes were identified in *Brassica* genomes. The total estimated copies from the whole genome indicated that the Copia family contained the highest numbers of copies in comparison to other LTR retrotransposons and the numbers were much higher in *Brassica oleracea* (7540) as compared to *Brassica rapa* (1596) (Figure 10.4). The TRIM elements were fewer in numbers in both A and C-genome *Brassica*. The SINEs were also abundant in *Brassica oleracea* and less frequent in *Brassica rapa*. Among DNA transposons the non-autonomous hATs were very high in numbers in both *Brassica rapa* and *Brassica oleracea*. The MITEs derived from Mariner (Stowaway), Harbingers (Tourist) and Mutator-like MITEs were very high in number; predominant in *Brassica oleracea*. Very few copies of Mutator and Harbinger-like elements were identified and Mariner are only represented by their MITEs (Stowaway) derivatives. The TE elements with unknown superfamilies were also identified in the present study. In general, the number of copies of each superfamily is higher in C-genome as compared to A-genome.

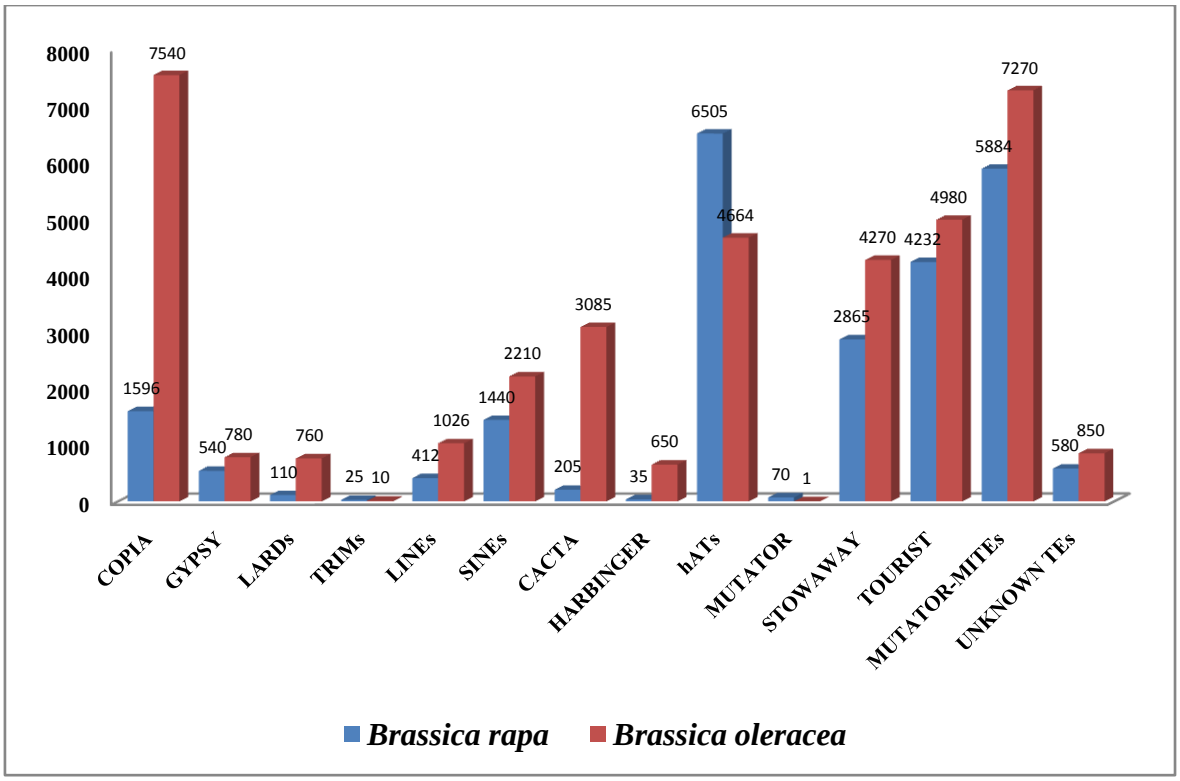


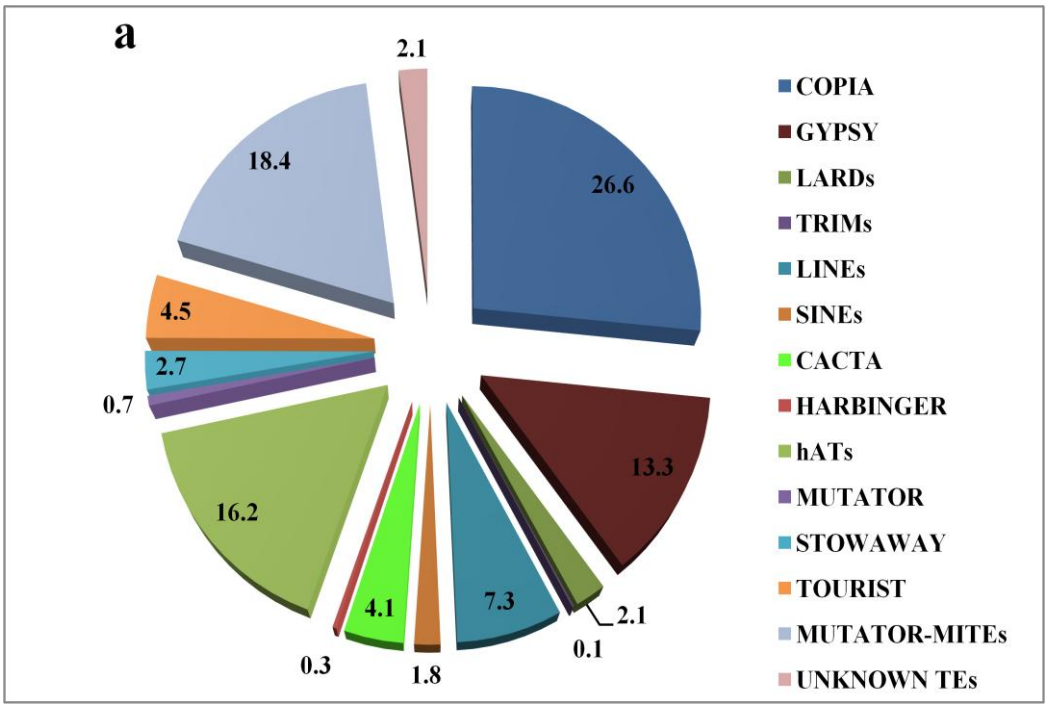
Figure 10.4: Estimated copy numbers of each superfamily of transposable elements in *Brassic rapa* and *Brassica oleracea* whole genomes.

10.4 Relative percentages of TE superfamilies in the *Brassica* genomes

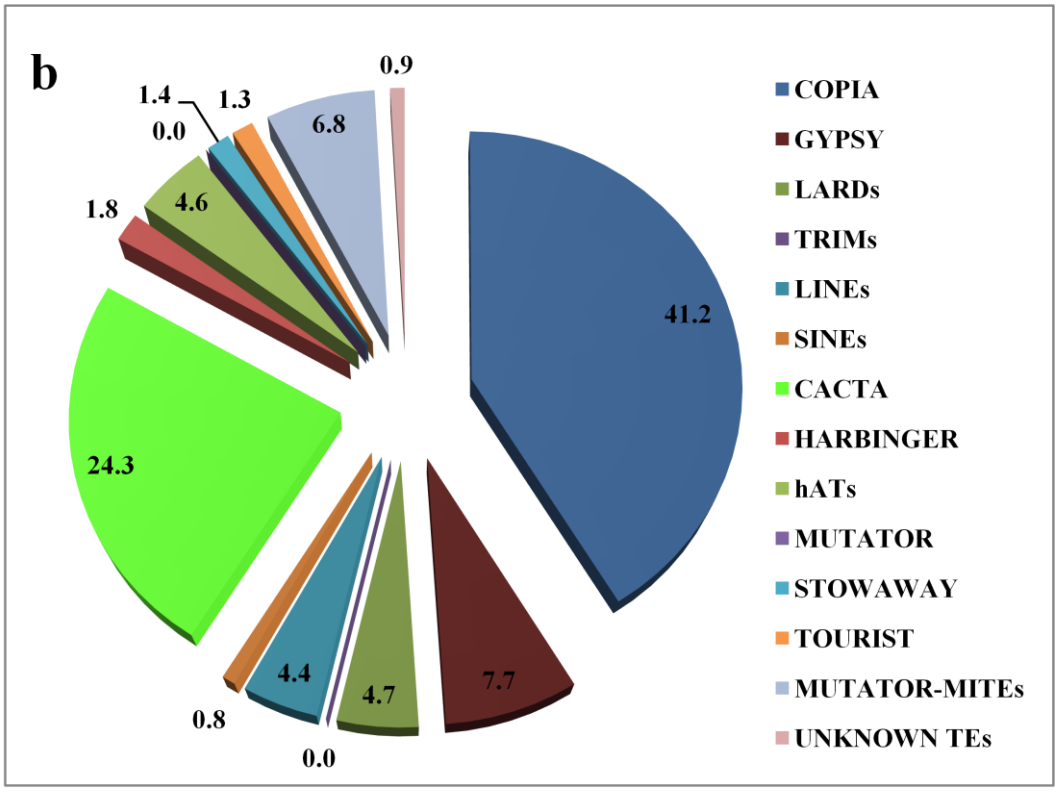
The relative percentage of each TE superfamily was calculated in *Brassica rapa* and *Brassica oleracea*. The Copia elements showed a very high percentage as compared to other TEs. The Gypsy superfamily covered a high proportion after Copia. The LARDs superfamily also covered a substantial proportion (2.1%) of A-genome and 4.6% in C-genome *Brassica*. TRIMs were less active in both genomes and covered the smallest proportion of all TEs investigated in *Brassica* genomes (Figure 10.5a & b). LINEs were more abundant in A-genome (7.3%), while in C-genome they are only 4.4% of all TE proportion. SINEs were very high copy numbered elements in *Brassica* (predominant in *Brassica oleracea*) but only covered 0.8-1.8% of all TEs due to their small sizes.

Among DNA transposons, the major proportion (24.3%) is covered by *Brassica oleracea* CACTA elements, while they are less frequent (4.1%) in *Brassica rapa* (Figure 10.5a & b). The non-autonomous hATs were distributed in high copy numbers but due to their small sizes, they cover very less proportion of TEs in *Brassica rapa* (16.2%) and *Brassica oleracea* (4.7%). Harbinger and Mutator-like TEs are very less frequent DNA transposons in both *Brassica rapa* and *Brassica oleracea,* while all the MITEs (Stowaway, Tourist, Mutator-like) covered major portion of TEs. Among the MITEs, Mutator-like MITEs constitute the major proportion in *Brassica rapa* (18.4%) and *Brassica oleracea* (6.8%) due to their larger sizes and high copy numbers. The Stowaway and Tourist-like MITEs have shown high activity in A-genome.

In *Brassica oleracea* nearly half the proportion is composed of LTR retrotransposons, while CACTA covered a quarter of all TEs. Several novel insertions were also identified with unknown superfamilies and were included in a group of unknown TEs. The unknown elements covered a small proportion (<2.1%) as compared to other superfamilies of TEs in *Brassica* genomes (Figure 10.5a & b).



Brassica rapa



Brassica oleracea

Figure 10.5: Percentage of each TE superfamily in a) *Brassica rapa* and b) *Brassica oleracea*. The Copia elements have shown very high percentage as compared to other TEs.

10.5 Future developments

During the work in this thesis, DNA sequencing costs have reduced many folds, although the effort required for assembly and analysis has changed by a much smaller amount. Indeed, many whole genome sequences published in the last two years have many gaps or miss repetitive and mobile sequences. The elements defined from BACs as in this work will be useful to assist with efficient and complete genome sequence assembly approaches in the future.

Markers for identification of chromosomes and genomes are important in wide hybridization and alien introgression programmes which have enable plant breeders to exploit variation from diverse germplasm. Both the genera studied here have interesting challenges for crop improvement. Both have several related species with various chromosome numbers, and polyploids are common. There is a need for a robust phylogeny showing the relationship of the wild and cultivated relatives. Germplasm resources are critical to collect, characterize at molecular and phenotypic (including biotic and abiotic stress resistance) levels. Interspecific hybridization is also likely to be very important to increase the diversity available to breeders. Banana has the additional problem of sterility and parthenocarpy, attributes required in the crop. In the next months, the genome specific markers here will be tested in both *Brassica* and *Musa* to identify transfer of chromosomes between species in intercrosses. As well as use of flanking PCR markers, it will also be important to test the mobile transposable element fragment further by Southern and *in situ* hybridization to see if they are high copy number but still retaining genome or even chromosome specificity by hybridization. The work will also need to be expanded to see if any sub-genome specific markers can be developed, and test the primers in more distant species of interest for breeders, whether in other *Brassicaceae* genera (eg *Erucastrum, Diplotaxis* and *Enarthrocarpus*) or *Musaceae/Zingiberales* species such as *Ensete* or ginger.

Understanding the past evolutionary behaviour of transposable elements within the genome is a key to exploiting them as both markers and sources of variability in the future. In this work, the genomic studies have made a significant advance towards identification of the nature, structure and mobility of transposable elements as a major genomic component of genomes studied here.

REFERENCES

Ahmed S, Shafiuddin M, Azam MS, Islam MS, Ghosh A, Khan H. 2011. Identification and characterization of jute LTR retrotransposons:: Their abundance, heterogeneity and transcriptional activity. *Mob Genet Elements,* **1**: 18-28.

Alix K, Heslop-Harrison JS. 2004. The diversity of retroelements in diploid and allotetraploid *Brassica* species. *Plant Mol Biol,* **54**: 895-909.

Alix K, Joets J, Ryder CD *et al.* 2008. The CACTA transposon Bot1 played a major role in *Brassica* genome divergence and gene proliferation. *Plant J,* **56**: 1030-44.

Altschul SF, Gertz EM, Agarwala R, Schaffer AA, Yu YK. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res,* **37**: 815-24.

Altschul SF, Madden TL, Schaffer AA *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* **25**: 3389-402.

Antonius-Klemola K, Kalendar R, Schulman AH. 2006. TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. *Theor Appl Genet,* **112**: 999-1008.

Bancroft I, Morgan C, Fraser F *et al.* 2011. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nat Biotechnol,* **29**: 762-6.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the *Brassicales*: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other *Brassicales*. *Genome Biol Evol,* **1**: 391-9.

Benjak A, Boue S, Forneck A, Casacuberta JM. 2009. Recent amplification and impact of MITEs on the genome of grapevine (*Vitis vinifera* L.). *Genome Biol Evol,* **1**: 75-84.

Bennett MD, Leitch IJ. 2011. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann Bot,* **107**: 467-590.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol,* **42**: 251-69.

Bernet GP, Asins MJ. 2003. Identification and genomic distribution of gypsy like retrotransposons in Citrus and Poncirus. *Theor Appl Genet,* **108**: 121-30.

Bundock P, Hooykaas P. 2005. An *Arabidopsis* hAT-like transposase is essential for plant development. *Nature,* **436**: 282-4.

Bureau TE, Ronald PC, Wessler SR. 1996. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc Natl Acad Sci U S A,* **93**: 8524-9.

Bureau TE, Wessler SR. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell,* **4**: 1283-94.

Bureau TE, Wessler SR. 1994a. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. *Proc Natl Acad Sci U S A,* **91**: 1411-5.

Bureau TE, Wessler SR. 1994b. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell,* **6**: 907-16.

Capy P. 2005. Classification and nomenclature of retrotransposable elements. *Cytogenet Genome Res,* **110**: 457-61.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res,* **37**: D93-7.

Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol,* **58**: 377-406.

Cheng X, Zhang D, Cheng Z, Keller B, Ling HQ. 2009. A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. *Genetics,* **181**: 1183-93.

Cheng XD, Ling HQ. 2006. [Non-LTR retrotransposons: LINEs and SINEs in plant genome]. *Yi Chuan,* **28**: 731-6.

Cheung F, Town CD. 2007. A BAC end view of the *Musa acuminata* genome. *BMC Plant Biol*, **7**: 29.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nat Rev Genet,* **6**: 836-46.

de Freitas Ortiz M, Lorenzatto KR, Correa BR, Loreto EL. 2010. hAT transposable elements and their derivatives: an analysis in the 12 *Drosophila* genomes. *Genetica,* **138**: 649-55.

de Jesus EM, Ochoa Cruz EA, Cruz GM, Van Sluys MA. 2012. Diversification of hAT transposase paralogues in the sugarcane genome. *Mol Genet Genomics*.

De Keukeleire P, De Schepper S, Gielis J, Gerats T. 2004. A PCR-based assay to detect hAT-like transposon sequences in plants. *Chromosome Res,* **12**: 117-23.

Deragon JM, Casacuberta JM, Panaud O. 2008. Plant transposable elements. *Genome Dyn,* **4**: 69-82.

Deragon JM, Gilbert N, Rouquet L, Lenoir A, Arnaud P, Picard G. 1996. A transcriptional analysis of the S1Bn (*Brassica napus*) family of SINE retroposons. *Plant Mol Biol,* **32**: 869-78.

Deragon JM, Landry BS, Pelissier T, Tutois S, Tourmente S, Picard G. 1994. An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. *J Mol Evol,* **39**: 378-86.

Deragon JM, Zhang X. 2006. Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst Biol,* **55**: 949-56.

Doak TG, Doerder FP, Jahn CL, Herrick G. 1994. A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc Natl Acad Sci U S A,* **91**: 942-6.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature,* **284**: 601-3.

Doyle, J.J. and Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. *Focus* **12**: 13-15.

Drummond AJ, Ashton B, Buxton S *et al.* 2011. Geneious v5.4, Available from http://www.geneious.com/.

Du C, Hoffman A, He L, Caronna J, Dooner HK. 2011. The complete Ac/Ds transposon family of maize. *BMC Genomics,* **12**: 588.

Du J, Tian Z, Hans CS *et al.* 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J,* **63**: 584-98.

Eickbush TH, Jamburuthugoda VK. 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res,* **134**: 221-34.

FAOStat. 2012. FAOSTAT http://faostat.fao.org/. Accessed March, 2012.

Fawcett JA, Kawahara T, Watanabe H, Yasui Y. 2006. A SINE family widely distributed in the plant kingdom and its evolutionary history. *Plant Mol Biol,* **61**: 505-14.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet,* **9**: 397-405.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet,* **3**: 329-41.

Feschotte C, Mouches C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. *Mol Biol Evol,* **17**: 730-7.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet,* **41**: 331-68.

Feschotte C, Wessler SR. 2002. Mariner-like transposases are widespread and diverse in flowering plants. *Proc Natl Acad Sci U S A,* **99**: 280-5.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet,* **5**: 103-7.

Flavell AJ. 1984. Role of reverse transcription in the generation of extrachromosomal copia mobile genetic elements. *Nature,* **310**: 514-6.

Flavell AJ. 1992. Ty1-copia group retrotransposons and the evolution of retroelements in the eukaryotes. *Genetica,* **86**: 203-14.

Flavell AJ. 1999. Long terminal repeat retrotransposons jump between species. *Proc Natl Acad Sci U S A,* **96**: 12211-2.

Flavell AJ, Dunbar E, Anderson R, Pearce SR, Hartley R, Kumar A. 1992a. Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. *Nucleic Acids Res,* **20**: 3639-44.

Flavell AJ, Jackson V, Iqbal MP, Riach I, Waddell S. 1995. Ty1-copia group retrotransposon sequences in amphibia and reptilia. *Mol Gen Genet,* **246**: 65-71.

Flavell AJ, Knox MR, Pearce SR, Ellis TH. 1998. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *Plant J,* **16**: 643-50.

Flavell AJ, Pearce SR, Heslop-Harrison P, Kumar A. 1997. The evolution of Ty1-copia group retrotransposons in eukaryote genomes. *Genetica,* **100**: 185-95.

Flavell AJ, Pearce SR, Kumar A. 1994. Plant transposable elements and the genome. *Curr Opin Genet Dev,* **4**: 838-44.

Flavell AJ, Smith DB. 1992. A Ty1-copia group retrotransposon sequence in a vertebrate. *Mol Gen Genet,* **233**: 322-6.

Flavell AJ, Smith DB, Kumar A. 1992b. Extreme heterogeneity of Ty1-copia group retrotransposons in plants. *Mol Gen Genet,* **231**: 233-42.

Friesen N, Brandes A, Heslop-Harrison JS. 2001. - Diversity, origin, and distribution of retrotransposons (gypsy and copia) in conifers. *Mol Biol Evol,* **18**: 1176-88.

Fujino K, Sekiguchi H. 2011. Transposition behavior of nonautonomous a hAT superfamily transposon nDart in rice (*Oryza sativa* L.). *Mol Genet Genomics,* **286**: 135-42.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell,* **19**: 3403-17.

Gbadegesin MA, Beeching JR. 2010. Enhancer/Suppressor mutator (En/Spm)-like transposable elements of cassava (*Manihot esculenta*) are transcriptionally inactive. *Genet Mol Res,* **9**: 639-50.

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**: 305-17.

Ge XH, Wang J, Li ZY. 2009. Different genome-specific chromosome stabilities in synthetic *Brassica* allohexaploids revealed by wide crosses with *Orychophragmus*. *Ann Bot,* **104**: 19-31.

Gierl A. 1996. The En/Spm transposable element of maize. *Curr Top Microbiol Immunol,* **204**: 145-59.

Gierl A, Saedler H. 1989. The En/Spm transposable element of *Zea mays*. *Plant Mol Biol,* **13**: 261-6.

Gierl A, Schwarz-Sommer Z, Saedler H. 1985. Molecular interactions between the components of the En-I transposable element system of *Zea mays*. *Embo J,* **4**: 579-83.

Goubely C, Arnaud P, Tatout C, Heslop-Harrison JS, Deragon JM. 1999. S1 SINE retroposons are methylated at symmetrical and non-symmetrical positions in *Brassica napus*: identification of a preferred target site for asymmetrical methylation. *Plant Mol Biol,* **39**: 243-55.

Grzebelus D, Gladysz M, Macko-Podgorni A *et al.* 2009. Population dynamics of miniature inverted-repeat transposable elements (MITEs) in *Medicago truncatula*. *Gene,* **448**: 214-20.

Grzebelus D, Lasota S, Gambin T, Kucherov G, Gambin A. 2007. Diversity and structure of PIF/Harbinger-like elements in the genome of *Medicago truncatula*. *BMC Genomics,* **8**: 409.

Grzebelus D, Simon PW. 2009. Diversity of DcMaster-like elements of the PIF/Harbinger superfamily in the carrot genome. *Genetica,* **135**: 347-53.

Grzebelus D, Yau YY, Simon PW. 2006. Master: a novel family of PIF/Harbinger-like transposable elements identified in carrot (*Daucus carota* L.). *Mol Genet Genomics,* **275**: 450-9.

Guignon V, Droc G, Alaux M *et al.* 2012. Chado Controller: advanced annotation management with a community annotation system. *Bioinformatics*.

Han CG, Frank MJ, Ohtsubo H, Ohtsubo E. 2000. New transposable elements identified as insertions in rice transposon Tnr1. *Genes Genet Syst,* **75**: 69-77.

Han Y, Korban SS. 2007. Spring: a novel family of miniature inverted-repeat transposable elements is associated with genes in apple. *Genomics,* **90**: 195-200.

Hansen CN, Harper G, Heslop-Harrison JS. 2005. Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenet Genome Res,* **110**: 559-65.

Hansen CN, Heslop-Harrison JS. 2004. Sequences and Phylogenies of Plant Pararetroviruses, Viruses and Transposable Elements. *Adnances in Botanical Research,* **41**: 165-193.

Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. *Genome Biol,* **5**: 225.

Hawkins JS, Hu G, Rapp RA, Grafenberg JL, Wendel JF. 2008. Phylogenetic determination of the pace of transposable element proliferation in plants: copia and LINE-like elements in *Gossypium*. *Genome,* **51**: 11-8.

Heitkam T, Schmidt T. 2009. BNR - a LINE family from *Beta vulgaris* - contains a RRM domain in open reading frame 1 and defines a L1 sub-clade present in diverse plant genomes. *Plant J,* **59**: 872-82.

Heslop-Harrison J. 2011. Genomics, Banana Breeding and Superdomestication. Proc. Int'l ISHS-Pro*Musa* Symp. on Global Perspectives on Asian Challenges. *Acta Hort. 897*.

Heslop-Harrison JS. 2010. Genes in evolution: the control of diversity and speciation. *Ann Bot,* **106**: 437-8.

Heslop-Harrison JS. 2012. Genome evolution: extinction, continuation or explosion? *Current Opinion in Plant Biology,* **15**.

Heslop-Harrison JS, Brandes A, Taketa S *et al.* 1997. The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. *Genetica*, **100**: 197-204.

Heslop-Harrison JS, Schwarzacher T. 2007. Domestication, genomics and the future for banana. *Ann Bot*, **100**: 1073-84.

Heslop-Harrison JS, Schwarzacher T. 2011. Organisation of the plant genome in chromosomes. *Plant J*, **66**: 18-33.

Hippolyte I, Jenny C, Gardes L *et al.* 2012. Foundation characteristics of edible *Musa* triploids revealed from allelic distribution of SSR markers. *Ann Bot*, **109**: 937-51.

Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J. 2010. Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biol*, **10**: 204.

Hu W, Das OP, Messing J. 1995. Zeon-1, a member of a new maize retrotransposon family. *Mol Gen Genet*, **248**: 471-80.

Huang J, Zhang K, Shen Y *et al.* 2009. Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice. *Genomics*, **93**: 274-81.

Jacobs G, Dechyeva D, Menzel G, Dombrowski C, Schmidt T. 2004. Molecular characterization of Vulmar1, a complete mariner transposon of sugar beet and diversity of mariner- and En/Spm-like sequences in the genus *Beta*. *Genome*, **47**: 1192-201.

Jarvik T, Lark KG. 1998. Characterization of Soymar1, a mariner element in soybean. *Genetics*, **149**: 1569-74.

Jiang N, Bao Z, Temnykh S *et al.* 2002. Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics*, **161**: 1293-305.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004a. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**: 569-73.

Jiang N, Ferguson AA, Slotkin RK, Lisch D. 2011. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proc Natl Acad Sci U S A*, **108**: 1537-42.

Jiang N, Feschotte C, Zhang X, Wessler SR. 2004b. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr Opin Plant Biol*, **7**: 115-9.

Jing R, Vershinin A, Grzebyta J *et al.* 2010. The genetic diversity and evolution of field pea (*Pisum*) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. *BMC Evol Biol*, **10**: 44.

Junier T, Pagni M. 2000. Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**: 178-9.

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE. 2005. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res*, **15**: 1292-7.

Jurka J, Kapitonov VV. 2001. PIFs meet Tourists and Harbingers: a superfamily reunion. *Proc Natl Acad Sci U S A*, **98**: 12315-6.

Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet*, **8**: 241-59.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, **110**: 462-7.

Kalendar R, Schulman AH. 2006. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nat Protoc,* **1**: 2478-84.

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics,* **166**: 1437-50.

Kamate K, Brown S, Durand P, Bureau JM, De Nay D, Trinh TH. 2001. Nuclear DNA content and base composition in 28 taxa of *Musa*. *Genome,* **44**: 622-7.

Kapitonov VV, Jurka J. 1999. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica,* **107**: 27-37.

Kapitonov VV, Jurka J. 2003. A novel class of SINE elements derived from 5S rRNA. *Mol Biol Evol,* **20**: 694-702.

Kapitonov VV, Jurka J. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol,* **23**: 311-24.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet,* **9**: 411-2; author reply 414.

Kapitonov VV, Tempel S, Jurka J. 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene,* **448**: 207-13.

Karlov GI, Fesenko IA, Andreeva GN, Khrustaleva LI. 2010. [Chromosome organization of Ty1-copia-like retrotransposons in the tomato genome]. *Genetika,* **46**: 769-73.

Kawakami T, Strakosh SC, Zhen Y, Ungerer MC. 2010. Different scales of Ty1/copia-like retrotransposon proliferation in the genomes of three diploid hybrid sunflower species. *Heredity (Edinb),* **104**: 341-50.

Kazazian HH, Jr. 2004. Mobile elements: drivers of genome evolution. *Science,* **303**: 1626-32.

Kempken F, Windhofer F. 2001. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma,* **110**: 1-9.

Khan MF, Yadav BS, Ahmad K, Jaitly AK. 2011. Mapping and analysis of the LINE and SINE type of repetitive elements in rice. *Bioinformation,* **7**: 276-9.

Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution,* **55**: 1-24.

Kikuchi K, Terauchi K, Wada M, Hirano HY. 2003. The plant MITE mPing is mobilized in anther culture. *Nature,* **421**: 167-70.

Koch MA, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*). *Mol Biol Evol,* **17**: 1483-98.

Komatsu M, Shimamoto K, Kyozuka J. 2003. Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma. *Plant Cell,* **15**: 1934-44.

Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. *Heredity (Edinb),* **107**: 487-95.

Kuang H, Padmanabhan C, Li F *et al.* 2009. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the *Solanaceae*: new functional implications for MITEs. *Genome Res,* **19**: 42-56.

Kubis SE, Heslop-Harrison JS, Desel C, Schmidt T. 1998. The genomic organization of non-LTR retrotransposons (LINEs) from three *Beta* species and five other angiosperms. *Plant Mol Biol,* **36**: 821-31.

Kumar A, Bennetzen JL. 1999. Plant retrotransposons. *Annu Rev Genet,* **33**: 479-532.

Kunze R, Saedler H, Lonnig WE. 1997. Plant Transposable elements. *Advances in Botanical Research,* **27**.

Kwon SJ, Hong SW, Son JH *et al.* 2006. CACTA and MITE transposon distributions on a genetic map of rice using F15 RILs derived from Milyang 23 and Gihobyeo hybrids. *Mol Cells,* **21**: 360-6.

Langdon T, Jenkins G, Hasterok R, Jones RN, King IP. 2003. A high-copy-number CACTA family transposon in temperate grasses and cereals. *Genetics,* **163**: 1097-108.

Le QH, Wright S, Yu Z, Bureau T. 2000. Transposon diversity in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A,* **97**: 7376-81.

Le Rouzic A, Boutin TS, Capy P. 2007. Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A,* **104**: 19375-80.

Lee JK, Kwon SJ, Park KC, Kim NS. 2005. Isaac-CACTA transposons: new genetic markers in maize and sorghum. *Genome,* **48**: 455-60.

Leeton PR, Smyth DR. 1993. An abundant LINE-like element amplified in the genome of *Lilium speciosum*. *Mol Gen Genet,* **237**: 97-104.

Lenoir A, Lavie L, Prieto JL *et al.* 2001. The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. *Mol Biol Evol,* **18**: 2315-22.

Lerat E, Capy P. 1999. Retrotransposons and retroviruses: analysis of the envelope gene. *Mol Biol Evol,* **16**: 1198-207.

Lisch D. 2002. Mutator transposons. *Trends Plant Sci,* **7**: 498-504.

Lisch DR, Freeling M, Langham RJ, Choy MY. 2001. Mutator transposase is widespread in the grasses. *Plant Physiol,* **125**: 1293-303.

Llorens C, Futami R, Bezemer D, Moya A. 2008. The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Res,* **36**: D38-46.

Llorens C, Futami R, Covelli L *et al.* 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res,* **39**: D70-4.

Llorens C, Munoz-Pomer A, Bernad L, Botella H, Moya A. 2009. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol Direct,* **4**: 41.

Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H. 2012. Miniature Inverted-Repeat Transposable Elements (MITEs) Have Been Accumulated through Amplification Bursts and Play Important Roles in Gene Expression and Species Diversity in *Oryza sativa*. *Mol Biol Evol,* **29**: 1005-17.

Lyons M, Cardle L, Rostoks N, Waugh R, Flavell AJ. 2008. Isolation, analysis and marker utility of novel miniature inverted repeat transposable elements from the barley genome. *Mol Genet Genomics,* **280**: 275-85.

Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res,* **14**: 860-9.

Ma XF, Gustafson JP. 2008. Allopolyploidization-accommodated genomic sequence changes in triticale. *Ann Bot,* **101**: 825-32.

Macas J, Koblizkova A, Neumann P. 2005. Characterization of Stowaway MITEs in pea (*Pisum sativum* L.) and identification of their potential master elements. *Genome,* **48**: 831-9.

Macas J, Neumann P, Navratilova A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics,* **8**: 427.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol,* **16**: 793-805.

Marsch-Martinez N, Pereira A. 2011. Activation tagging with En/Spm-I /dSpm transposons in *Arabidopsis*. *Methods Mol Biol,* **678**: 91-105.

McCarty DR, Settles AM, Suzuki M *et al.* 2005. Steady-state transposon mutagenesis in inbred maize. *Plant J,* **44**: 52-61.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A,* **36**: 344-55.

Menzel G, Dechyeva D, Keller H, Lange C, Himmelbauer H, Schmidt T. 2006. Mobilization and evolutionary history of miniature inverted-repeat transposable elements (MITEs) in *Beta vulgaris* L. *Chromosome Res,* **14**: 831-44.

Menzel G, Krebs C, Diez M *et al.* 2012. Survey of sugar beet (*Beta vulgaris* L.) hAT transposons and MITE-like hATpin derivatives. *Plant Mol Biol,* **78**: 393-405.

Mills RE, Bennett EA, Iskow RC *et al.* 2006. Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet,* **78**: 671-9.

Minervini CF, Viggiano L, Caizzi R, Marsano RM. 2009. Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. *Gene,* **440**: 42-9.

Miura A, Kato M, Watanabe K, Kawabe A, Kotani H, Kakutani T. 2004. Genomic localization of endogenous mobile CACTA family transposons in natural variants of *Arabidopsis thaliana*. *Mol Genet Genomics,* **270**: 524-32.

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature,* **411**: 212-4.

Moisy C, Garrison KE, Meredith CP, Pelsy F. 2008. Characterization of ten novel Ty1/copia-like retrotransposon families of the grapevine genome. *BMC Genomics,* **9**: 469.

Momose M, Abe Y, Ozeki Y. 2010. Miniature inverted-repeat transposable elements of Stowaway are active in potato. *Genetics,* **186**: 59-66.

Monteiro A, Lunn T. 1999. Trends and perspectives of perspectives of vegetable *Brassica* Breeding World-wide. WCHR- World Conference on Horticulture Research ISHS. *Acta Horticulture,* **495**.

Motohashi R, Ohtsubo E, Ohtsubo H. 1996. Identification of Tnr3, a suppressor-mutator/enhancer-like transposable element from rice. *Mol Gen Genet,* **250**: 148-52.

Muehlbauer GJ, Bhau BS, Syed NH *et al.* 2006. A hAT superfamily transposase recruited by the cereal grass genome. *Mol Genet Genomics,* **275**: 553-63.

Muszewska A, Hoffman-Sommer M, Grynberg M. 2011. LTR retrotransposons in fungi. *PLoS One,* **6**: e29425.

Muthukumar B, Bennetzen JL. 2004. Isolation and characterization of genomic and transcribed retrotransposon sequences from sorghum. *Mol Genet Genomics,* **271**: 308-16.

Myouga F, Tsuchimoto S, Noma K, Ohtsubo H, Ohtsubo E. 2001. Identification and structural analysis of SINE elements in the *Arabidopsis thaliana* genome. *Genes Genet Syst,* **76**: 169-79.

Nacken WK, Piotrowiak R, Saedler H, Sommer H. 1991. The transposable element Tam1 from *Antirrhinum majus* shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol Gen Genet,* **228**: 201-8.

Noma K, Ohtsubo E, Ohtsubo H. 1999. Non-LTR retrotransposons (LINEs) as ubiquitous components of plant genomes. *Mol Gen Genet,* **261**: 71-9.

Noma K, Ohtsubo H, Ohtsubo E. 2001. A new class of LINEs (ATLN-L) from *Arabidopsis thaliana* with extraordinary structural features. *DNA Res,* **8**: 291-9.

Novikova O. 2009. Chromodomains and LTR retrotransposons in plants. *Commun Integr Biol,* **2**: 158-62.

Novikova O, Mayorov V, Smyshlyaev G *et al.* 2008. Novel clades of chromodomain-containing Gypsy LTR retrotransposons from mosses (Bryophyta). *Plant J,* **56**: 562-74.

Okada N, Hamada M, Ogiwara I, Ohshima K. 1997. SINEs and LINEs share common 3' sequences: a review. *Gene,* **205**: 229-43.

Oki N, Yano K, Okumoto Y, Tsukiyama T, Teraishi M, Tanisaka T. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. japonica. *Genes Genet Syst,* **83**: 321-9.

Okpul T, Harding RM, Dieters MJ, Godwin ID. 2011. Occurrence of LINE, gypsy-like, and copia-like retrotransposons in the clonally propagated sweet potato (*Ipomoea batatas* L.). *Genome,* **54**: 603-9.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature,* **284**: 604-7.

Osborn TC, Pires JC, Birchler JA *et al.* 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet,* **19**: 141-7.

Ostergaard L, King GJ. 2008. Standardized gene nomenclature for the *Brassica* genus. *Plant Methods,* **4**: 10.

Ouyang S, Buell CR. 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res,* **32**: D360-3.

Park KC, Jeong CS, Song MT, Kim NS. 2003. A new MITE family, Pangrangja, in *Gramineae* species. *Mol Cells,* **15**: 373-80.

Park M, Jo S, Kwon JK *et al.* 2011. Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. *BMC Genomics,* **12**: 85.

Pearce SR, Harrison G, Heslop-Harrison PJ, Flavell AJ, Kumar A. 1997. Characterization and genomic organization of Ty1-copia group retrotransposons in rye (*Secale cereale*). *Genome,* **40**: 617-25.

Pereira A, Cuypers H, Gierl A, Schwarz-Sommer Z, Saedler H. 1986. Molecular analysis of the En/Spm transposable element system of *Zea mays*. *Embo J,* **5**: 835-41.

Perrier X, De Langhe E, Donohue M *et al.* 2011. Multidisciplinary perspectives on banana (*Musa* spp.) domestication. *Proc Natl Acad Sci U S A,* **108**: 11311-8.

Peters SA, Datema E, Szinay D *et al.* 2009. *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *Plant J,* **58**: 857-69.

Plasterk RHA, van Luenen H. 1997. Transposons. In: Riddle DL, Blumenthal T, Meyer BJ, Priess JR eds. *C. elegans II.* Cold Spring Harbor NY.

Pollefeys P, Sharrock S, Arnaud E. 2004. Preliminary analysis of the literature on the distribution of wild *Musa* species using MGIS and DIVA-GIS Montpellier, France: INIBAP.

Pozueta-Romero J, Houlne G, Schantz R. 1998. Identification of a short interspersed repetitive element in partially spliced transcripts of the bell pepper (*Capsicum annuum*) PAP gene: new evolutionary and regulatory aspects on plant tRNA-related SINEs. *Gene,* **214**: 51-8.

Price Z, Dumortier F, MacDonald W, Mayes S. 2002. Characterisation of copia-like retrotransposons in oil palm (*Elaeis guineensis* Jacq.). *Theor Appl Genet,* **104**: 860-867.

Proels RK, Roitsch T. 2006. Cloning of a CACTA transposon-like insertion in intron I of tomato invertase Lin5 gene and identification of transposase-like sequences of *Solanaceae* species. *J Plant Physiol,* **163**: 562-9.

Queen RA, Gribbon BM, James C, Jack P, Flavell AJ. 2004. Retrotransposon-based molecular markers for linkage and genetic diversity analysis in wheat. *Mol Genet Genomics,* **271**: 91-7.

Rajput MK, Upadhyaya KC. 2009. CARE1, a TY3-gypsy like LTR-retrotransposon in the food legume chickpea (*Cicer arietinum* L.). *Genetica,* **136**: 429-37.

Ramallo E, Kalendar R, Schulman AH, Martinez-Izquierdo JA. 2008. Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. *Plant Mol Biol,* **66**: 137-50.

Remigereau MS, Robin O, Siljak-Yakovlev S, Sarr A, Robert T, Langin T. 2006. Tuareg, a novel miniature-inverted repeat family of pearl millet (*Pennisetum glaucum*) related to the PIF superfamily of maize. *Genetica,* **128**: 205-16.

Roccaro M, Li Y, Sommer H, Saedler H. 2007. ROSINA (RSI) is part of a CACTA transposable element, TamRSI, and links flower development to transposon activity. *Mol Genet Genomics,* **278**: 243-54.

Rocheta M, Cordeiro J, Oliveira M, Miguel C. 2007. PpRT1: the first complete gypsy-like retrotransposon isolated in Pinus pinaster. *Planta,* **225**: 551-62.

Rubin E, Lithwick G, Levy AA. 2001. Structure and evolution of the hAT transposon superfamily. *Genetics,* **158**: 949-57.

Salina EA, Sergeeva EM, Adonina IG *et al.* 2011. The impact of Ty3-gypsy group LTR retrotransposons Fatima on B-genome specificity of polyploid wheats. *BMC Plant Biol,* **11**: 99.

SanMiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* , **82**.

Santiago N, Herraiz C, Goni JR, Messeguer X, Casacuberta JM. 2002. Genome-wide analysis of the Emigrant family of MITEs of *Arabidopsis thaliana*. *Mol Biol Evol,* **19**: 2285-93.

Sanz AM, Gonzalez SG, Syed NH, Suso MJ, Saldana CC, Flavell AJ. 2007. Genetic diversity analysis in *Vicia* species using retrotransposon-based SSAP markers. *Mol Genet Genomics,* **278**: 433-41.

Sarilar V, Marmagne A, Brabant P, Joets J, Alix K. 2011. BraSto, a Stowaway MITE from *Brassica*: recently active copies preferentially accumulate in the gene space. *Plant Mol Biol,* **77**: 59-75.

Schmidt T. 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant Mol Biol,* **40**: 903-10.

Schwarzacher T, Heslop-Harrison JS. 2000. Practical in situ hybridization. Oxford: Bios.

Schmidt T, Kubis S, Heslop-Harrison JS. 1995. Analysis and chromosomal localization of retrotransposons in sugar beet (*Beta vulgaris* L.): LINEs and Ty1-copia-like elements as major components of the genome. *Chromosome Res,* **3**: 335-45.

Schranz ME, Osborn TC. 2000. Novel flowering time variation in the resynthesized polyploid *Brassica napus*. *J Hered,* **91**: 242-6.

Schulman AH, Flavell AJ, Ellis TH. 2004. The application of LTR retrotransposons as molecular markers in plants. *Methods Mol Biol,* **260**: 145-73.

Schulman AH, Flavell AJ, Paux E, Ellis TH. 2012. The application of LTR retrotransposons as molecular markers in plants. *Methods Mol Biol,* **859**: 115-53.

Seibt KM, Wenke T, Wollrab C *et al.* 2012. Development and application of SINE-based markers for genotyping of potato varieties. *Theor Appl Genet*.

Sergeeva EM, Salina EA, Adonina IG, Chalhoub B. 2010. Evolutionary analysis of the CACTA DNA-transposon Caspar across wheat species using sequence comparison and in situ hybridization. *Mol Genet Genomics,* **284**: 11-23.

Shimatani Z, Takagi K, Eun CH *et al.* 2009. Characterization of autonomous Dart1 transposons belonging to the hAT superfamily in rice. *Mol Genet Genomics,* **281**: 329-44.

Shirasawa K, Hirakawa H, Tabata S *et al.* 2012. Characterization of active miniature inverted-repeat transposable elements in the peanut genome. *Theor Appl Genet*.

Shirsat AH. 1988. A transposon-like structure in the 5' flanking sequence of a legumin gene from *Pisum sativum*. *Mol Gen Genet,* **212**: 129-33.

Shu Y, Li Y, Bai X *et al.* 2011. Identification and characterization of a new member of the SINE Au retroposon family (GmAu1) in the soybean, Glycine max (L.) Merr., genome and its potential application. *Plant Cell Rep,* **30**: 2207-13.

Snowden KC, Napoli CA. 1998. Psl: a novel Spm-like transposable element from *Petunia hybrida*. *Plant J,* **14**: 43-54.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene,* **167**: GC1-10.

Staton SE, Ungerer MC, Moore RC. 2009. The genomic organization of Ty3/gypsy-like retrotransposons in Helianthus (*Asteraceae*) homoploid hybrid species. *Am J Bot,* **96**: 1646-55.

Syed NH, Flavell AJ. 2006. Sequence-specific amplification polymorphisms (SSAPs): a multi-locus approach for analyzing transposon insertions. *Nat Protoc,* **1**: 2746-52.

Takahashi H, Akagi H, Mori K, Sato K, Takeda K. 2006. Genomic distribution of MITEs in barley determined by MITE-AFLP mapping. *Genome,* **49**: 1616-20.

Tam SM, Causse M, Garchery C, Burck H, Mhiri C, Grandbastien MA. 2007. The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species. *J Evol Biol,* **20**: 1056-72.

Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A. 2000. The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell,* **12**: 381-91.

Tian PF. 2006. Progress in plant CACTA elements. *Yi Chuan Xue Bao,* **33**: 765-74.

Tomita M, Asao M, Kuraki A. 2010. Effective isolation of retrotransposons and repetitive DNA families from the wheat genome. *J Integr Plant Biol,* **52**: 679-91.

Tomlinson P. 1969. Anatomy of the monocotyledons. III. Commelinales–Zingiberales *Oxford: Clarendon Press; 1969*.

Town CD, Cheung F, Maiti R *et al.* 2006. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell,* **18**: 1348-59.

Trentmann SM, Saedler H, Gierl A. 1993. The transposable element En/Spm-encoded TNPA protein contains a DNA binding and a dimerization domain. *Mol Gen Genet,* **238**: 201-8.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. 2009. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature,* **461**: 423-6.

Tu Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci U S A,* **98**: 1699-704.

U.N. 1935. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn. J. Bot. 1935,* **7**: 389-452.

Ungerer MC, Strakosh SC, Stimpson KM. 2009. Proliferation of Ty3/gypsy-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. *BMC Biol,* **7**: 40.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet,* **10**: 725-32.

van Leeuwen H, Monfort A, Puigdomenech P. 2007. Mutator-like elements identified in melon, *Arabidopsis* and rice contain ULP1 protease domains. *Mol Genet Genomics,* **277**: 357-64.

Vershinin AV, Druka A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS. 2002. LINEs and gypsy-like retrotransposons in Hordeum species. *Plant Mol Biol,* **49**: 1-14.

Vukich M, Giordani T, Natali L, Cavallini A. 2009. Copia and Gypsy retrotransposons activity in sunflower (*Helianthus annuus* L.). *BMC Plant Biol,* **9**: 150.

Wang GD, Tian PF, Cheng ZK *et al.* 2003. Genomic characterization of Rim2/Hipa elements reveals a CACTA-like transposon superfamily with unique features in the rice genome. *Mol Genet Genomics,* **270**: 234-42.

Wang H, Liu JS. 2008. LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. *BMC Genomics,* **9**: 382.

Wang X, Wang H, Wang J *et al.* 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet,* **43**: 1035-9.

Waugh R, McLean K, Flavell AJ *et al.* 1997. Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol Gen Genet,* **253**: 687-94.

Wessler SR, Bureau TE, White SE. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr Opin Genet Dev,* **5**: 814-21.

Wicker T, Guyot R, Yahiaoui N, Keller B. 2003. CACTA transposons in *Triticeae*. A diverse family of high-copy repetitive elements. *Plant Physiol,* **132**: 52-63.

Wicker T, Keller B. 2007. Genome-wide comparative analysis of copia retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res,* **17**: 1072-81.

Wicker T, Sabot F, Hua-Van A *et al.* 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet,* **8**: 973-82.

Witte CP, Le QH, Bureau T, Kumar A. 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc Natl Acad Sci U S A,* **98**: 13778-83.

Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *Embo J,* **9**: 3353-62.

Xu M, Brar HK, Grosic S, Palmer RG, Bhattacharyya MK. 2010. Excision of an active CACTA-like transposable element from DFR2 causes variegated flowers in soybean [*Glycine max* (L.) Merr.]. *Genetics,* **184**: 53-63.

Xu Z, Dooner HK. 2005. Mx-rMx, a family of interacting transposons in the growing hAT superfamily of maize. *Plant Cell,* **17**: 375-88.

Yaakov B, Ceylan E, Domb K, Kashkush K. 2012. Marker utility of miniature inverted-repeat transposable elements for wheat biodiversity and evolution. *Theor Appl Genet*.

Yamashita H, Tahara M. 2006. A LINE-type retrotransposon active in meristem stem cells causes heritable transpositions in the sweet potato genome. *Plant Mol Biol,* **61**: 79-94.

Yang G, Hall TC. 2003. MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res,* **31**: 3659-65.

Yang G, Weil CF, Wessler SR. 2006. A rice Tc1/mariner-like element transposes in yeast. *Plant Cell,* **18**: 2469-78.

Yang TJ, Kim JS, Lim KB *et al.* 2005. The Korea *Brassica* genome project: a glimpse of the *Brassica* genome based on comparative genome analysis with *Arabidopsis*. *Comp Funct Genomics,* **6**:138-46.

Yang TJ, Kwon SJ, Choi BS *et al.* 2007. Characterization of terminal-repeat retrotransposon in miniature (TRIM) in *Brassica* relatives. *Theor Appl Genet,* **114**: 627-36.

Yang YW, Lai KN, Tai PY, Li WH. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J Mol Evol,* **48**: 597-604.

Yano ST, Panbehi B, Das A, Laten HM. 2005. Diaspora, a large family of Ty3-gypsy retrotransposons in *Glycine max*, is an envelope-less member of an endogenous plant retrovirus. *BMC Evol Biol,* **5**: 30.

Zabala G, Vodkin L. 2008. A putative autonomous 20.5 kb-CACTA transposon insertion in an F3'H allele identifies a new CACTA transposon subfamily in *Glycine max*. *BMC Plant Biol,* **8**: 124.

Zerjal T, Rousselet A, Mhiri C *et al.* 2012. Maize genetic diversity and association mapping using transposable element insertion polymorphisms. *Theor Appl Genet*.

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR. 2001. P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc Natl Acad Sci U S A,* **98**: 12572-7.

Zhang X, Jiang N, Feschotte C, Wessler SR. 2004. PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics,* **166**: 971-86.

Zhang X, Wessler SR. 2005. BoS: a large and diverse family of short interspersed elements (SINEs) in *Brassica oleracea*. *J Mol Evol,* **60**: 677-87.

Ziolkowski PA, Kaczmarek M, Babula D, Sadowski J. 2006. Genome evolution in *Arabidopsis/Brassica*: conservation &divergence of ancient rearranged segments and their breakpoints. *Plant J,* **47**: 63-74.

Zou J, Fu D, Gong H *et al.* 2011. De novo genetic variation associated with retrotransposon activation, genomic rearrangements and trait variation in a recombinant inbred line population of *Brassica* napus derived from interspecific hybridization with *Brassica rapa*. *Plant J,* **68**: 212-24.

IMPORTANT WEBSITES USED

***Arabidopsis* Genomic tRNA Database:** http://gtrnadb.ucsc.edu/Athal

Conserved Domain Database: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

European Bioinformatics Institute (EBI): http://www.ebi.ac.uk

GC Calculator: http://www.genomicsplace.com/gc_calc.html

Gypsy database: http://gydb.org/index.php/Main_Page

National Center for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/

ORF finder: http://www.ncbi.nlm.nih.gov/projects/gorf/

Primer3: http://frodo.wi.mit.edu/

Repeat masker of Censor software: http://www.girinst.org/censor/index.php

The Repbase database: http://www.girinst.org/repbase/index.html

WebLogo: http://weblogo.berkeley.edu/logo.cgi

CHAPTER 3
IDENTIFICATION AND CHARACTERIZATION OF NOVEL LTR RETROTRANSPOSON FAMILIES FROM *BRASSICA*

The nucleotide sequences of LTR retrotransposons identified in *Brassica* genomes are given below. Only few representatives from each superfamily are given below. The TSDs are shown in red while LTRs are represented by blue colour. The red colour in internal regions indicates the reverse transcriptase region. The details of the elements are given in table 3.1.

***BrCOP1*: Copia retrotransposon in *Brassica rapa* (AC189222.1) from 120554-30970 (3′-5′)**

GTGAA***TGTTGAAAAAAGGATCGTTTTATACTTTGTTTCGTTCTCTTTGATTTCGGTATCTCTCTCACCTCTCACGTTGTTT***
CATTCGTGTGTATTACGTTTGTTTGCGGAAAGTAAAGACACACAGATTTAACCAGTTCACGCCTCAGTGTGAGGACGTTAC
GCCTGGTCCGAGGTTATCTCGGAAATCCACTAGAAAGCTTGGTTACACTTAACCCTTAGAGACCGAACAAACGCTAGACTT
GCGGCTGCTCTCTGTGTTTCTCAACTGCTCTCTTAACCCTAATGTCTCAGCTCTGTGTCCGCTTTATAGCATTTGAAGAGG
CGGTGCCACGCTCCTCTCTCTTCTCTGTTTTACTCTGCTTTACCGTACAGAGGAAGAAGACAAATGCCAGCGATAAGGGAG
ACATGGTTAATTACTCCGTTGCCCCTTAATACCGTTAAGTCTCGCACGTGAGCTCTCCGTGCGAACTACTGTTTCCTTAAC
GACCAGCTGAGACCCTCTGCTTCGGGTAGAGACTCTTCGACACTTAGGCAAAACTCAACAAAACTCCACCTTGCCGTAGTG
TCAACTTCCATCTTTCTCTTTAGCTTACTTTCTAATGTTACTTGATTTCCTCCTCGGATCCTGATACTCTTATCAGTTACT
GGTAATCCTGAGCTTCTCCATTGCGCTCAGGAAATTGTTCACCGGCAAGACTTTAGTAAGCATGTCTGCCGGATTGTAGAC
TGTTGATATCTTCAACACCTTCGTTTCACCATCAGTAATTGTGTCTCTGATGAAGTGAAACCTTCTCTGTATATGCTTTGT
CCTTTCATGGTGGACAGCATTCTTTGCTAATGCTATAGCACTTTGAGAATCGCAGTGAATCTCCACTGCTCCTTGCTTAAA
CCCCAACTCGTTCATCAAACCCTTTAGCCAAATGGCTTCTTTTGCTGCTTCTGTTAGGGAGATGTACTCTGCCTCTGTTGT
AGATAATGCCACCACAGGCTGTAAACCTGATCTCCAACTTATAACATTACCACCAACTGTAAACGCATATCCTGTAATGGA
CCTTCGCTTATCAAGGTCGGCTGAGTAATCTGAATCGCAATATCCTGTTACTACAAACTCTCCTTCTCCCTTGAACTTTAG
TCTTGTGTCATGTGATCCTTGTATGTACCTTAGTACCCACTTAACTGCTTGCCAGTGAACCATCAGTGGCTTTCCCATGAA
CCTGCTTATCATTCCAACTGGATAAGCGAGGTCTGGTCTCGATCCTATCATCGAATACATGATGCTACCAACTGAATTTGC
GTAAGGAATAGACTTCATCTGATCAGCTTCCTCTTCCAGTTCTTTATCAGTTGTAGATTTGAGTCTCATGTGTGCCCCCAA
GGGCGTTTTCACCGGTTTACAGTTCACCA**TTCTGAAGGTTCTCAGTACTTTCTGAAGATACTCTTTCTACGACAGTTCCAC**
AACTCGTTCTTCTCTGTCTCTCTTTATCTCCATTCCCAATATCCTCTTGGCTGGACCAAGGTCTTTCATTTCAAAAGTAGC
ACTTAGACTTTCCTTCAACTCCAATACTGTGTCTTTGTTCTTTGATATAATCAACATATCGTCCACGTACAGCAGAAGGTA
AGTCCTCTGCTCTTCCCTGGTCTTCTTAAAGTACAAGCAACTGTCCTTCAGACTTCTCGAATATCCTGTAGACCTCATGAA
AGCATCGAATCTTTGGTTCCACTGTCTTGGTGATTGTCTCAAGCCGTATAACGATTTTCTTAACAGACAAACCTTCTCAGG
TGCGTTCTTGTCCACATATCCTTCAGGCTGATCCATGAATATCTCTTCGTCCAAGGTTCCGTGCAAGAATGTTGTCTTAAC
GTCCATTTGCTCTAGTTCTAGATTGAAATGAGCAACTGCAGACAGCATTAGACGGATCGTGACGTGTTTGACTACGGGTGA
ACAGATTTCTTGAAAATCAATACCTTCCCTCTGAGAGTAACCTTTGGCAACAAGTCTAGATTTGTGTCTTGGTCTCTCAAC
ACCCGGGATTCCAACTTTCCTCTTGAACACCCACTTACAGCCTATTACTTTCTTCTTCACTGGTTTCTTAACTAGATCCCA
TGTATGATTCTTGATAAGTGATAACATCTCTTCATCTGTTGATCCTCTCCAGAGTTTGCTCTCTGGACTGAGCATTGCTTC
AGCATAACTTTGCGGCTCAAGATCTCCACCATCCTCAGTTAAGTTAAATGCAAAACCAAGGTCTCCTAGTTCATCATATCT
CTTAGGAGGTCTTGTAGTTCTTCTTTCTCTATCTCTCGCTAACTGGTAGTCAGCAAGACTTGGTGGCATATAGTCCTCCTC
TGAATCTGTCTCAGACTCTGATCCTTGTTGCTGTTCTGTAGGCACGTTTCCTTGGTTAATGTTGTCTTGCTCGTTAGCTCC
ACCTTCTTCTGTTACTTCATTTTCTGTAGTTACAGTAGGAGTACTTATTACAGTAACCTCAGGTTCAGTTTCTGTTCTTTC
TTCAGTCTGTCCAGTAGAGCTCTTGAACATAACTTCCTCATTAAACACAACGTCTTTACTTATTACCACTTTCTTTTCTTC
AATTAACCACACCCTGTAGCCTTTTACCCCTGTTGGATAACCAAGGAACACTCCTTTCTTGGCTCTTGGATTTAACTTCCC
TTGATCTGAATGTACATAGGAGATACATCCAAAACTTCTCAGACCACTCAAGTCAGGTAATACCTCTGTCCACACTTTCTC
TGGTATCTCATAATCGATCGGAGAAGCTAGTGTTCTGTTGATTACATACACAGCTGTTGAAGCCGCTTCTGCCCAGAACAT
CTTATCCAAACCAGACTCGCTGAGCATGCTCCTAACCTTCTCCATTATCGTTATGTTCATCCGTTCAGCAACGCCGTTCTG
TTGTGGAGTGTAAGCACACGTTCTGTGTCTTACAACTCCCTCTTGCTTGCAGTAACCATCGAACTCAAGATTACAAAACTC
AAGCCCATTGTCAGTCCTCAACTTCTTAACCTTCCTGCTTGATTGATTCTCAACCATCGTTTTCCATTCGACGAACGTTCT
AAAGGCTTCATCCTTCTTCTTCAGAAACTATATCCAGACTTTTCTGCTGTAGTCGTCGATGAATGTCATGAAGTACTGACA
GTTTCTCAGAGAACTTGGTACTGACGGTTCTCCCCAAAGATCTGAGTGTATGTATTCAAGTTTTTCTTTTGTAACGTGCTT
CCCTTGACCGAAGCTGACCTTATGAGATTTACCATACACACAGTCTTCGCAGAATTCAAAACCTTTCATCTTATCTGCTTC
TAGGCATCCTTTCTTGATCAATAAATCTGTGTTCTTCTTACTCATGTGACATAGTCTGCTATGCCACAGCTTTGACTCGTT
CTTGGAACTCACTACTGCATTTGCGCTTCCTGCAACTACCTTTCCTTGTAAGATGTATAGAGTTCCTACTTTCTCCCCTTT
CAACAGCGTTAAGCAACCTTTGGTGACTTTCAAAGAACCGTTCCTGGACTGAAACCAACAGCCTTGATCTTCTAATGTTCC
CATTGAGATCAAGTTTCTCGTCATGCTTGGAACGTATCTGACGTTTGAAATCAGGACAGTTGATTGATCATCATTCTGAAT
CCGTATACTGCCAATGCCTTTGATCTCAGAATGAGTGTGATTAGCCATCTTCACTCTTCCTGTTTTTGAGTTATCAAACTC
AACAAACCAATCCCTCCTTGGAGTCATATGAAATGAGCACCCTGTGTCCATGATCCACTCGTCTTCTTGTCCTTCACATAC
TGCATTAGACTCTTCTTCCACATTCAAACCAGCTGCATCGACAACTTGACCCACAACATTCGAAGTTTCACCTTTCTCTGA
ACTGCCGTTCTTCTTGTTCCTTTCTTTCCAAACAAAACACTGCTTCTTGTAGTGACCGTCTTTACCACAAATCCAACAAGT
CATCTTCTCTCTTGACTGTGACCTGTTCTGGCTTTTTCCTCTGTCTGAACTCCTTCCTCTTCTATCTCTCCTGCCTCTTTC
TTGAACATATAGAGCCTCTGAGTTACTCTTGCTCAATTTACCACTCGCTCCAAGTTCTAGACTTTTCGACCTTATAGCTCC
AGTGATTACTTCCAGTGCTAGCGTTGTCTTACCATACTTCAAAGTTTCTTTCAACTGATCAAACTGTTTCGGTAAAAACAT
CAGTAACACAATCGCTTGGTCTTCATCTGAAACTGTAACCTTCACGTTCTCCAGGTCAGATATAAGCTTGAAGAAGTCGTT
TACATTCTCCTCCATAGTCATGCTATCTGACATCTTGTACCCATATAACCTTTGCTTAAGGTATATCCTGTTGGGTAACGA
CTTAGCCATGAACAACTTGTCTAAGATCTTGAGCATACCAGCTGTGGTCGTTTCCTTGATGACCTTTCTGAGAATATGATC
CCCCAGACTTAGTATGATCGTAGATCTAGCTTTTTCTCTCTTCTCCTTGAGGGTTTTCTCCTTTTGAATCTTCTCTTCCGG
CTTGTCAGAAGCTTCAGTCGCAGACATCCCTTTCTCTGCTGTAGAGTCTTCAACATCCACGCTTTCTTCATCTTCTAACCC

TTCTAACAAACACAGTAGCTCCAGATGAGCCAACAGCTTCTCTTTCCAGAGAACATAATCGCCTTCTCCGTCGAACTTCTC
AACCTCTGAACGTACTGAAGTCATGATACGCATCAAAGGACAAAGCTTTACAGAAAAGCTTCAATCTTTACAAAGCTTCAA
TCTTTCCTCCTTCCCAGCCAGCCAACCTGGCTCTGATACCAAT***TATTGAGAAAAGGATCGTTTTATACTTTGTTTTGTTCT***
CTTTGATTTCGGTATCTCTCTCACCTCTCACGTTGTTTCACTCGTGTGTATTACGTTTGTTTGCGGAAAGTAAAGACACAC
AGATTTAACCAGTTCACGCCTCAGTGTGAGGACGTTACGTCTGGTCCGAGGTTATCTCGGAAATCCACTAGAAAGCTTGGT
TACACTTAATCCTTAGAGACCGAACAAACGCTAGACTTGCGGCTGCTCTCTGTGTTTCTCAACTGCTCTCTTAACCTTAAT
GTCTCAGCTCTGTGTCCGCTTTATAGCATTTGAAGAGGCGGTGCCACGCTCCTCTCTCTCCTCTATTTTACTCTGCTTTAC
CGTACAGAGGAAGAAGACAAATGCCAGCGATAAGGGAGGCATGGTTAATTACTCTGTTGCCCCTTAATACCGTTAAGTCTC
GCACGTGAGCTCTCTGTGCGAACTACTGTTTCCTTAACGACCAGCTGAGACCCTCTGCTTCGGGTAGAGACTCTTCGACAC
AGGCAAAACTCAATA**GTGAA**

***BoCOP22*: Copia retrotransposon in *Brassica oleracea* (AC149635.1) from 23364- 32285 bp (3′-5′)**

TAGCT*TGTTGATGTAAGATAAACGTACACTAACTAGAACACACAGTAACACAGAGATATTTCTTTTCAATAAATTCAACTA*
GATTAATATTTTTAAAAACACATACAAGTTCTTCTTAAAAGCAATCATTCAAACATGTTCAAGACAAACCTTGACTTATTG
ACTAGTCTTCCTCGTTCACAGACTTAACGTAGTCTATGAATCACTCAAACCCTTAACCCCTAGTGGCTTTGTTTCTCCTTA
TTCGGACTTGCAAACCTCCTATTGCTAAAAGCTCTCTCTGTGTGTTCAGCAGTCACACCGTGCCAAGCCTCCTCTTTATAT
TCTTCAGAGGAGAAACCCTAGACATAAAATCTATTTAATGGAAACTTTTCATATCTTCTTCTTCAGCGAATAAGCCAATCT
TCCTTTCTTCTGAGTTAGATTGCTTCTTCTCCAATTCTTCCTTTAAACTGACTCCTTCAGTAATTCTTCCTTGAATGCTCA
TTAACTCACCCGACCAGCTTCTTGTATCCTCATCTGATGTCGTAGTTCATGTCGTGCTTCATGAACTACTACAGCTCCGTC
TTGACTATACATCTTCAATCTCCCCCTTTTTAATCCTTGCGTCTCACAAGTACCTAAAATGAAGTCACACACACAACACAG
GGATTCACAACAACCACGAGCAAACACATATAGGCATGAAGCAATCAAAAGCATACTTATATTATCTTAATAGTTGGGTTC
AAGTATTTAATACATCTTAACAAAGACAACACACATAAATACTCCCCCATAAACCATGATCAGTCTGAATAATAAGATAAA
CACAATTGCAGCAAGTAGTGCAATCGGAAAGTACATAAAGCAAATAAACATCAAAGATTTCTCCCCCACAAACTTAGCCAT
TCTTCATCTTAATCCTCATCTTCTCCCCCTGCTTGTGAGCACAAGGGTTAAAAACGAAAGTAAACACAAGAGTCACAAAAG
AATAAGACGTGACTAACCTGCAACAGAACGACTCAGATCCTGAACAATGTCAGTCAGTGCTTGAAGAGCTCGGGTGATTTG
CTGTAAGACCTTGTGAAGATCATCCTGAGTCAGTACTCCTTGAGGTATTGAAACAAAGCCAAGGTCATGCACACTGAAACG
TCGCGGTCCTGTGGTGGTGCTTGGAGTCGGAGCACTTGGGGATGATGGCAGAGGGTGTGACACGCGAGCTGATCTTGAGGA
AAGTTTTCCTTCCCCTCCCTTCTTTGCGTTACCCTTTGCTTTAGCAAGACGCTGTTCATAGATCTCACCAACCCGTTTGTC
CTTTGTATAGACCACTGGACCCTGAGATTTCTCACCCGGAAGAAGAGAGACACGATGTTGCTTCTGGAGGACACTCATGAT
CAAGCGAGGAAAGATCAACCAACGAGAGTCATTCTTCTGCACCACACCAAGATTCAGGACTTGCCTGAAGATCAATCTTCC
CATATCAACTCGGATTCCTCGAGCCATCTTGTAGATCAACCGTGCTCGATCAACAGAGACATGTGTCTTGTGTGTAGACGG
GATCCAATTGTAGGCAGCAATGATCATGAGGGCACCGTAGCAAGAAGACAAGTCAGCGGTGGTCAGATTGTCCCATTCTTT
TCGAGTTCCTTCAGTGAGAAACTCAGCCAACTCAGACTTCGAGATGCTATCAAGAGTGGTCTCTTCTTCTACCTCATCTTC
ATCCAGAGGTTCGACATGGAGAGCCTCGTTGATCATTGTTGGTGAGAACTCATACGTGTGACCACGTACTTGGACCGCTAT
CTCCTCTGCATCTGCTTCAACCTTTGTGGCAGGAAGACCTGCATAGAACTCATCAACGACTTGTTCAACATAGTTTCCAAG
TGCTGAGACTGTGGTTCCCATAGGCCCTTTTGCGATGATCTCTATATAGCCCCACTGATCCTTCTCAGTCATATCCACAGA
TCTCTCGGCTATGAGGCTCCGACGCATAAATTGCTGAACCCTCTCGGTGATTGGGGACACTTCAACCACATCACTTCGCTC
AGATGCTTTCCTGGATCGAACACCGCTTGAAACAGAGGTATTCTGAGCAAGGACTGGAGGACTCAACGCATCAGTCGATGG
TGTCTTAGATGTGGACTTAGGGGCGCCAGATTTACTCGTCTTGTGTCGCTTTGAAGATCCTAATCCCAATTTCTTCATCAC
CCGCTTCTTCCTGCTTCGTTTCTCAGAGATTTCTTCAGCCGATGACCCAAGAGAAACTGCATCCTCTGCCCCTGCCTCAGA
TCGATCCTGACTCTGCTCTTCTGCCTTGTCACCAGAATCGATCTTTTCATCAGGGCAAAAAACAAAAGAGGGTCGGATACC
TGATTTGGATTTTCAGATCTGAGTGATTCAGACGGCGTAGCGTTGAGTGGGGCAGATGGAGACTTAGGCGGTGTTGGGTCA
GTCAATGTCGTAGCATCTGGAGCAATCAGACCTTTCTCTTGGACAAAAGAGACTTGTTGATCACCGCTTTCGTCAGACAGA
GCCTTCTCACGTGCAGTGTTGAACTCGTCTAGTTCCTCCGTGACCTCTGGGAGCAGGTTTATTGTCGGTAACTCATTTCTC
GGAGGAGTTGCCTTAGATTCTTCATCAGTGAGATGGGGTCTTGGTGATTCATCCTCTGCATGTGACGAGGAAGCAATTACA
ACATCACGTTTAGCCATCTGCTTCGTTCTCACCATAGATGTCAAGAAATCGCCTTTTTGATTGAGAGCAAAAGAGACGGTG
AAGAGCAGTTCTTGATTTTAAAACAGTAACCATTTAGGGTTTCTGATTCACAACGTCCCTTAAGAGGGAAACTCAATCAAC
TTGTCCGAAATTGCAACTCTGGTCCCTCCTTGTTAGCCTTTTTCCTTATTTGACATGTACTTTCAAATTCTCTTATCCAAA
CCCAAGTCATTCAAGAACATTTGTGAATCAATCGTGTATCACAAGACTTTCAGAATAGTAATCAAACATGAGTTAACAAAC
TCCCAATTGGCCAATTTTGATCTCATAATTTAAACAAAATTCGTTGTACCTTTTAAGATGAGTACTGCAAGACTTGGCTCA
GCACTTGATTCAGGTTGCACGTCTCACTCTTTGGAGATTTTACGATTTGGTGCACCTGATTATCTTGTTATGGCTAGCTTT
CACATTGCCATGCCTTTATCAATCATGTGGGTTTTCAACCTTTATCCGCGTGCTGTGAAACCCATATTGTCACCCTTGTTT
CCTTTTTTTTTTTAATTTCTTTTGAGCATCTTTCTGAAACATGACTGGCTCTTTTAAGATTAGGGAGGGGCCACAGATTGAC
GCAGAACCTGTACCTCACATATTGACGGCACAACAAACAGCTCTTTTCCACAGCGTTCTCAGGTCCAAACATAATCTGAAT
TATACGTACATCCGGAGCAAGTACCTGCACTAACATCTGCACATCTTGACTCAATGGTCAAGCTTACACACTCCAATTGCA
TTTCTTAGAGTAATAAAACGGGTGTATTCAAGAGATTAAGTAAATATATCAGCCAATTGAGAATTAGTAACTACATGATCG
ATGACTACCTGATTGTCTTCAACTAATTCTCTGATGAAATGGTGCCTGATGTCAATGTGCTTTGTTCGCGAATGTTGAACT
GGATTCTTTGAGATATCGATTGCAATTTTATTGTCGCAGTATACAAGAAGAGGACCAGCGTGCATTCCATAATCAGCTGAC
ATTTGTTTCATCCAGATCAGTTGTGAACAGCAGCTTCCCATAGCGATATATTATGCTTCTGCTGTAGAAAGCGATATAGAG
TTTTGCTTCTTACTCAGCCATGAAATGAGATTGTTTCCAAGAAAGAAACATCCTCCACTTGTGCTTTTACTGTCATCAGCA
CATCCTGTCCAGTCTGCATCACAGTAGCCCACCAAGTTGTCATTCGAGTTTCTTGAGTAGTACACTCCAAGGTTCTCTGTT
CCTTTGACATACTTGATGATTCTTTTCACAGCATTTAGGTGAGACACTTTCGGTTTAGCTTGATATCTCGCACAGACCCCA
ACACTGAATGATAGATCCGGTCTACTAGCAGTGAGATATAACAAACTCCCAATCATCCCTCGATAAAGCTTTGTGTCAACA
TCCTCTCCTGCTTCATCGTGTGCAACTTTCAAGGTAGCACTCATAGGTGGTTTTTGGA**GATCTTTGCTGGTCCGTAGTCCA**
ACATCGTTTTCACCAGACTTTGTGCATATACGGCTTGAGATATAAATACTCCTTTTTCTGATTGATCAATCAGTAGCCCCA
AGAAGTACTTCAGCTCACCACACATACTCATCTCAAATTCTTGTGTCATGTTCTTCACAAAATCGTCTACCATTGATTGAG
ATGTGCTCCCAAAGATTATATCATCCACATAGATCTGAACAACAATTATGTCCTTCTGTTTTTCCATGAAGAACAAAGTCT
TGTCTACACTTCCTCTGATGTAGCCACCGTCCACAAGGAATCCAGTAAGTCTCTCATACCAAGCTCGAGGTGCCTGCTTCA

ATCCATACAAAGCTTTCTTAAGTCGGTACACATAATTGTGATGGATTGGATCCTCAAACCCCTTTGGTTGTTCTGCATACA
CTTCTTCTTGCAATATCCCATTTAAGAAGGCACTCTTCACATCCATTTGATACACCTTGAAGTTTAGTATACATGCCATCC
CTAGAAATAACCGTATGGATTCCAGTCTAGCAACTGGAGCAAAGGTTTCTTCAAAATCAACTCCTTCGATTTGTGAGTAAC
CTTGTGCTATAAGTCGTGCCTTGTTGCGAACCACTTCTCCATGTTCATCTGTTTTGTTTTTGAAGATCCACTTGGTTCCAA
CAATGTTGACTCCATTAGGTCGAGCCACCAACTCCCATACATCATTCCGTGTGAATTGTTCTAATTCTTCCTCCATGGAAA
CAATCCAATACTCATCTCGAAGTGCTTCAGTATGTGTCTTAGGTTCGACAACAGAGACAAAACATGCAAACTGAACCATGT
CCCTGAAGTTGATCACTTTTCCACGAGTTTTGCGACCTTCCTCAACTCCTCCAATGACATCTGACACTGAATGATTCCGAT
GAACTTGCTGAATGACTTGTTGAGGAACTGAAGCTTCTGACTCAATCACAGCCCCTTCTATCTCGCTCTCTTCAGGTGATT
TGTCACTGACCCGAGTTTCTTTCTTTTCAGTGTCTTCATCTGACTCTCATTGCTCCCAAGTTACAGATGTCATGTCATCAA
ACACCACATTCACGGATTCCACGATGACTGCTGATCTCTTGTTGTAAACTCGGTAAGCTGTGCTAGTTCCTGAATAACCCA
GAAAGATTCCCTCATCACTTCTGGAGTCGAATTTTCCAAGATAGTCCTTGTCATTTAGGATGTAGCACCTGCAACCAAAAA
CATGAAAATAACCAATATTAGGTGTTTTTCCATTCCACATTTCATAAGGAGTTTTTGTGGTCCCTTTTCGAACATAAACTC
GGTTAATAATATAGCACGCAGTACTCAATGCCTCTGCCCAGAACCTCTGAGGTATTTTGTTTTCATGAATCATTGCACGAG
CCATTTTCTGTAGAGTCCTGTACTTTCGTTCAAGCACTCCATTCTGCTGCGGGGTTCTTGGTGCTGCAAACTGATGACTAA
TGCCCTTTTGTTCACAGAATTCCATCATAACTCCATTCTCAAATTCTCCACCATGATCGCTGCGTATCTTCTTGATTCCTC
CTCTTTCATTTATGAGTTGAAGAGCCAGGATTTTGAAGCTTTCTGCAGTTTCCGATTTCTCTCTGATGAATCGAACCCAAG
TGAAGCGAGTGTAGTCATCCACCAGCACAAACACATACTTCTTGCCTCCAATGCTTTCTTTCTGCATTGGCCCCATGAGAT
CCATATGAATTAGATCAAGTAGTGCTGAAGCTTGTACATCTGTAACCTTCTTGTGTTGAACTTTTACTTGCTTTCCCTGAT
TGCAAGCTCCACACACCATCTTGTCATCAAATCGGAGCTTCGGTACTCCCCTAACTAGATCCTTATGCACCAGATATGTCA
TATTCCTGACATTCATGTGTCCCAATCTTTGATGCCATAGTTCAACATTTCCTTGCGCTGTCATGCATTTGATCTTCTTTT
CCCACATGTAGCAGTTGTTACCAGATCTGACACCTTGCAAAATCTTCACGCCATATGCATTTATTGCACGACAGTCAGTCG
CTGAGAATACTACAGTGAGTCCTTCATCACACAATTGACTAACACTAATCAGATTTGCTTTTAGTCATTTGACGAAGTATA
CATTTGCTAGCTTAGGTAACTCTGAGTTGCACGTGATGCCTTTGCCTTTGATGATTCCATATCCTCCATCTCCGAAGGTAA
CCTTTCCACCTTTAACTTTTGACACTTCTTTGAGATACTCAGCATTGCCTGTCATATTTCTGGAACACCCGCTGTCAAAAT
ACCATGGTTCATCTGAATCTAGCTCATCTGTGACACGTGCCATGTTACACCTAATTCCACCAGCTCCACTGTTAGCTGCTC
CTGGATACAGACCAGTCTTCTTCATCCAGACTTGATTCGTTTTTCCAATCCTCCCAGAACTTCCGTTGGCTTCCACATTTG
ATTGACCCTGTTGAGGTACTTATAGCAGAACCTTTTGTAGTGTCCGAACTTTCCACAGAAGAAACACCCAGTAGTCTGATG
CACAGTTGGCTTTGAGTTGAATTCCTCTCTATGTGCATAGCCTCCTGACACAAATTATGTCTTTCCTGTATTACTACTCTG
CCTCCCAGTGTAGCCCAGACCCATGTGTGAGCTTTCTGTTCTTCCAGCACACAGAATCTTGTCCAGTTGCTTGGTTCCAGT
GAGCATCTTGATGTTTTTGTTCTGATGGTCGAGTTTTGCCTGGAGATCTGCTGACACCTGTTTCTCAGCTACCACTTCCTG
TCTGAGTTCTTCAGCTTGTTTGGCAAGAATTAGTTTTTCCTTGATTAGTCCGACACTCTCCTTAGCCAGTTCAGCTTCTCT
CTGTAGCTCAACTCGAACAATGGACAGTTCAGCTTTTAACCTCTCTTTTTCCTTAATGAGTGAGAGGTTCTCCATACTGAG
TTTGATTAGTGTTTCACGCACCTCCTTGTAGCTTTCATCTAACTCTCTTTCTGGTTCACCATCGCTTTCAGAGCCTGATTC
AAAGTCAGTTATTCCAAGAAATGCCACAAAGTTATTCAACTCTTCCTCACTGTCACTTTCTGAGTTATTGTCATCGATTCC
AATCATGGATCTTTCCTTTTTCCTTTTTCCATCTGCCTCACACTCAGCCTGAGTATGTCCAAAACCTTGACATCCATGACA
CTGAAATTCTCTTCTCTTGACCGTTGGACAATCGGTTCTGAAGTGACCGTATCCCTTGCATTCATAGCACTTGACATCAGC
CTTCTTGTTCCCCTTGTCAGATTCAGCAGCTGTCTTCCTCCATGGAACAAACCTCTTCTGTCCTTGCTCTACCCTGCGTAG
AGCTCTGTCAAATCGCCTGACCAATAGGCTTACTGGATCATTATCTGCAGCTTCAGTTTTTTCTACTGATGCTAAAGCGAT
TCTTTTTTCCTTCTTTCCACCAGAGACCTCCATTTCATGAGCTTGAAGCATCCTGACTACTTCATCAAACGTCATCTCGTC
AGTGTTAAGAGAAACGCTCATAGCTGCCTTGTATGGCATGAATTTAGCTGGTAGACACCTTAGGAACTTCTTCACTAGATT
CTTTTCTTTGTATTTTTTACCAAGCGTTAATGCTTCCTGAGCTACTGAGCTGAGCTTGGAGCTAAAGTCACCAATCGATTC
ATTTTCATCCATTTTAAGTTCTTCAAATCTGGTTGCAAGCATGTCTTTTCTGGAGCTCTGAACCTTCAGTGTTCCTTCAAA
ATGTACTTGGAGAATGTTCCATGCATCTTTTGCTGACTCACACCCTTGAATCAGCTCAAATTGCTTTCTCGCTACGCTGCA
GTGAATAACAGTAAGAGCTCTCACATTGAATTTAGCCATCTTGTTTTCTTCATCTGTCCATAGCTCTTCTCCCTTCGGTAC
ATCAGTTCCATCATCACCTCTGGCAACAGGTGCCGTCCACCCAGATTCCACAGCTTTCCATGCCAGAGGATCTATCCCCTT
GATCGTTGCCTTCATACGAGCCTTCCAAAAGCTGTAGTTTACTACTTCAAGGATCAACTTTGAGTGCGACAACATCTCACT
GTAGCTGTCCATCTTACTCCGCATGATCTTACCTGTTTTGAAAAAAAATATAGATACCCGCTCTGATACCACT***TGTTGATG***
TAAGATAATCAAGCTACACTAACTAGAACACACAGTAACACAGAGATATTTCTTTTCAATGAATTCGACTAGATTAGTCCT
TTTAAAAGCACATACAAGTTCTTCTTAAAAGCAATCATTCAAACAAGTTCAAGACAAACCTTGACTTGTTGACTAGTCTTC
CTCGTTCACAGACTTAACGTAGTCTATGAACCACTCAAACCCTTAACCCCTAGTGGCTTTGCTTCTCCTTATTCGGACTTG
CAAACCTCCTATTGCTAAAAGCTCTCTCTGTGTGTTCAGCAGTCACACCGTGCCAAGCCTCCTCTTTATATTCTTCAGAGG
AGAAACCCTAGACATAAAATCTATTTAATGGAAACTTTCCATATCTTCTTCTTCAGCGAATAAGCCAATCTTCCTTTCTTC
TGAGTTAGATTGCTTCTTCTCCAATTCTTCCTTTAAACTGACTCCTTGAGTAATTCTTCCTTGAATGCTCATTAACTCACC
CGACCAGCTTCTTGTATCCTCATCTGATGTCGTAATTCATGTCGTGCTTCATGAACTACTACAGCTCCGTCTTGACTATAC
ATCTTCA**TAGCT**

***BrGYP5*: Gypsy retrotransposon in *Brassica rapa* (AC189430.2) (5′-3′)**

CTAGG*TGTAGGAGATGGATTACATCCCTCTACAAGGCCCATCACATAACAGGCCCTAGCCCGACAAGGTCGATCAAGCTTA*
GTCGGCTTAGCGGCTAGAGACGTCAGCTCGACCTCAGAGTACTTTAAGCCGGTCAGCTAAGCCGATCAGCTATACTCGGCT
TGGATGAAACAGTACTAAGGCCCACATACTCGGCCTTTGGGCGACAAGGCCCAAAGGATAGGCGCGATTAGGGCACCACGC
AGAAGGGCACTATAAGAGAGAAGGAGGAGGCAACGAAAGGAGGACTTGGGGGAGAAATCACATACTAAGCGGCTAGATTTA
GGGTTTCCTAATCATCTCTTTGATCTTGTCGTTTAAACCTCTGATCTTGTCTCCTTACCTTCGATCAGTCTTGTAACCCAC
TGTGTTGATTCTAATAAAAACGTCTTTAGTCATCCCATTTTCTAAGTTATCATTCTAATCGACTGAATCCTGTACAAACAA
TTGGCGCCCACCGTGGGACCGACTAAAGCAACGTTCTAGATCTACATGGCTCAAGACGACGCCGCCTTCGGCGCCCCAGGC
GAGGAACCGACGCCTACGCCTGCGGCCGCGCCGCCCATCACCTCAGAATTCATGAGCTCCGTCATGGCTCGACTCGCCCTC
CAAGACGAAGTCCAAAAGACAACCAACGACCAACTCGCTGCGTTGGTCGAGGCGCTCACAGCCCCTGAGGGACAAACTAGC
CATCCCCAGCTGACACGCCGCCGCCTCTTCAACACAAATCCTACGGCAGCCGGAGTCGACCATATCTCTGACGACTCGGAG
CCTAACGAAGCCTTTCTCGCAGACGCTCCCCCAGCAGGCTCAGATCTCACGACAATACGCGAGCTCGCCGAGCTCAAACTC

AGCCTTCAACAAATGGGAGAGAAGATCCACCATGTAACCAGCGCAGCTCCGCAAATAGAGAGCGTACTCGCCGCAACCTCG
CGCACTCCTTTTACTCGCGCGCTAACTAGCGTCCAACTCGGAAAGATAGAAAAGCTGCGCCTACCTGAGTACAAGCCCGGC
GGAGACCCGGTAGAACATATGACCGCTTTTAACATCGCGATGGCGCGAGCTCGTCTCCCCGACGACGAAAGGGACGCAGGT
TACTGCCAGCTGTTCGTCGAGACTCTTCACGAGCAAGCCCTGACTTGGTTCTCCCAGTTGGAGGAGAACTCAATCGGATGT
TTCCGCGACCTATCAGCAGCTTTTCTCAAGACATACATCATGTTCACAAAGCGCAGCGCCACCGCGTCCAGCCTATGGAAC
CTCAATCAGAAAAAGGATCAGAGCTTGCGCGACTACATGGAGAAATTCAAAGCCGTAGTGTCAAAGATTGAGATCCCAGAC
GGAATCGCTATCGATGCATTGCGCAACACCTTGTGGGTCCACTCCAAGTTCCGAGAAGACCTGTACCAGAATCCAACCAAG
TCGCTCCAAGACGCTATCGCACGCTCCGATAACTTCATCCGAATGGAGGAGGACACCAACGCAATCCTCAGCAAGATGAGC
GCACCCAAGGCTCCAGCGGCTAAGAACGCAAATGCGCGACAAGAACCGCGCCAGCACGCTCCAAACGACAAAAACGGTCCC
AAGGATGGTTACATGTATGTCGTGAACGAGAATAACACACCAATCTCCACTCTCGTAGTTCGCGGGGAGGGGTGGAACAAG
TGGGTAAGAGAACTCGAGTCATCCGACCAAAAAGTCGATTCTGTTTGCACCACCCAACCCGCAGCTGGAGTCGGATCAGCA
GCAGGACCTTCCCGGACCGTCGACCTCACCAAGCATTGCAAGTATCACGACGTCAAAGGACACGATACCTCAGAATGCAAA
TCTCTCTACGCGCATTACCTCTCGTCCCTTGCAAGCGGCGAGTTTAAGTTTGAGCCATTGAAAGCCAAACCAAAGAACGGT
AAGAGCTGGAGCAAGAACAAAGAACGAAGGTCCCAGCGCAAAGCCACTGGCAGAGGTCGACAAAACGACGCTCCGCAACGA
GACGACGAGGAAGAAACCCCAAGGGATAACGGCGGGGGAGACTCCTCAGCCGACGAAGAGCATCCGGCTAATCGCAGACGC
ATTGAGGTTATACTCTCTCAGCAATCCTTGTCGTCCGACGAAGATAATGACGATTCGCCTGTACCCGGAGACCTGAGAGAC
AGCCTTAAACGGCGGCTCGCACCGGAAAATGGAAGCGATACCACACGCAGAGATCTCCGGACGATGCTAGATGCACGAAAG
TCTCGGCGCATCTCGACAAGCGTTGGCAACAACAACGAAGGGCCAGTCGGCGACCTCCGAGACAAACTCAATGCCGGAGTA
AGCGATCTCCGCGTGAAGCTTAACAAGTCAAAATCAACAGACTTACGACGACAGTTGGAACGAGCTAACGGTCAACCTCAA
CTTCCACCTCCTGATACCAGCGTACCAAAAGACCTCCGCGCCTTACTGAACTCCAAGCGAGTCCAAACGGGACAGTCTTTA
AACGTCATCATGGGGGGGTTCCCCTAGCGGCGACTCAGTCCGTTCCGTGAAAGACTATCGTCGACAAGTCACGACGTCCCA
GAAGTGGCCGAGTAAACCGTCGAGTCATCCTCCAATAACCTTCTCATCAGATGACGCTGAAGGTGTTCACGCGCCCCATAA
TGATCCCCTCCTCGTCGTCCTCGGAATTGGAGAGTATGATGTCACCAAGATCCTTATTGACACCGGGAGTTCCGTTGACCT
CATCTTCCGAGGAACTCTGCAGAAGATGGGAGTCGACCTCGACGACATAAAAGCGTCCTCCAGAACGCTAACAGGATTTAA
CGGGTCATCCGAAACCATCTTGGGAACGATCCGCCTCCCGGTGCGTGCATGCGGTGTTACTCGAACGGTCAAATTCGCCGT
CGTTAGCACAAAAGCTCCGTATCACGCTATACTCGGTACTCCTTGGTTACACTCGATGCAAGCTGTCCCTTCCACCTACCA
TCAGTGCATCAAGTTTCCCGGCGCGGACGGGAAGATAAAAACATTGCGTGGGGACCAAAAGGCCGCTAGGGATCTCCTAGT
CGCCACGGTCAAACTCCAACGAGCGTCGCCACTCGTGAACTCAGTGTCCCCTCCAACCCCAAAAGTCTACTCCCAGGAAAA
CGAGGTCCTCGAGTTACCTATTGATGACGCCGACCAGAGCCGCACCGTGAGAGTTGGCGCATACCTCTCCGAAGAAATGCA
GCAGTCAGTTCTGGATTTCCTCAGACAGAACGTATCCACGTTCGCTTGGTCCATGGCAGACATGAAAGGCATTGACCCAAC
TATAACGACGCACGAGCTAAATGTCGACCCAACTTTCAAACCTATCCGACAGAAGAGACGTAAGCTCGGCCCCGATAGGTC
TAAGGCCGTGAAC**GAGGAAGTCGACAGGTTACTCGGTGCAGGTTCGATTGCCGAGGTCCGCTACCCCGAATGGTTGGCAAA**
CCCAGTAGTCGTCAAAAAGAAAAACGGCAAGTGGCGCGTCTGCGTCGACTTCACCGACCTGAACAAAGCCTGCCCAAAGGA
TAGCTACCCTCTTCCCAACATCGACCGCTTAGTCGAGTCTACAGCTGGAAACGAGATGCTAACCTTCATGGATGCCTTCTC
CGGTTACAACCAAATAATGATGCACCCGGATGACCGCGAGAAAACGGCCTTCATCACGGATAGGGGAACCTATTGCTACAA
AGTCATGCCATTCGGCCTGAAGAACGCCGGAGCAACCTACCAAAGGCTTGTGAACAAAATGTTCGCAGATAAGCTGGGTAC
CACCATGGAAGTGTACATCGACGATATGCTGGTTAAGTCGCTCCATGCCCCCGATCACCTCCGCCATCTACAAGAATGCTT
CGAAACTCTCACCAAGTATGGCATGAAGCTAAACCCAGCAAAGTGCACGTTCGGGGTCTCTTCTGGCGAGTTCCTTGGTTA
CATTGTCACACAGCGAGGAATCGAGGCGAACCCAAAGCAAATATCTGCAGTTCTAAACCTCCCGAGTCCGAAGAACAGCAG
AGAAGTGCAGCGGCTCACGGGCAGGATAGCCGCTCTAAATCGATTCATCTCCAGATCCACCGACAAGTGCCTGCCATTCTA
TGATCTCCTGCGAGGAAATAAAAAGTTCATTTGGGATGAGAAGTGCGAGGAAGCGTTCACTCAACTCAAACAGTACCTGAC
CACGCCCCCAGTACTCGCTAAGCCAGACGTCGGTGATGTTCTATCTCTCTATGTCGCAGTATCACAGGCTGCAGTCAGCAG
CGTTCTGATAAAAGAAGACCGCGGCGAGCAAAAGCCCATCTTCTATACAAGCCGACGCATGACAGCACCAGAGACGCGTTA
CCCAACTCTAGAAAAGATGGCTTTAGCAGTCGTTGAAGCAGCGCGAAAACTACGACCATATTTCCAGTCGCACTCCGTGGA
AGTACTGACTGATCAGCCTCTCCGGACAATACTCCAGAACACTAACAGATCTGGCAGACTCACAAAGTGGGCTATCGAACT
CGGCGAGCTCGATATCATCTACAAGAACCGCACGGCAGCGAAATCCCAGGTCCTAGCCGACTTCTTGGTCGAACTGGCCCC
GGAATTAGAGCAAGATCTCACATCCCCAAGCTCAAACTGGACACTGCACGTCGACGGATCGTCGACCAACAAGGGGGCAGG
CGCCGGAGTCCAATTGCAGTCCCCGACCGGTGAGCTAATCAGACAATCTTTCAGCTTTGGCTTCCCCGCGTCAAACAACGA
GGCAGAATATGAATCTCTGATTGCAGGACTCCGCTTAGCAAAAGCCGTCAAAGCTAAACGACTAAGCGCTTATTGCGACTC
CCAGTTAGTCGCCAGTCAGTTCAGTGGCGACTACGACGCCCGCAACGATCGAATGGACGCCTATCTCAAAATAGTGCAAAG
CCTGGCAGCAGAGTTCGAATTCTTCGAACTCATCAAAGTTCCGAGAGGAGAAGAACGTCTGCGCCGATGCCCTCGCCGCCCT
TGGCAGCAAGCTTCGTGATCAAGTGAAAAGAACCATCCCGATACATCGTATCGAGAAACCAAGCATCGACGTCCTAACCGA
TCAAACCCTCATCGCCCAAGTCATTGAACCCGCCACTCCAGACGACGATGGGTTTGGCCCTGACTGGAGAACTGAATTCAT
CAACTACCTCTCGAAAGGGGAACTCCCAGCAGAAAAATGGGCAGCTCGCCGGCTAAAAACCCGCAGCGCCCATTACGTTGT
TCTCGACAATGAACTGCATAGATGGACTGCGAGTAAGGTGCTCTTAAAATGCATCCATGGCGACGAGACAGCAAGGGTTAT
GGCGGAAACGCATGAAGGCGCCGGTGGAAATCATTCGGGCGGACGTGCGTTAGCAATAAAAGTAAGGAGTTTAGGTTTCTT
CTGGTCAACAATGAACGCCGATTGTGAGTCTTATGCGAGAAGCTGCGACAAGTGCCAACGGCACGCACCCAGCATCCATTG
TCCAACCGAACTGTTGCGAACAACCACCGCTCCCTACCCGTTCATGCGATGGGCGATGGACATCATAGGACCACTCCCTTG
TTCCCGCCAAAGACGCTTCATCCTCGTCCTCACCGACTACTTCACCAAATGGATCGAAGCTGAAGCATACGCTCAAGTCAC
AGACAAAGAAGTCCGCGGCTTCGTCTGGAAAAACATTATTTGCCGCCACGGACTGCCCTACGAGATCGTCACCGACAACGG
GTCACAGTTCATGTCAGGCAACTTTAAGGAATTCTGTAGCAAGTGGAACATTCGGCTAAGCCCGTCCACTCCACGTTACCC
GCAAGGTAATGGCCAAGCCGAATCCTCCAACAAACTCATCATCGACGGCATTAAAAAGCGTCTAGACCTCAAAAAAGGTCA
CTGGGCTGACGAACTCGACGGAGTCCTATGGAGCCATCGCACGACTCCACGGGGATCGACTAAATCGACACCCTTCTCTCT
CGCTTACGGTGTTGAAGCCATGGCTCCTGCTGAAGTTAACGTTTCAAGCCTCCGACGTTCCAAAATGCCTCAATACGTCGA
GCTAAACAAGGAGATGCTACTCGACGCTCTCGATGAGATAAGAGAACGGAGAGATCAAGCCCTGCTGCGCATCCAGATTTA
CCAACATCAGATTGAGAGCTACTATAACAAAAAGGTCCGGGCCCGACCTCTAGAACTCGGTGATCTCGTCATGCGCAAAGT
GTTTGAAAACACCAAAGAGCTTAACGCCGGTAAACTCGGCGCCAGGTGGGAAGGACCATACAAAATCATCAAAGTTGTTAA
ACCTGGCGTATACCGGCTCCAAACCTCGCGCGGAGAAGAAGTCCCGCGATCATGGAACTCAATGCATCTACGACGTTTCTA
CTCGTAGAAGTATCTTAGTAAAAAAAAAAAAAAAAAAAAAAAAAAAACGAGTAGATGCACCCCCGTGGTCACTTCTACTCGA

CCGAGTAAATGCGCCACTTACGGCCACTTTTACTCGGGAGAAAAACGAACTACGAATGGCTTGATCCTCAACCGAGGTACG
TAGGCAGCCTTAACAGGTCCAGCTGTAACAAAAAAAAACGAGTAGATGCACCCCCGTGGTCACTTCTACTCGACCGAGTAA
ATGCGCCACTTACGGCCACTTTTACTCGGGAGAAAAACGAACTACGAATGGCTTGATCCTCAACCGAGGTACGTAGGCAGC
CTTAACAGGTCCAGCTGTAACAAAAAAAAACAGAGTAGATGCACCCCCATAGTCACTTCTACTCGACCGAGTAAATGCGCC
ACTCACAGTCACTTTTACTCGGGAAAAACGAACTACGAATGGCTTGATCCTCAACCGAGGTACGTAGGCAGCCTTAACAGG
TCCAGCTGTAACAAAAACAAAGTCAAAACCTTACCTAGGTCTCAATTGAATAAACAATGACTTTCACATCGTGGTTCCCGT
GCCTCCAGCAGAGGATACGCGGACACTCACATTCCTTTGCTCGCGGCTATGTCCTGATCAGACACTTGGCTCCAGGACTTT
AACAGTCACGGTTCACCTATGCCTAAGCATTGAACCGACTAAACAGAGTGAATCTTTGCCTTTTCTCGACTAAAATGTAAC
GGCTTAGCTACCCGATTGCTTAATTGTCACTGAGGAAACGGACAGAAGGACCTTTCACTCTTGTACTCAATTCTTTGTTAT
CGACTTTTGGACAACTCCAAAACGAAACATGATGACAGCCGAGTAAAACCAATTACCGGTTCATCACTTGTCTATTTGCAA
AAATTAAGAATATGAAAATCTGATGTTTCAAAACACAACAGATAAGTGAAATCGCTTAAGGGTTGCAAATACTGACAAGAA
ACACAAAGTAGAAAGAGTTTTAAAGTACAACAACAAGGTCTATCAAGACCACCGATGAAAGCATCCTAAATATAACATACA
ACAAGATCAGACGACAACGTCTTGGTCATCCCTACTCGGGTCAAGCGCTTCGATCGTCTCTTCTTCAACGACTTGATCGCC
AAGGCCTTCAACCTGAGCAGAAGGGTCTGGAACTGGCAGGTTCGGGACCATCTCTTCAACAAGCGCCGGCGATGACCCGTT
CTCCTCCACGCCCACTCCGTCCTCTTCCCGCTCCTCAGATGAAGAAACCAAGACCGGATCTTCAGCGGCTACATCCCCCGT
ATCCGTTTGAATCGCTTCCCCGGTCTCCTCGGGATCGCTCCCGTCGGGGCGCTCCTCGATGGTAGTTCCTTCGGGAACGAC
GATACGCTCCGACAAAGCAGAAGTAATGTCTACCATCGGCTCATCGACTGGATCTTCAGTCGCCTCGCGGGAAGTGATCAA
CTGGGAGGCTGATTCGTGGCCAATCAGACCAACATTTGATCCATACGGGTCGAGTACCCCCATGAGCTCTTCATTCACAAA
TCGAGAAGGGAGGCTGAGAGGAGAGAGAGCATAATCATCCTCAGAGAACGGCTCGAGGTAAAGATTGGCGACTTCAGCCTC
ATACATCTTCTCCTGCTCCGAGAAGATGTTGATCATACTCGGCGGAATCGCGATCCCGCTATCTTTAATCATCTCAAGGCA
TTTTTTGGTTCCTGAAGCCTGCCCATATAAGTTCTTGGCCTTCTCTCGCGCGATAAGGCGATCCAGATAGCCCTTCATCTT
ACCGATACAGCGAGTAGACTTGTCCGCCATAGCCGTTTGGACCTTGATCCTCTCGCGGGTAACCTCCTGAACCCTGGAATC
CCTCAGGCGTTGCCTCTCTTTGACCAGCGTTCCGACGACGGCGTCCCTCTCTCCCTCGAGCTCAGCCTTCTCCTTCTCAAG
GGAAGCGACCAATCGCTCTAAGCGAGTTTTCTCGCGGAGAGCGTCCTTCTTCGCAAGACGGTCAGACTTCAGCTTGTCTTC
CAACTCCTCGAACTTAACTCGGAGGACCTCTTTCTCTTTGGCCGCTTTGTCGCTGGCCTTTTTGTTCTCGGCGCGTAACCT
CTCGATCACGCCCAATCTGGTCCGTGCGAGTTTATCCGAGGCGCCCAGCTGGACCATCGTCTGCTTCAAAGTGCTGTCGTA
CTTCTCCACGAGATAGTTCATGCTCCCGTCACTCTGCGTGAAGCAAAAGAGACTTAGAAAAAATTTTTAAGAAAGAGAAAG
GGAAACAAGGGGAAAACAAGTTTACCCGTCTGCCCGCCATAGCCGCGTCAATGTACTCCTTTTTGAAGTATAAGTCGTCGA
CCGGCGGCAACTCTTTGGTCCCGCCACGAATCTGACGAGTCAGTTCAGCACACGGAAGTGGATTCAGAACCAACGGAGTCG
CCTCGTCGTAAAGGAATTCTACGTGATCCGGGAAGTGAACTCCTCCTCCCGCCTTCCGCGGAAGCGAACCTCCGCTTCCCG
CACCAGTCTCCCTCGCAGAAGACGAAGGGGCTCCCTTACTCGACGGAGATTCATCCCGTGAGCCATCTCCGACTTCGGTAG
CTGGTTCCCCACTTCTAGAAGGAATCTCGGGAGCAGAAGTCCCATCGCCTTCCGCTGACTTCTTCTTCAGTTTCTTCTTGG
GACGCTCCTCAGGCGATCCCCGAGTAGAATCGGCCGGATCGTTTCTTTCCGCACCATCATTCCCTGCGGCAGTTGGAGTCT
CCATAGCCCCAACAGAAGTCTCTTCACGAGACCTTTTCTTGCTCTCCTTTTTCTTCTTCTTCTTAGCCGCGACTTCAGAGA
CATCGGCAGAAGGTTGTGCCTCTTCCAAAGGGACACTCCTGTTCTTGGCCTTGGCCCTATTGGCCTTCCTCTTTGGGGAAC
TCTGAGCTAGCGGCTCAGAGTTCACATCTCCATCTGTGGAAACAGGCCTGACTTCTGAGGCACCGGCAGAAGACGATTTTG
TCGAGAGCATTTGAAGCTTCCCTTTCAACAGGGCACTTAAGTCGGGAACTCCCTCCATCTTTCTAGCTTGGTTAAGGAGTT
TTTGTTGCTTCCGAGTAAATAGGGAGAGACGTGATTTGCGGGGACCGAGTACACAAGGAAGTCTCGATTCCCAATCAACTA
TATAAAAAAGAGAAATCAGACACGGAGACTATGAAAAAGGTTCAAGAGCATCTCTCTAGGATAAACGGCAACTTTACCTCT
GGCGATCCTAGCTTGTTGCCGACGTATCCACTCCCGACTAAGATCCGGCCACCTGAGATGACTATGCGTCGCGATCAGTTG
AGCAGTTTCGAAGAATTTCTCTGGATAGGCAATTGTATTGGGATGACGAACTGCAAAAGAAAAAGGGAACATCAGAACCAT
CGACCAGAATACGAGTAAACAGACAAATTTCTACCAAGTTGCTGATTCCATAAAACGCGATAATCGTCCCCCGGCGGCTCC
TCGAAGGCATGCTCGTCGGACTTAATGAAGAAGTAAGCGCGTTGCCAATCCTGCGTCTTGTTGGGATGACCAGTCAAAACG
TTGTAACTCGCTCTCATCTTTACCGAGAAGATCCCGTTAGGCTCCGCCTTCGTGAAAGTCAGCTCCTCAAACACCCTCACG
CTCATCGAAACATCCATCTCCGCTGCCATAACCATCAACATGACGGCTATGCGCAGCGACCCGTTCAGCAGCTGAGAAATG
GCGATATCCCGACGAAATGCGTAAGACGTGATCAGTCTGGGGATTGGGAACCAGAGCTTCGTCTGATCCTTGAAGTAAGAT
TCATACACGCACTGGTAACCAACCGGTGGCGACCACGGCCGCTGATCGGTTGACGGAATGAGGAAGGTGACTCCAGCGCCA
CCACTCTCTCTTAGCAATCTCTTCACGGTCTCGTTGGAAGAACAGGAACTGAAAACGTTCCCCCACGATTGAACTCGCGGG
TCGCGTAACGCTTCGGGAGGCAGCGAAGGTAGTTCTTCGAAAATCCCACCAGGGTGATATATCACAGGACGACAGTCTACT
TCGTAGAACGGAACACTTCTCGGAGCAGGGAGACGGCCTGACTGGTCGATACGACCCCTCCGACGAGCCCGTCTCCTAGGT
CTTACGACAAGGCTAGGCGCTTCAGAACCACTCGCCTCTCTATCCTCGTCCTCAACGTTTTCTTCGATTTCTTCGCGAAAC
TGTCTATGGGCATCAGCGACCAGAAGGCGCTGGGAAAGGCTCATGTTCTCCGTATCCCGCAGAGCATCACGATGGATTATG
TCGAAATCCTCCAACGGACCCCCGTCCGCATCCCTAGCCGGGCTTGGCGAAGTAGCTATATCTTTCCCTTTTTCCTCGCGC
GATAGCCGATTCCCTGAAGCCATATCTCTTCTACTCGGGGGGACGACTAAACCAGCTAGATCGCTAGCAAAAAACCCTATG
CTACGGAAGGCTTATCTAGAGAGAGAAAATAAAAGTAGAGAGAAGGGAGAGAAAGTGGAATACCTGAGCTTGCAGAGAAGA
ATGACGAAAGCAAAGGAGAGGCGCCTATTTATAGCAAAAACGAGGCGCATCCTTCGAAGCGTCATCATTAGGTCTACAAAG
GCCTAAATGGGCTTCAAAGGCCGCCTAAATGCCCAAAAACCTACTCGCGACGGTTCGGTTTCTCGCCAAAACCGGTTTTCA
ATCAAGAAACCGGACCGAAATTAATCTCCGGTTCTAGCCCGAACCAATTCGACAGAGTTAACCATCCAAAGACCGCCTGGA
CTCCTAATCCGAGAAGTTTACCTCCCGAGTAAAACACAACGCATCGTGAGAAGACGTAACGCGGAGCGACTAGACGCGGCA
CGCGACTAAAACCATACACCAGTAAAACGAGCTGTCGTCGGCATTAGGCCCTGTTATATGAGCGAACCTGACCCACAAATT
CACTGAGACCGACAAGCCGCTCACAAGTGAACTGGGGGGGGACTTAT***TGTAGGAGATGGATTACATCCCTCTACAAGGCCC***
ATCACATAACAGGCCCTAGCCCGACAAGGTCGATCAAGCTTAGTCGGCTTAGCGGCTAGAGACGTCAGCTCGACCTCAGAG
TACTTTAAGCCGGTCAGCTAAGCCGATCAGCTATACTCGGCTTGGATGAAACAGTACTAAGGCCCACATACTCGGCCTTTG
GGCGACAAGGCCCAAAGGATAGGCGCGATTAGGGCACCACGCAGAAGGGCACTATAAGAGAGAAGGAGGAGGCAACGAAAG
GAGGACTTGGGGGAGAAATCACATACTAAGCGGCTAGATTTAGGGTTTCCTAATCATCTCTTTGATCTTGTCGTTTAAACC
TCTGATCTTGTCTCCTTACCTTCGATCAGTCTTGTAACCCACTGTGTTGATTCTAATAAAAACGTCTTTAGTCATCCCATT
TTCTAAGTTATCATTCTAATCGACTGAATCCTGTACAAACA**CTAGG**

***BrGYP14*: Gypsy in *Brassica rapa* (AC189233.2) in 3′-5′ orientation**

ATCAT*TGATGTTAGGAGTTTTCAAGGCTCCTAAGACAAATGTTGTAGTATAAAAGATTGTCGAACCAGTTCTGAGGGATAT*
CAAAGCACTGAGAATGCAAGTACTCACTTAATCTAAGTGCAACCAATGATTTAGAAGGGTTTTAAACTATGACTAAAACTA
GAAAGCAATAACAGAATGATACTTTCTTGACTAAAGGAAAAGAGAACTCATGGGTATAGGGATTAGACCTTGGGTGATCAA
GTATCGAACTAAGGATGGCAGATGATCAATCAAACTATCAACCTTAAGCCTAGACACAATTCTAAGCAAGCTCTATGTCTA
GATGAATGCTCATTTGCTAACATATCTCAAACATCAAATGTCTTTGGTTGAATAATATGAAAGCAATCATTACTAACAAAT
CTATTAGCTATCTTAGTACCTTTAACAACAAATGTCTTTGGCAAAGTATACTAAAAGCCTAGGAGAGTTGTCTCGGGCATT
TCATCGAACACCTTTCGGGTGAGAAATGCCTAAGGATCAACAACTGAGTGGCCAACTCAGAAGATGCATTATGATTACTCT
ACTAGCAAGGAAATATGAATGATCTACACTAAAACATCCTAGCTCTAACCTAATCACCCTTAATCTCCCTAACCCATGAAT
TCAAAAGGTGATTACTCACTAATCTCCATGATTCCTCTTAAACCCATATTGGATTTCAGATTAATCATGTAAAGAAATAGA
TAAGAAATCAACAAGAACACAAGAACATAACAATCAAAATCCAAGAGATGAACTTCTCAAGAGAGTTCTTGTGTATTTCTC
AATAGATCAAAAGATAAAAGATAATCTGCCTCTGGTGGCTACAAAAGATGTTTAAAACATAGGTTTTTCCAAGTGCAAAAC
GTCCAAAATAAATTGCAAAAAGGTCCTTAAGAAATCATGATTTTCGGCAGCAAAATAACGCGGAGCGACTTGCAGGTGTCG
CTCCGGGAAGTCGCTCCAGGGCCGATTTTTGGTGTCTCCGGGCGAGAGGTCGCGAGCGACTTTGGTGTGTCGCTCCAACGG
GTCGCTCTGGATCGGGAGCGACCTTGGTAGGTCGCTCTGAGAGGTCGCTCCAGGCTTCGCTTCGTGTCGTCTCTCCATAAA
GACGCGAGCGACCTCGGGGTGTCGCTTTGGGAGGTCGCTCTGAGAGGGGTGTGATAACGGAGCGACTTCGTGGTGTCGCTC
CGGGAGGTCGCTCCGGGCTCGTTCTCGCGTCTCCGAGTGATGAAACCGCGAACGACTTCTCCCTGTCGCTCTGGTAAGGTC
GCTCCAATAGGGAAGTCAGAGCGACTTGGTGGTGTCGCTCCGGACTGGTCGCCCCATGCCTTGCTCGCCCAATGACCACTC
TAAACACTCCATTTTGAGCTCCAAATGCACCCAAATGTCTCCAGAAACTCCATGTGGTACTCAAATACCTGATAGAGACAT
ATGTATGCAAAATGCAACCTAAACATGTCTAAATTCTAATCTATATGATGAAAATGTTTATGAATGAATGGATAAAACAAT
GTAAATATGCAAGATATCAACTCCCCCAAACTTGTTCTTTTACTTGTCCACAAGTGAACTTTCTAGAACTCATAGGGAGAG
AGGTTGAAGGTGGGAGCACATAGCCAAAAGAAACACAACTAGCACACTCTCTTATTACACTCTAGCTTCTCTAGGCTCTTT
TTGTTCTTACACTCATCATCAAGCATCCACACATTCAAATCAACCAACTCTCACATTCATTAAGCACAAAACATCAGGTGA
ATTCTTGCAAATGGTCAGTTGGTCCAAGTCATTTGGTTGGGTAAGGGAAGGCTTTTATTCAAGTGATTCAAGAGGTTCAAA
ACATATGATCTTTAAGGTGGTTTACTCTCAAAACAAGTAGCCTTGACATTGCACATAATATATCTAAGAAAGGGACCAACT
CATGCATACAATGCTCAATCTCCATTGTTCTACCCTTTTCTCAAACATACAAATCACACAATCATTTTCCAATGTCAAACC
CAACTCACATCTCTCACCAAAAGATCCTAAGAACTTTAACTCCTTCTCTTTGAAATCTACAAGGGATTTTTTTTCACAACC
TCAAAACATTTCTTAGCTCCTAACAACTTTGCTAGCCCCCTTTTCTTTCTTTTTCTCTTTTTTTTTTTTTTTTCTTTTCTC
TTTTTTTTCGTTTTTCTTTTTCTCTTTTTTTTTTTTTAGGCTGGGGGCCAAGACTTTTCAAAACTTGAGCTAGAGGCTTCTA
TACTGGTCCCAATAAAACAATTCACTTAAGCACAAAGAGTCTATTCTTTTCTTTTTCATTTCTCCCAAATCATAATCACAA
CACTCACCCCCACCTATAGCTAGACAATAGAGTGTCCAATCTAGCAAGAATGAAGATCAAGCATTGTCGTTCCCGATACTC
TCAACATTATGCACATGTAAGACTTTCCAAAAAAGGCCTCACTCATCAAACAATGAAAGCTTAAAAGGAGGAAAAGGTTTT
GGGAGTGGTCTACCACTAGAGTTTGTCAAAAAAAGATTGGTATAAAAGATGTGACAACTCAAGTGTGTATAGCCATGACTC
AGTACACAAGGGACCATGAGCAAGAAGCATTAAGTTCGTTCAGTTCAAATAAGGTTGTAGTTGGCTTCAAAGACTGAGTTT
CAGCAATCAACAAGTTTCAGGAAGAGTTTTCAAGGCTCAAAACATACAAGTCTTTTTGAGAGGTGCAAAAGCTACTCAGGT
GCAACGTAGTGTTCTTTACAAGGCATTTAAAATCATTGCTCCCAATGTAAGTGAATGCAACCTATATGCTCTAGACTCTCC
TAGAAATGCAAATGATGCAAACTAAATGATTGATTTTTTTTTTATATGCAAATAGGTATGCAATGCATGACTCAATGAAAC
AACATCAAAGCAAACATGATCAAATACTTGGGACCTCCCCCAAACTTGAGTTACACAGTCTCTGTGTCGTCAAGTAGAGAG
AGATACCGAAAAGAAGACTAATATGCAAAAATGAAATGGTATATACAAGGGAGTGTGGTGGGTTACCTTCTATGGGGAATG
AGTAGAGGGAGATGCTCCCTCCTCGTCGTCCTCAGGTGGTAACTCTTCAAGGTGCTCATCAGCAGGAGTGTCCATCTCTTC
AATGCAGAAGGCTTGCCCGAACACAGTTGGTTTCCTCATTTTCCCCTTGATGTCAAAGTGGAGGATGTTCTCTTTACCCAA
ATGGAGATCAATCGTGCCCTCCTTCACATTAACAATAGCTCCTGCTGTAGCTAAGAATGGCCTTCCTAGAATCAATGGATC
TTGAGCCTCCTCACCCATCTCAAACACCACAAAATCTGTAGGTATCTCATACCTTCCAATCTTCACAGGGAGGTCCTCTAA
GATGCCCACAGGGTACTTCACTGAACGATCAGCTAACACCAGAGAGAGTCTACACTTCTTGTACTGAGTGAATCCAAGCTT
CTTTGCAACAGACAAATGCATCAAGCTGACACTAGCTCCCAAATCGCAGAGACTTTTCTCAAATACCATAGGTCCAAGAGC
ACAAGGTAGTGTGAAACTTCCTGGATCCTCTAATTTCTCTGGAACATCAAGCCTCTGGATGATGGCATTGCACTCATGGGT
AAGAATCATCATGCCTTCCATCTCCTTCTTCTTTGCAGCTACAACATCTTTCAGGAAATTGTTGTATTGAGGAATCAACAT
GAAAGCATCGATGATGGGCATTGCAACCTGAGCTTCACTCATTTGCTTCTCAAACAGAGCTTTGTATTTCTCTAGCAGCTG
CTTCTTGAATCTACCAGGGAATGGAAGTTTGGGTTCATAGGGAGGAGGAACAAAAGAACTTTCACTTGCTGGAGTAACAAC
TTCACCATCCTTTACTGTCTTCTTCTCCTCTCCAACCTTTCCTTTACCTTTGGCTTCCACTATCTTCTCCAAGATCTCGTC
ATTGATCTTCTCATCAACAATCACCACTTCATCATCTATGTTGATGGCAACCCCCTCACCTAATTTCTCAGCATCCTTGGT
GAGGGTTCTAGGAGGTAACTGCTTACCACTCCTGAGGGTGATAGCTTTGGCCTCCTTGGGATTTTGGTCAGACTTTCCAGG
TAGAGATCCTTGCTGGCGATTCTGGTGAGTGTTCATGGAAGCAAATTGATTCTCTAAATTCCTGACTGTAGAAGCAAGGTG
TGAGAATTTGTTGTTGAGCTCATTGTAGCCCCCATCAATCTTGGAATGAAGGTTCTTCAACTCATAACCAACTTGCTTCTC
ACTTCTAGTCTGAGACTCCAAGATTTGCTTCAGTAAGGTATCAGTGCTGCTCTCTTGAGGAGCAGAGGAACCAGACGAGGG
GTTTTGCTGAGGCTGATAGCTGCCTTGCTGGTTGTTTTTTGGCGGATAGCCACTCTGTTGGTTGTTTGGATAGGATTTCTG
TTGGTAGTTGTTGTACTGAAAGTTGAGCTCCTTTTTGTACCAGCTACCATTGTTGTTGATGAAACACAACTCCTCTTGACC
TTCCAAACCCTCAACCTCATTGACAGCAAGTGGATCTTCCTGCTTGGAGTTACCAACAAACTTCAGCTGCTCTTGGGTTGC
TTTTTCAGCAATGAGTAGGTCGATCTTGTCTTCCAAAGCTTTGATCTCTTTCCTCGTCTGCTTATCATCTGTTCTACTGTC
TCTGTCGTGGTCTCCACTGTAGACTGCATCACTCTTTACCATGTTGTCAACCAGTTCTTCTGCATCCTCCTCAGTTCTCCC
CAAGAAGAACCCATTGCTAGCTGTATCCAGTCTGGCTCTGTACTTAGGAAGAGCACCACGGTAGAATGTGCTCAGTAAGCT
CTCCTTAGAAAAGCCATGGTGTGGGCATTGAGCTTGGTAGCCCTTGAATCTCTCCCAGGCTTCACTGAAGCCTTCCAAGTT
CTTCTGTTGAAAGCTGGAAATCTCGTTTCTCAGCTTAGCAGTTCTTGAAGTAGAGAAGAACTTCTCCAAGAATGCTTTCTT
GCAGTCATCCCAAGTAGTGATGGAGTCACTGAGTAGAGACTTCTCCCACTGACGTGCCTTATCCCCCAAAGAGAAAGGGAA
TAGCTTGAGCTTTAAGGCATCTTCGGACACACCATTGGTTTTTGACAACCCACAGTAGCTGTCGAACCTGTCCAAGTGATC
AAATGGATCCTCTAGAGCCAAGCCATGATACTTGTTGTTCTCGATCACGTTGAGGAGTCCTGATTTGATCTCAAAGTTGTT
GGCTGCCACAGCGGGTGCTCGGATTCCCAATCTATGACCATGAATGTTGGGGCGGTCGTAAGTGCCAATGGGTCGAGCTGC
TCGCTGTTGGTTTTGAGGTATGTCTCCCATATCAGTACTCAATCTCTGCAAGTGAGTCTGTTGCTCTTCTTCTCTTCTATT

TCTAGCACACTCTCTCTCTAAAGCTCTGATGTCTGCAGCTATTGGAACTAGGTTTGATGGACCCCTGCTCCTCAAGTTCAT
ACACCTGTAGATTAAAGGGAGGTGAAGAAAGAGAATCAGTAACAAAAGAAAATAAAAATGACTTAGTCTCAAGCAAGTGAC
TAAATCTCAATGTTCAAATCTACTCAGAATTTGGCAACGGCGCCAATT***TGATGTTAGGAGTTTTCAAGGCTCCTAAGACAA***
ATGTTGTAGTATAAAAGATTGTCGAACCAGTTCTGAGGGATATCAAAGCACTGAGAATGCAAGTACTCACTTAATCTAAGT
GCAACCAATGATTTAGAAGGGTTTTAAACTATGACTAAAACTAGAAAGCAATAACAGAATGATACTTTCTTGACTAAAGGA
AAAGAGAACTCATGGGTATAGGGATTAGACCTTGGGTGATCAAGTATCGAACTAAGGATGGCAGATGATCAATCAAACTAT
CAACCTTAAGCCTAGACACAATTCTAAGCAAGCTCTATGTCTAGATGAATGCTCATTTGCTAACATATCTCAAACATCAAA
TGTCTTTGGTTGAATAATATGAAAGCAATCATTACTAACAAGTCTATTAGCTATCTTAGTACCTTTAACAACAAATGTCTT
TGGCAAAGTATACTAAAAGCCTAGGAGAGTTGTCTCGGGCATTTCATCGAACACCTTTCGGGTGAGAAATGCCTAAGGATC
AACAACTGAGTGGCCAACTCAGAAGATGCATTATGATTACTCTACTAGCAAGTAAATATGAATGATATACACTAAAACATC
CTAGCTCTAACCTAATCACCCTTAATCTCCCTAACCCATGAATTCAAAAGGTGATTACTCACTAATCTCCATGATTCCTCT
TAAACCCATATTGGATTTCAGATTAATCATGTAAAGAAATAGATAAGAAATCAACAAGAACACAAGAACATAACAATCAAA
ATCCAAGAGATGAACTTCTCAAGAGAGTTCTTGTGTATTTCTCAATAGATCAAAAGATAAAAGATAATCTGCCTCTGGTGG
CTACAAAAGATGTTTAAAACATAGGTTTTTCCAAGTGCAAAACGTCCAAAATAAATTGCAAAAAGGTCCTTAAGAAATCAT
GATTTTCGGCAGCAAAATAACGCGGAGCGACTTGCAGGTGTCGCTCCGGGAAGTCGCTCCAGGGCCGATTTTTGGTGTCTC
CGGGCGAGAGGTCGCGAGCGACTTTGGTGTGTCGCTCCAACGGGTCGCTCTGGATCGGGAGCGACCTTGGTAGGTCACTCT
GAGAGGTCGCTCCAGGCTTCGCTTCGTGTCGTCTCTCCATAAAGACGCGAGCGACCTCGGGGTGTCGCTTTGGGAGGTCGC
TCTGAGAGGGGTGTGAGAACGGAGCGACTTCGTGGTGTCGCTCCGGGAGGTCGCTCTGGGCTCGTTCTCGCGTCTCCGAGT
GATGAAACCGCGAGCGACTTCTCCCTGTCGCTCTGGTAAGGTCGCTCCAATAGGGAAGTCAGAGCGACTTGGTGGTGTCGC
TCCGGACTGGTCGCTCCATGCCTTGCTCGCCCAATGACCACTCTAAACACTCCTTTTTGAGCTCCAAATGCACCCAAATGT
CTCCAGAAACTCCATGTGGTACTCAAATATCTGATAGAGACATATGTATGCAAAATGCAACCTAAACATGTCTAAATTCTA
ATCTATATGATGAAAATGTTTATGAATGAATGGATAAAACAATGTAAATATGCAAGATATCA**ATCAT**

***BoLAR1*: LARD-like element in *Brassica oleracea* (AC149635.1) from 8183-14365 bp**

TAATA*TGTTATAAATCAATTAGTGGATGTGCATAACCGTCCCGGTCCATAACCAGTTCATATACCTAGAGAGAGAGTCGGC*
CACGACTTAGAGAGAAAGAGAGAGAGAGAGGCGGCTTGTAGAAAAGGAAACCATCTTTCCTTTCTTTTTGTAATTGTTATC
TTTCCATTATTATGTATTAGTAGTTTTCCTAATCTTAGTGGGTTTAGGTTTTAGATACTTTCATTTTTACTTATCTTGTAA
TCCCCTATATAAGGGAACACTTATTCATTAATGATAGATAGAAACATTCAGTTCTCAAACCTTCGTTTTATAACATATTAT
CAGCTCGATAACCTCTAACACCGTGAGCCTAAAACCAAAATCACTAAAAACCCTAAAATCCCTAAACAAAGAGGAATCTGC
CCCAGAACTTCAAAGATGTTGTTCTCCCAAACCGTAAGGACCCAGAAAACGATCAAGATATCAAATCGAAGCTCTTGCCGA
AACGAATCAGTCAGTGCAAACCGTTTGTCGATCTAACCACTGACGCGTCCTCACGCACCGATACAGTGCGCGCCGATTTGC
TTTGGAACCCTAATCTGTTCTTGCAAGCGTTCCAGCTCGTGTTCATCCGATCAGCCCCAACCCTAAGGCATTCCTGATACT
AGACGAACTCAGCCTCAGCTCCATCTGTTCTCAGCTTGCATCCCGGACAGCTACAACTCGCGTTCCCGATCAGACGCATTC
CGTTCCAGCTCGCGTTCCATTCCAACTCGCGTTCCAGCTCGCGTCCGATCGTCCAGCTCGAGGTTCTCTTGGTGGTCCGGT
TCTACAATCTTCAAAACCCTAAAGGTAATTCGAAATCTGAAAACATGAATCGAATCTGGTTGTTTTGTAAAGATTGAAACC
CTAAAATATAATCTCTAAAAAGCTGAAAGCCATAGGATAATAGATCAATACCCTATGAGTAAATCGAAGCTCTATAAGTTC
GATCCAAACCCTAAAATTGAGATTTGATCCATTGATCAATTAAAATCAGAACCTTGAAGGTTTGAATCTCAAAATTAAACT
CCCTTGATCTAAGATCAAATCGAATACCCAAAACCCTAATTTGAAAAATCGATTTTAATTGTTTGGAATTTGAAACTTTGT
TTGATTGACTGATCACCTAGAACTATGATTATGATTGTTTAAACTTGAATCTTGAAATCTGAATTGATTGGCATGTTTAAA
CTAAAACCCTAATGATCTAGTTTAGATTGTTTAAAATTGAAATCTGAAACTATAGGATTAGGTTAAACTATTCTAGACATA
ATCATATTTGTGGCCGTGTGGCCTTGTTGTATTCATACCTAAACCTAATCACATTGATTGATTGCAATTTCACTTAGAATC
TCATGTTAGAATGGTATTGAAATTGAAATTGAAATCAAGTATGATTGTTCTTTGCTTGGCCGATTAGCTTGCTTGTGTTGA
GTACATTGTATAAACTGATAGAATGCATCTATGCTAGGATTGCAATTACACAACAAACTATATGCATTGATCCTGAATTGA
ATCATTAGCCGTGTGGCTTATTTTCCTGTTTTGTCCGTGTGGCTTCCTCATGGCCGTGTGGCTTCTTCTTGGCCGTGAGAC
ATAAACCTTAGCCATAAGGCTTGCTTGCTTGATTGATTCTATTAATTGTGAGTTAAGTCAGCTAGATTGATTGCATACTGA
ATCTGAAACCCTAAATCTCGAATTTGAATCCCTAAGGGCCTTGAAACCCTAAATCACATGATATTAATCCAAATTGAGTTT
AGGTTTGATTCTTGTTATTTTGTCTGATCTAATTGATGTCTTGAAAGTTAAGGTGCTAGATTATTTTGATTCTCATAAAAT
CTGAAACCCTAGCCACATTTTGTATTGCTAAGATGATAAAACCCTAATTGGATCAACATAGATTGTTTACATCATAGCTTG
GCCGTGTGGCCTTTATTGCACCTTACTATCATAGCCGTGTGGCTTGCATACATCACATCACCTTGATCACATATAGAATCC
AAAAGCTAAGATCGTATGCATCATATATATCTATCTAGCCGTGTGGTCTCTTTGCTTACATCATGATTGTATGACAGTGTG
GCCTTATTGCACATCACCCCTATAGCCGTGTGGCTTGTTCATTAGAAATGCATTAACTCGATTGTTTACTTGGTCACACGA
TTTATTTTCTTATAAAGATTGAGTGATATATGATTACAGATGTCGAAAATCAGCAATCAAGATTATGCTAGCCTAAATCTC
TTTGGAGATAATTATTTACAATGGGCATTAGACACATAGATTGTCCTGAAATCAAAGGGCCTCGGTGAATGTATCATCGAG
GACAATAATGCAAGTGAAAGTAATAGATACATGGCAATAATGATTACTCGCCAACCCCAAGCCAATGAGGCTGCAGAAAAA
AAAATCTAAAGAAAGCAACCACGTCCATCATGATAAACCATACGGATGTAACCGTGATGGATGGAAAGGACGTGGGCATAG
CCAAACCACCTTATACAACCATGGACATAGGAATCACTATAACCGTGGTTGTGGACACAGTTTTGGCCATGGTTAAGGGAG
AGGAGGTTGTGGCATTTCTAAGCCACCACACTCGGCCAAGTCAGTGTGCCATAAGATATGTAATGGGGAATCATTAGGCTA
AAACATCTGATTGTCTAAAAAGAATAATGTTGATTTCGATTTCTGTTTTTGCTTTGCTTTATGTTTGTGATTTTCTTGAAT
TAAGAACTTATATTATAAGATGAATGAATGATTTTAGATATGACTAAGAAAACACCAATATATAAGGCAATGTTGCGGGTA
TAGTCAGTCAGAGAAGGTCATGCGACCAGTATAGAATTGCCTAAAGGGCTACGGCTAGACTTAGTACATTGGCACCTAAAG
AGGCTAAGTACATTGGTGCCTAAAAAGAGCACGCATAAAAGACTTAATGCTCTATATCCACCCATAGAAGCCCATTGAGTT
TATAAGATATATGATAATGGTTTCCATATAGAAACAATGGGCAAAGGAAACAAAGAGAGATCGAAACTATACATGCATTCT
CTACTGATCTAGACTATGCAAAGATCAATATGATAAAGGCAAAAGCCATGGTAACCAGATTGGTATGCCCCACGAAAATTA
TACACTATATGGCATGACCGGATTAGCCATTCTGGTCTAAACTTGATGCAAAGATTGATATTGAATAGGCACACAGAGTTA
TCCCATAAGATCTCACGTTGTGTAACATGTACACAAGGGAAACTCATGAGGCTTTATTGTCATAAGCATCACGAAATTATA
GTATGTTCATGAGGGGGAGAGAGGATAAACTCGTGATCCATGATCACACATGATCAATACTACAAATTTGTGTACTACTAT
GGTCTATGACCATGCTATAAACGGCTCGGCCGCAAAGGCCATTACCTACAAAGGTTCGGGACATCACGAGTTTTACATGTA
TATAAGGTTAATATGCATCAGGCCATATAAGTGAGCATAAATATTCCCATCTCAGATTGAATACGGGTCATGAGCCAGACA
TATACCATCTTAAGAGACTATTGAATAAACCACCACAGACATATGTACCGTCTAAGATGGAACCTTAGAAGAGGATTGAGA

ACATATGTTGGATAAGAGTATGACTTTCTCCCACGACTTCTAAGAGACCTTGAGCCAAATATGGGTGATCAGAAAGTGGCC
AGGTACACGGATTGCATGATCAATGAATCTGACTATCCAACATTAAAGGGAGAATGCTATAAGCTGGTAAAGAGTAAAGAA
TGATAAGAAACAGAATGGTATCAGCCATCATTGTCTTGGCAAAGATCCTCGGATTAGAATTATAGACGTCCAAAGAAAGAA
AAGATTATACATAGCTAGCTAATCGAGATGCCAGACACTGACCCGAAAAGAAAAGAAAAGAATGACTAAGTCATAACCAGC
TTAGCACCAAACGAAATCTGATGTCCTAGAAAGAGACACAGTCAAGTTGTTACAGAGTCTATACAAGATAGACCAAAGTGT
TCCATAAGATAAGTAACCTCGGAAAGAAAAAGAGAAAGGTGCATAGAATACAAATCCGAGGTCATAAGGAAACCAGACCAG
ACATTGAGAAAAGGCTGCGCAGCTAAACATCTAAGGTACCAAACCATGTAGTTTGGGACGCCAAGCTACTAGGTAACAAAG
GTCCTGGATAATAAAATCTCAAATCGATTAGATCATGTCTGGAACACAATGGAACCATATATAAGTGTCGACACATAAAGA
TATATTTGTACATAAGAGAGTAGCACTTGAACATAAAGATATAAACGAGGATCATGAACGAAAGAGTACTCATAAAGATTA
AAGATTAGATTGAAAAGATAAACGTGGGGTTTAAAGTATCTAAAGAAGAGATAGGCGTATTGGCCATACACATATACATAA
ACACCATCTGATAAACCAGTGAAATAGATGGGTCTTGTGAGGGAAAAAAACCCGTGAGATATGGTACATGATCACGTGGTC
TAAAGGTGTTCATGTTTCCTACTAACAGCATTCGATCATAGCTTGTTACACAAGGATACTCACAGAGACCAAATGAACAGA
TTATAAGGAGATAAACTCCTATGTGGTGGATGCTGCTATATATTTTCGAAATTGACAAAGGTCTGGTCATAAGAAAGAAAT
TAGATACAAAGATGCAGTAAGCAACATATGGATCACTGGATAAAGAACAACATTGTTCCTAAGCTGTTCATGGATAAAGAA
CAACATTGTTCCTAAGTTGTTCATGGACTGAAACAAAGTAGCTGCATATATTATATAAAGGAATTGTATAAGGACAATCCA
AATTCAGTCCATACATATACTTATAAAACAGATTTGAATTCTTTGTGTTTACTTTATACTAAACCAGCCTAAAGAGAGGTT
AATTAGTTGATACATATTATTTATTAATTCTGTCTGGCTAGAGGTGTCCGGCCCTTCTTCTTTGATCATGGCTAGTGGAAA
AGAAAGATCAGCCATCATGATTGCCGTCCATAAGCTCGTGTAGTCGAGGTCCATGACCTAATAAATCTTCTGGTGCGTGGT
ATGATTCACGGCAACAAAGGTCATGGGACACCATCAAGGTATGGATAAAAGACTGCAATAGCATTACCGAATGCAAGAGAG
ATTACTGATCGAGATCCATGAGTGTGTTCGGCCGAAGAGCCATAGCTTGATGAGATTATAAAACTCATTACAGCAATAATC
ATTGATCACATCTTGATCTAGAACACTCATGAGACAAGTTATAAACATGTATAGAAGAAGTCTATGACTCAATATGGATCG
ATCAGATTAGTGCATAGTCGATGATAATAAAGAAAGATCTATGGACAGATACATTGGTACACATCCGTGTAGTAGATACAC
CGGAACATCGGTTTACGACCTGAGACTATAAGTTCATAAGTACCATGCTCAGTATATATTGTCCACAGTGTGTTTAGTCTG
TCCGACCTAATTAAGAGTACTCGACAGCCAAAGGTCCACTTCTTGACCATGCACACACGATGTCAGAAGACATGAGAAGAG
ACCGGAAAGGTCAAAGTAATCCAATTGCGGTTCAGTAATGAACTTGGCCGATTTGTTTACTCCCTACCTGCACGTTTAGGA
AAGATCACGCATCAGATGCGTAGACTGAAGAAACTTCCACTGAGGTCCACATCAGGGGGAGTAGTACGTGTTGTATTCTTT
TTCCTTCATCATGTTTTGTTCCACTGGATTTTTCATGATAAGGTTTTAATGAGACAACATTAAGCGTATTACAATCCCTGA
ATGGTTATGGCATCCAAAGGAGAG***TGTTATAAATCAATTAGTGGATGTCCATAACCGGCCCGGTCCATAACCGGCCCCGTC
CATAACCGGCTCATACACCTAGAGAGAGAGTCGGCCACGACTTAGAGAGAGAGAGATAAGCGGCTTGGAGAAAAGGAAACC
ATCTTTCATTTTCCTTGTAATTGTTATATTTCCATTATTATGTATTAGTAGTTTTCATAATCTTAGTGGTTTTAGGTTTTG
GATACTTTTCTTTTTACTTATCTTGTAATCCCTTATATAAGGGAACACTTATTCATTAATAATAGATAGAAACGTTCAGTT
CTCAAACCTTCGTTTTATAACA***TAATA

CHAPTER 4

CHARACTERIZATION OF LINEs and SINEs: UBIQUITOUS COMPONENTS OF *BRASSICA* CROP GENOMES

The sequences of reference LINEs and SINEs identified in this study are given below. The details about the elements are listed in table 4.1 & 4.3. The TSDs of elements are shown in red colour and poly(A) in blue colour.

A 6382 bp *BrLINE2-1* in *Brassica rapa* (AC189630.2) in 5′-3′ orientation

GTTACATGAATACAAGAATACTACTAAAGAATAAAAAAGCCTTTTCTCTCTCTTGCGATTTTTCTCCCAGACACGGTTCCA
GAGAAAAACCTTTACACCCTGTACCCAAAGCTGACCTAGCCTTTTAACCCACGGTTCTTTCCCTTTGTCGGCTTTTATCCT
CTGTAAAACCCTGCAAATTTTCAAATATTTACTACTTTAAGCTATATCCAACGGATAATAGATAGCTCCCACAAATTATCC
TACCGGCTCTGCAGTCTCCTCCGCCGGGTTCAGCCCCCGACTCCGTCGGTGGGTCAGCCAGGGTCCTCCTGTGCCGTCTCC
GTTCTGTTCTCGCATCTCAGCCTCATACAGTCTCTGTACTCTCTCTATACCGGAGCTGGTCTTTATGGATGAGATAAGATA
GATACCTTTCTCAACTCTTGGGTTGTTATAACCTCCTACTAGTACTGCCTCCAACAACTTGATATAGCATTAATCACTTTG
AGTTACTCACCTATATCAGCATTTACTGACCTTGGAAAGTCGAGATCTCACCATCTCTTTAAGATACTCTCCACGCTGGGT
CTGGTTATACTAAATCTCCTTTCAGCTTGAATTATTTGTGCCACTTCCACCATGTCTCGCCGCTACTCTCGTTCGGACAAG
GAAAAATGGAGTGCGGCCGCCCCCCCGCCACCTCCGCCTACTCGCCGTGCTCCTATTCGCATTCCAGACAGTGACCCTGCC
GCCCTAATAGCGGAAAACAAGCTCACTATCATTGGTAGGGTTACAAACCCACGGTTCCAGAGACCTCGAGCTGTTATCGAC
TTCTTGCCCCAAGTCTGGAATTTGGAAGGGCACGTTGAAGGCCGTGAACTAGGGCTAGACAAATTCCAGTTTCGTTTCGAT
TCAGAACACGAGCTTCAAACGGTTCTTGACAAAGGCCCGTACCATTACAAAAGATGGATGTTGATACTACAAAGATGGGAA
CCAGTGGTGTCGGAAAACTTTCCTTCTCAGATATCCTTTTGGGTGAAAATCCATGGTATCCCCCTCCACTACTGGAATGAA
AAAGCTGTTGAAACTATCAGCGATGCACTTGGACATGTCTCAAGCCGAAACGCCAAAGAAGCAAAGTTTCGTGTAGAAGTG
AATGGGCTCCTTCCACTTGAAATGAAAATGGAAATCTTGCTCCCCTCTGAAGAAGTTACAGAAGTGGAGTTTCAATACCTA
AAGATAGAGAAACACTGTTTCACTTGTTTCTCTCTCCTTCATGAAGAGGAGGACTGTCCCTCTCGCCCAAGAGGAGCTCGA
GCTCCTAAGGATAGACCCCTCGGTATTACTCAGGCCATTGCTCTCCAGCGAATCGAAGCTGACAAAAAGAGACACGATGAT
AGAAGAGGCTATAGACGGCCTGCTCCTCAACAATCTGCCCCAAATAATCCTTCCCTGACGAATCGGGAGGAGCATCGCCAT
TACTCTGATAATAGAGACTCTCATCTGTTGCCTTCAGACCCTGCCTCCCATCGGGATTACCATCGCCATTACTCTGAGAAT
AGAGGCTCTCGTTCGATGCATGTAAACCCTGCCTCGCATCACTCTAACAGACCTAGCGGGGGCTCTTTTGAGCCTCGAAGC
CAGCATGATGAAACTGTTAGAACCCCTGCAGCATCAGCCACCAGAGGATCGATCAGGAGGGGATCAGGAGGATCGATCGCA
GACGGCCAAATCATCTATCCAGTGAATCAGAACAACAGTGGAGGGCAATCTAACTCTAGGGAGAGAATTCCCGCTCGTGAT
CGGCTGTCAGGACAATCCCAGGAAGACAGAGTTCCTGCCATGGAACGCCTCTCTGGTGGAGATACATCAATGGTTCCGGTA
TTTGAATTACAGGACATTGGAGGGGAAGTTGCGGTGGAACCCGCGTTACAACCCTCACACTTGGCATCTGGCTCCAGGGTT
CCGGCCTCTCTGCGTCTAGGAAGCCCAAGTGTCTCAGGCAACAGGAACAAAGCAAAGGCTACTGCAGCAGCAGCGTTAAGC
AAACAAGCTGGGAAGCGAAAGGTGTCAAAAACCACGAGCAATAAACGTGTAGCTAGAAGCCCAATGCAAGTACTAAGCCTC
TCAAAAAGCATTGCAGCACGATCCAAGATTTCCACCCGTAGGAGACTATGCCCAGCTCGCTATGCAAATGAATTTGCGGGG

CCAGGTTCAAGCCGCAAGAGACAGGATTCGAGTAATCCCATTACGGTGAATATACCGTCGACTAACAAGGAAGGGGCGGAT
TTTCGGTCCCATCTTCCTCCTCTTCCTTAGTAATTATGAGCTGGAATTGTCAGGGCTTGGGAAACGTTCTGACAATTCGCA
GAGTGAAGGAGTTACATCGGACTCTCTCACCGGATATCATGTTTCTCATGGAGACAAAAAACTCCGATGAATTCATCAAAT
CAAAGCTGGACAGTATACAGTATCCCAATTATTTCTCAATCCCACCAGTGGGCTTGAGTGGCGGCCTCACTTTGCTATGGA
AGCAGGGGGTTGAGATCAAAATTATAGAGTCATGTGCCCATTTCATCGATGCTGAAGTGGTTTTCAAAGGAACTTCATCCT
TTGTAACTTTTGTTTATGGCGAGCCAGTGGCAGGGAATAGAGCGGATTTCTGGAACACTCTTACGAGAGTGGGTGCAAACA
GAGAGGAAGCATGGCTATTGACAGGTGATTTTAATGATATATTGGATAATTCGGAGAAGGAAGGAGGGCCTGAGCGATGGG
AGGGCTCCTTCACGGCATTCCGATCTTTTGTTTCACAGAACGGACTATGGAACCTTAAGCACTCCGGAAATCAACTCTCAT
GGAGAGGAGTACGATACACTCATCATATCAAGTCTAGACTGGACCGCTCCTTGGTTAATTGCTCCTGGTCAGAGTCATTCC
CAATGGGTAGGAGTGTGTACCTCAGATATGAAGGGTCTGACCACAGACCAGTGATCACTTACTTCAATACTGCAATGCAAA
AACGCAGGGGTCTTTTCAGATTCAATCGATCATATACTGAGAATGAGGAGGTTACACAGCTAGTGGATGCTGCCTGGAATC
ATCACCCTCTCGATTCAGTCATTACAAAACTCAACTCTGTCCGTCGAAGCATCATCAAATGGGCTAAAGAACAGAATGTAA
AGAGCAACCTTGTTATCCAATCTGCCCAGCAAGCGCTTGAAGAAGCACTATCAATGCCAGTTGCGGACTTACCTCATATCC
AAACTCTCACTAATACATTACTGGTTGCTTATAGAGAAGAGGAAAGTTTTTGGCTACAAAGAAGCCGGATTCAATGGCTTA
AGAAGGGGGACAGGAACACCGGCTTCTTCCATGCGGCCACTAGGAAGAGGAGAGCGCTCAATAATCTCTCAGTGATAGAAA
AAGAGAATGGCGAGGAAGTGTTTGAGGAAGCTCAGATCTCTTCGGTCATTGCAGATTACTACACGGAGATGTTCACCACAA
ACAGCAACTCTGACTTCTCTCTAGTCCAAGCCTGTCTTGTTAATAAGATTACTCCAGAGATGAATAGTCGGCTTATAGAGA
TTCCATCAGACAAGGAGATTGAGGAGGCAGCTCGGTCGATAAATGGAGGAAAAGCCCCGGGCCCGGATGGTTTCTCAGCAA
AGTTCTATCACTCTTATTGGCATATCATTGGGAAAGATGTAATCACTGATGTCAGGAGCTTCTTTGTCACCGGAATAATAC
ATCCACAG**CAGAATGAGACTCATGTTCGATTGATTCCCAAAGGCTCGGGACCGCGAAAGGTTGCGGACTATCGACCGATAG**
CTCTCTGCAACACTCATTACAAGATCATTGCCAAAATCCTTACTCGACGCCTCAAGCCTCTGCTGCCAGATCTTATCTCCA
AGACGCAATCGGCTTTTGTAGCGGGACGAGCTATCTCGGACAACGTCCTCATTACGCATGAAACCTTACATTATCTCCGAA
CTTCAGAGGCGAAAAAGTATTGCTCCATGGCGGTGAAAACCGATATGAGCAAGGCTTATGATCGCATAGAGTGGGGTTTTA
TTAGAGCAGTCCTGGCCCAGCTTGGTTTTGACCCGATATGGGTCTCTTGGATCTTGGCGTGTGTTGAATCAGTGTCGTACT
CTTTTCTGATCAATGGTTCACCTTCCGGTCATGTTACTCCATCGCGTGGAATACGGCAAGGTGATCCGCTCTCACCATACT
TGTTCATCATGTGTACAGAAGTCCTCTCAGCGATGTGTACTAAGGCGCAACTAAACGGTACTATGGCGGGGGTGAAGGTAG
CACGGAACTGTCCTCCGATTAATCATCTATTGTTCGCGGACGATACGATGTTTTTTTGCAAATCAACACCCGCCTGTGTCA
CTACTCTGAAGACGATCTTAGAGAGTTATGAAAATGTATCTGGACAACGGATCAACCTGCTGAAGTCGTCTATCACTTTCT
CGGCTAAGACACCTGGTGAGGTTAAAGCTCGAGTGAAGACTGAGCTGTCTATCCCGGCAGAGGGAGGGATTGGCAAATATT
TGGGCCTCCCCGAAAACTTTGGCCGCAAGAAGAGAGACATATTTGCGGCAATTCTTGATCGCATCAGACAGAAGGCGCATA
GCTGGACATCTCGGTTCTTATCAGGCGCCGGTAAACAAGTTCTCCTCAAATCTGTCCTTGCTGCAATACCGTGCTATACGA
TGTCATGTTTCAAGTTGCCTATCTCTCTATGCAAACAAATACAATCCCTCCTTACTCGATTTTGGTGGGATGCAAACCCGG
AAAAGAAAAAGATGTGTTGGGTCGCATGGTCTACTCTCACACTGCCCAAATATGCTGGTGGCTTAGGTTTCAAAGACATTG
AGACTTTTAATGATGCCATGCTCGCTAAAATTGGCTGGCGCCTTATACAGTTTCCAGACTCGTTGCTGGCACAAGTGCTGC
TCGGAAAATATGCTAAAGCTTCCTCTTTCATGGAGTGTGAAAGCCCATCATCAGCTTCCCACGGATGGAGAAGTATCTTGG
CAGGTCGGGAGGTTCTAAGGAAAGGACTCGGTTGGGTGGTTGGGAATGGAGAAAACATTAAGGTTTGGGGGGATCCTTGGC
TCTCCTCCTCCTTTCCTACAGCCCCTATAGGTCCACCAACTGAGAATGCAGTATCGATGACGGTTAGTGAGCTTCTATGCC
CTCTCACTAACGCCTGGGACATCCAAAAGATCCAGAACTATCTTCCTCAGTATGAAGGCATTATTAATAGGATTATCACAA
GCTCTGCCCCTGCTCGCGACTCCCTTGCCTGGCTAGCTGAGAAATCAGGAGAGTACACTGTGAAAACGGGATATGGAGTGG
AACGTGTTGGCTTGATCCCTCCTCACACAAATGAGCAAACCTTTGATTGGCTAAAGAACATATGGAACCTGAATACTTCAC
CCAAGATCAAGGATTTTCTATGGAAAGTCAAAAGAAAGGCAATCCCTGTTAGTTCCAATCTAGCGACCAGAGGTATGGCTC
CGTTTCCTTGCGCGAAATGTGGAGGAGTTGAAGATGATCTCCACGTCTTCCTTCTCTGCCCCTTTGCTTCTCAATCATGGA
GCCTACTACCAGTTTATGAAATTCCGGATGGTTCAAATAGCTCCATGGCGAGCTTATTTGCAAATGGTAAGAAGTACACCA
ATCTGCCTCCGGTGGGGTGTGATACCCCTGTCTGGCCCTGGCTCTTATGGAACCTTTGGAAAGCTCGAAATAAGTTGTGCT
TTGAAAACAAAACCTTCACGGAGTGGGAAGTTGTGAACAAAAGCGTTACTGATGCGAAGGAATGGGCTGCTGCTCAACTAC
TCGCTGAAGAACATAACCACGCTACCAAGCGACAAGGTTCTCCTCCGATCATTCCCCAACCATCTCCTGGCCAAGTTCTGT
GCCATGTGGATGCAGCTTGGGATCTGCGTACTGGTAACTGTGGCATAGGGGTTCTGTTTTCGGAACTGGAAGACACTAGGA
TCCAGCCCCTTAAGGTTTCTCGTTCTTTTGTTTCATCAGCTCTCATGGGAGAAGCTCTTGCTGTTCGTTTGGCGGTGATGA
CCGCCTCCTCGTCTAACGTCCGATCGCTGATAGTTCTTTCGGATTCCCAGGTTCTTATCAACATGATTAGAGCTAAGGAAT
CACGCCCGGAACTGTTTGGCATCTTGTTTGACATCTATCATTTTAGCTTATCTTTTGAGGATATCTCCTTTCATTTCATTC
CTCGCTTATGTAATGTTGCTGCTGATAATGTGGCTAAGGCAGCCCTGGCTTCTGTAAACTCTACCTCCCTAGTAGGAGGCT
GACTCTGTTTTCTTAATTGAATGAATCGTAGTTTGCTAC**AAAAAAAAAA**GTTACATGAATACAA

A 7313 bp *BoLINE3-1* in *Brassica oleracea* (AC240078.1) in 5′-3′ orientation

CTATGATTGGGGACTTAGCGTTGATTGGGTAGAATTAGTATTTACTGTGGATAAAGCGAACGCTCTCGATGAAACTTGCGT
TCTAGATCTTTAAATATATGATAACATCGACAAGTAATCATAGTCGTTGTATGGTCTAGTTTGGATGCGAATTCAAGTTCA
ATGTTCCAATTACCAAATACCAACTTGAAGTAACACATAATGGCCAGAGTAGGAAATTCAATATTATAAGATAAATTATTA
CTGAAATGTTGCTTTTTACTATTGTCACATTTCTCAAAGTTGATTATACACACAAGAAAATAAATTGAATGATTGTGTTGG
AAGCAAGGAAGCACTGACTAGAGAAAACAGCTTAATCAAGCACCACTAAAATCTTATTAGACGGAGTCAGCGTCATTAGAT
ACCATTAATTTATTGTTAATGACCCATCTACCAAAGTTTGATTGATAAACTAGTGCATTCTATTTTAATAACCACTGAGTC
GATTGATAATTTAAATTACAATAATTAGTGTTCTTTTGCGGTCAAAGAAAACGATTAAAGTAATTTTCTTAAAGTGATGTT
TAGACAGAAATAGTATACCTTCATCGAGAAAGAGAAATTTTCCAAATTCAGTTAGCAAACAGTTAGATTATGCAAAATATT
ATTTACCTTGTCTAACCAAAAAAAAAAACTAATAAACTTTTATAGTTACAGGATTTTTTTTCATTTTAAAAATTACATGAC
GTGAACATTTACATATTTTTTGAACTTTGTAAACATTTTCATTTCCATTTGAATTGCTTATCACACAGAAAGCAGAATGGT
ATAGCACTAGAAGAGTGTTGCTCTATTGAACTCTTTACGCCCTTTAGTTCGTTTTCTACTCATATTTTTTGTTCATGAAGA
GCGTAGAGATCATGCATGTAAAGATTTGTTTTGGTCCCAGCTAGGTTGATATACTGGTTTATACTCTGGTGTACATATGGT
CTTTTTGGTTTCTTAAATCTGTTTTTGTTGTTGGTGTAATGAAAAGACAATGACACCATTGGTGTGTGGTTCTTCTGTTAA
ACTTTATGTTCTCAGACCGAACTTTGTGATTTACGATTTTTTAGAAAGTATTCAAAATAATTTTTTTTTGAGTGCTATGTT

ATTACATTGAATTAAATAATATTGTGTAGGTCTCAAGATGTTGAATTATAGAGGAGTTAGAGCATCATTAATCCGGAGTTC
TTATGATGGAGTTCTTAGCGGAAGTTAAAAAACTGTTTCTTAACTTCCGCTTAAAACCCTATCCTAATAACTCCGGGTTAA
TCATGATCTCATGTTACGGATATGGTGAAGATCCCCACCTCCACTAGAGTGGAAAGTTCAAATGGTGCATGAATGTTTACC
CAAATACGTAGTGCGTGAAATGCATGGTTTGACCAAGTTTCTTTCTTGGTTTGTTTTAACCAAACAAAAATTAGAAAAGTC
GACGAGTGTGGGTAAACATAAACTAATAATAAATAATTATAGAACGTGTTCTAGTTAATAACACAACATGCGACTGTGAGT
GAATGATTTACATGAATACATGACGTTTTTTTTTTTTTGCTAAAAAGATATTGGTATACACAAGTAGGTGTAAAAAGATAT
TGCGGGTAACCTTGATATGATCAAAGATTCATTAAAAACATCTAGAAAAATATGTATAAATTAAAAAAATTAACTCTTTCA
AGTAATGAACTACGGTTAACAACTTTAGATGTTAGTGTTTATGGTTTCAGAGTGTTTTCAAAATTATTCTTAATCATGAAT
GCATTATTTCACTCAAACTTGGATCAAATAACACGTTAAACCTCTTTCAAAAAAAGCTCCGTCCTTTCTCCTCTTCTCTGT
CTCGCCTTGTCACTTAGCCATGCCTCCGAAAAAGAAGAAAAAAACTCGTAATCTCCTCCGGGGCTCTTCGAAGATGGCCCG
TCTCCAAGGAGCGTTGAAACCCTCCACTTCGGTCCGGTCGTTACAACAGTCGTCTCTCGCCGGCGCTTCAAGCGCTGATGC
TGTAATTGCTGCTTCGGTGGCACCCACGACTGAGGTTTCGCCGTCGACCCCTGAGGTCTCTGTAGATCTGAGTTTGGGTCA
CCTACCACTACCGCTAGCTCCTGTTCTTGAGTTTGGCAATTCATCACCTCTCACGCCATGTGAGCTTGACGCTACCCCTAC
AGTATCGGAGGAGGTAACTCCCGACGCTACAATAGAGGTTTCCCTTCCAATCCAGGAAGTGGTTAGTAGAGACATCCCTCA
TGGAGCCCTGCAGAAGGCTGCTACCCTAGGGGAGTACTCTGGAACTAAGAGCTACGCCAGTTTAGTCAAGGACTCTGTGAC
CTTGGAAGAGCTAGGAACACCCTCTGAACACGTATCGGGGGTACCTTTCGTTCTTATCCCGGATGAAAATATAGAATCAGC
TAAGGAGGAGTTTCGTGATTACATCTTCGCTAGATTCCACGGGGACTGGCCTTCCATGGGTCGTATAATCGGTGTTGTTAA
TGCCGTATGGGCTAAGACAGGGCCGCGAATCTTTGTCCACATGGTGGGCGCTGGGGAGTATCTATTAAAAGTCACGTCTGC
TAAGACAAGGGAGCACCTCCTCTCCAGGACTTGTTGGAACATTGCGGGATTCCCGATGTTTGTTGCTCCATGGTCGCATGA
TTTTACTCCGGAGGAAGCTCCGATTACAAGTGCGGTGGTTCCAGTAGAGCTAAGGGGAGTCCCTTACTTGCTGTTCAACAA
GGAGAGTTTGAGTCGGCTGGCCACAGCAGTAGGCAAACCTGTGTCTTTGGCTCCGGAAACAGAGCGGAAGGAAAACTTTCA
GGTTGCAAAGTTATTTGTGCGTGTGGATCTCACACGGGAGCACCCCTCAAAGATGATCTCTGGGTTTTCAAATGGGAAAGA
AAATGAGATAACTATTTCTTACCCTTGGCTCCCACTAAAGTGTAATGCTTGTGGTAAATATGGTCATCTCAACACCAAGTG
TCGTGCCCTGCCTCGTAGTAACACGGAAGGCAGAAGACGTTCTCCTAGTCCAACGAATGAGGAGGATAAGGGGAGGAAACA
GTCTAGGCAAGGACGTCGTAGCAGAGGAGGTAAAGCTGGTACTCACAACAAGGAGCGATCGGTAGATGGTGATGCGAAGAA
GGGTGTCACATCCTCTCAGGGTTTGGAAGATGGTGAAATACCTCCTGAGGAGCATACAGAGGACACCACTGTAACAACTCC
GGTTAGGGAGAATGGCATTCCTGAATCATCTGATAAGACTGTCCCACCGGATATCTCTCTTCAGAAATTCCCACCTGCGCA
CGACTTGATTACTTCTTCGTATGAGTCAGGTCCATCCTCGGGGGTCCCTAGTGCTGAGGCAGACGGCTCAGATGAGCATGA
AGCCCCGTTTTTGCTGGTTAACCGCCAGAGCTGTGGCCGCAGGGTCGCAAAATCTATATAATTATAAATGTCGAAATTTTT
TGCATGGAATGTGAAAGGGCTTAACGATCCCAGACGCCACACCATGGTTAGAAACTGGATTAATATCCAGAGGCCGCTCTT
TGGAGCTTTTTTTGGAAACACATATTCAGGAAAGAAATTCACAACGGATTAATAATGCTCTCCCTATAGGATATGGAGCTTT
TTCGGCTACTATGATCATCATCTCTCGGGTCGAATAATTGTTGTTTGGGACCCGTCTGTTCGAGTCTTCATCTACAAGTCT
TCTGCTCAGGTTGTTACTTGTGGGATTTATCTTATGGCGGAAAATGTGAACTTCACGGTCTCTTTTGTTTATGGGTTCAAC
ACTGTACTCGAGCGTAAAAACTTATGGGAAGAGATGGTCTACATACATGATTCCACTCCGGTTGTTAACTCTCCTTGGTCG
GTGCTGGGAGACTTCAACCAAATCTTTCGTCTAAGCCAGCACTCTGATTACCCTCTCTCAGTTATTGATCCATCAGGTATT
GATGATATGGTTGCAGCTTTGTAGGACTCTGAGCTCTTCGAATGTCAAGCGAAAGGCATCCCTTTCACTTGGTGGAATAAC
AGTGGCTCCAACCCTGTCTCCAAAAGAATTGACCATGCTCTTATCAACCACTCGTGGGCCGCTTCATTTCCGGATTCATAC
GCAGACTTTCTTCAGCCAGATCAGTCTGATCATGCGCCTTGCCTTCTCAGGGTTCCCTCAATCAGTAGACGTATCCGCAAG
CCTTTCAAATTTTATCACCATCTTACTGGCCATCCTGACTACTCTTCTGTTGTCTCAGACGCGTGGTCAAATGCTGAAGTT
CAAGAATCTGAGCAGTTTAAATTGGTCCGACGAATGAAGCTTCTTAAGACAGATCTGAGGAACTTAAATAAAACGCACTTT
AGTGGCATCACAGGAAGGGTAAAGCAGCAGTCTGTGAGGGTAGCGAACCTGCAGCAAAGTCTTCTCACCTGCCCGGACCCA
GCTACTGCCTCTGAGGAGCACCGTCAGCGTGATATCCTTAACACCTTACTCAATGCTGAGCAAAAGTTCTTCAGGCAGAGG
TCGCGTGTTCGCTGGGCTGATGTGGGAGACAGGAATACACCTTTCTTTCATAAAACAGTAGCACAAAGGAATAATAGCAAC
CACATCCATTATCTGATAGATAAGTCTGGCCAGTTCCTGGGGTCCATCGATGATATAAAAGCGCACTCAGTCTCTTATTTC
CAGGATATCCTTGGCCACACAGAGCTTCCTGTCTCACCAGTTCCTCTAAATGCTTTGGAAGGTTTGCTCTCCTTCAGATGC
TCGGACCTACAGAAAGCTTACCTGAAGAGAGATGTCCTGGAAACTGAGATCAAAGGTACTATCTTCTCCATGCCCCTAAAC
AAAAGCCCAGGTCCCGATGGCTATTCTTTTGAGTTCCTTAAAGCATCGTGGGATACTGTTGGAGGGGATGTGATTGTTGCG
GTTTCTGAGTTTTTTCGTAATGGCCGATTGCTG**AAAGACCTAAACACCACAGCTATTGTTCTCATTCCTAAGACCACTACT**
GCTTGTAGCCTTTGGGACTACCGGCCGATAAGCTGCTGCAACATTGTGTATAAGATCATCACCAAGATAATCGCCAACAGG
CTCAAGCCGATTCTCAAGAGTTCGATAAGTCGTGCTCAGTCGGCTTTTCTTAAAGGGCGCAGCTTGGGTGAAAACGTCCTT
CTTGCAGCTGAGCTGATCCGCAAGTATGAGAACCAAAATTGCAGCAGGAGCAGCATGCTAAAAATAGACATCCGCAAAGCT
TTTGATACAATCTGCTGGGACTTTGTCATTAAGATCCTCCAGGCTCAGGGATTCCCTCCGATTTTTGTTACCTGGATCAGA
GAATGCATCTCGTCACCCAGGTTCTCTGTGGCCATTAATGGTGAGTTTGCAAGTTTCTTCCCGGGGAAAAAGGGGCTGCGT
CAGGGTGATGCAATCTCGCCGTACCTCTTCATTATGGTGATGGAAGTACTGTCAAAACTGATTGAACGAGCTGCTGCTGCT
GGACATTTTCGCCTCCATCCTCGGTGTTCTGAGCCTATAGTCACGCATCTACTTTTTGCTGATGATCTTCTTGTGTTCACT
GATGGGTCAAGGCACTCTATCTCCGGTGTTAAGAACGTGATGGCAGGGTTTAAAGACTGGACAGGTCTGGATATAAATGCT
GAAAAATCGGAAATTTTCTTTGGTGGGTATCTAGATATTGAGGCTGCTGTCATCAGTGACATCTCAGGTTTCAAGCGTGGT
AAATTCCCAACTCGGTATCTAGGCCTACCACTATGCCCTAAGAAGATCAGTTTTGCGACGCTGCAGCCTTTTCTTGAGCGA
ATAACTGCCAAGCTAAACAATTGGACAGTTAAATGCCTCTCCTATGCCGGCAGGATCACTATGATCTCATCCGTCATTTAT
GGAATGGTAAATTTTTGGAGCTCGGTCTTCACACTGCCAAAGAGGTTCTATGCTAAGGTTGACTCCGTCTGTGGGTCCTTC
CTCTGGAAAAATAAGACTACATCGGCTTCGGGGGCTCGAGTTAGCTGGGACGATATATGTAAACCGAAGAATGAGGGCGGA
TTAGGAATCAGGAAGTTGGAAGATTTCCAATCAGTCTTCCAGCTGAAGAGGGTTTGGAATTTTTTCTCTGAGGCAGGCTCT
TTATGGGTTGCTTGGCTCCAGCATAATGTCTTCGCGGGGTCATCTTTTTGGACTGCTGAAACTTCTACTACTTTCTCATCT
ACGGTCAACCAAATGCTGAAGCTGAAGCCGAAGCTAAATGCCTTAATGAGATGCAATTTAGGGGATGGGAAGTCTGTTAGC
TTCTGGTTTGACTGGTGGACTGATCTTGGTCCCCTAATCTCTGTGTGTGGCAGGAGAGGTACTAGGGACCTTCGTATTCCC
ATTGACGCTACTGTTAGTGCTGCTGCACCTAATGGACACTGGTCTCTTCCACCGGCGAGATCCGACGAAGCTGAAACCTTG
CAGGTGGTTCTCTCAACGATGCAGCCACCATCATCTTTGCGCGGAAAGGATTGTTTCCTTTGGCGGAGTGGTGCTCACTCT
TTTCTGCCCAAGTTCTCTTCTAAAGCTACTTGGAACTTTGTAAGACAACCCTCCCCTCTGGTATCGTGGTATAAACTTATC
TGGTTCAAAGAAGAGATTCCTCGCTGTTCTTTTGTCTCTTGGATGTCAGTTTTATCAAGGCTGCCTACTAGAGATAGGCTC

ATCTCATGGGGAATGCAGGTTCCTCCGGTGTGTCCCCTATGTTCTCTGGACCATGAATCGCACGCTCATATGTTCTTCTCC
TGCTGCTTCTCAACTGCAGTTTGGTCTCACTTCACGGGCTGGTTGCTCCCTACCACTCCGGATTCACTGCCTCTTTGGAAA
CCACCATTGCTCTGCCACATATTGCTTCCTGCTCGGGTGTTGTGTCTGTTATTAAACTTTTGATGCAGGTCATCGTCTACA
GCCTGTGGAGAGAGAGGAACGTCCGTATCTTTAAGAATGTTTCCTCTACCGCAGCGGCGATCGCCAAGTATGTTGATCGTT
CTATGAGGGATCGTCTCTTGTCCTTAGCCCCTCCAGCTGATGGCCACGACACGCTCCTCCTGCTCTACCTTTCCATCAGAC
GTCTCTTTCCACTTTAGTGTCGCTTTCTCTTATTTTGATTGTTGTTGTTTAGTCTGTTTCCTTTGTCTTTCTTTTCTGCTT
CTCTGTTCTTCTCTTGTAATAAGTTGTTTAACAACAACAGTGTATCCTTCTTTCAGAAAATGGTAAGCTTAACATTTTACC
AAAAAAAAAACTATGATTGGGTA

A 6560 bp *BoLINE4-2* from *Brassica oleracea* (AC240089.1) in 5′-3′ orientation

AAATTGTTTCGACCTTGTGGTTATTCAACAAGTTAAAAAAGAATCGTTAGAGAAGGAAATCAAAGTAACTTCTCTCCCAGA
CGACCTTTTCGAGAGCAGTCTTTATCCAATAGACCCAAGCACATTCCTGTCCAGCAACCTGCTTTAGTCCAAAAGTCAACA
GAGTTTCTTTTCGAATTTCTTTCTTCTTTTGGAAAAAACCATTGGATATCTCCACGCTCTGGTCTATTACACATCGGGTTC
TTCCCCCGACCATGGTCGGAGGGACCGCCGGAGCCGCCTTCACGCCGCTTAACTTGCCTGCTGGAGTTTCCCAACCGGTCG
TTTCAACCCAAAGCCACGCCTTTCCCCGCAGCGGAACCTGTTTCCTCTTCAGGGCCCAACTTCGCTGAAGATCTCCACTAC
ACCGAGGGTTTCCTCAGAACAGAGACGTGAACACATAAGCCTTTCAAGGAGCTCACAAATATGCTTAAGGGTTTCCTTATA
TATTTTTCTATCACAAACTAGAAGAAGATCTGTCACTTAAATCTCAACCACCAAGAAGTTATTGCTCACAAATTGACTCCA
CCAGCAACTACTTAAAGCCTTCAAAGCGGCTTCTCTAAGGAACTAACTGACACAGGGGTTCCTGTTACTACAATAGTTTCT
TCGTGCTTTACACCTCACAACAAAAGATGTCTGGAAGACTTAGGAGAGCGGATAAAGGTAAATAAATAGCTACTGACCCCT
CTCAAGCCCCAAGAACTGCTCGCATCAGAATCCAAGACCCGGATAATGCTGAGCTCATGCAACGCCACTCTCTAACACTCA
TAGGCCGTGTTACCAACAGAACGGCGCAGCGAGTCTGGTCTCTCATCCCCTTTTTCACCGATCTCTGGAAGGCAAAATCCA
AACCGGTCGGCTCTGACCTAGGAAACGGAATGTTCCAGTTTCAATTTGACAGTGAGGAGGATCTACTAACGGTCTTGGAAA
AAAGACCCTACTATTATGGAAGATGGATGGTTATTGTTCAGAGATGGGAACCCACGGTCTCCAAAAATTTCCCATCCCTCT
TGCCTTTTTGGATTAAAGTTCAAGGAATTCCTGTACATCTATGGACAGAGGAGACAATCCAGAAACTTGGCGAAGATCTGG
GTGTTTTTGAAAAGATGGAGATCACTTCATCAACGGTGAGAATGAGGGTGCAGGTTAATGGACTGCTCCCGCTAATCAAAT
CATCGGTGATTGAGTATGCGAATGGAGATGAAGTTACAGCGAGCTTCGTATATGAAAAGCTCGATAGGCATTGTCCTAAAT
GTTTCCGCTTGGATCATGACTTAAAAGATTGCTTGGAGGCCAAGCATGAAGCAAGAGCTCTTAAAGCTCAGGAAGCTAGCG
TGGGAAATGAGGAGCCTGAGAGAAGGGGAACGCACTACAAGCAACACGACTCAGGTTCTAATATTTTCCATTTCACAGCTC
AAGGGTCAGAGCACAGAGGAAGACAAGACTATAACCGGCGAGATCGCCAAGTTGATGCGCGAGATGAACTAGAGGCCCGAC
GCAGGTCTCGCTCCAGCCAAGATACAGTCCCTCGAAGATACTTTAGTGAGGACCCCAAGCGGGGAAGAGAGGACTATCGTA
GCCAAGATAGTCGGTCCTCCTATCACCGTGACACAAACCTGCTCCTGCGTGAGGTTTCTTCTAGGCCACGGGACCTGAGAA
GGGACATTTCGGATAAATACTTATCCAGGGACCGCAGCCTGCAACCAACAAGATCCAGAGATAGATCAATACCAGGAAGAG
AGGAATCTCCTCCACGGCAGTCAAGAATCAATCCAACTAGGGGGATTCCCCTGGAAGAAGTTCAAGCCTCTGTACCAGTGG
AAGTCTTTAATGCAGCAGTGGGAGAAGTCAGGGAGGCCAAGATACAATACACCCAATGCAACGATCCAACGGAGAGCGCGG
CTCGCAAGGAACGTATGAGAAGAGCAGAGGAAGAAGGTGAAATGGAGGAAACAACAGCACTTATGCTTCAGGCAACTTTGA
TAGCACCTACAGATACCTTGCAAAGCCCGGAACAGCAACCCACAGCAGAAAGAGTACCAGCTGCTCTCCGGCTAGGGCCTA
CTGTCCAACATTTACAAGGATCAGGCCAAGATGCATCCAAGGAAACAGGCAAGAGAAAGCCTGGAAGACCACCAGGAAGAC
GTACGGTGCAAGGAAGCCCAAAGTTAATCAGAGGATCGACTTCCAAAAAGAGAAAACTCCCACAAGATAAACCACCTCTCA
CCAGAAGAAAACTTATTCCGGAAACGGATCAAAGAAACCCGCAGAAGGCGAAGTCAAAACCCTCCTCATCACGAGGCTCCC
GAGGAGCCCGAGGAGCAAACGGAAAAGCAGAGGTGCAAACAAACTCGGAGGATCAGCCCATCTGCAATTTGATTCCGGCGT
CATCAAGAAGAAGAAAGATGGATTTTCAAAATCCATCTTCTGTCGTTCCTTAAAAGTTGCGAGTTGGAACTGTCAGGGTTT
GGGGAATCCCCGGACAGTTCGACGCCTAAAGGAGATGAAGAGAAACATTTCTCCTGACATTCTATTCCTTATGGAAACTAA
AAACCCCGACAGCTTTGTTGCAAAGAAGACGGACAAGCTGCAATATGAAAACAGAGTTCTAGTTTCACCTGTGGGACACGG
AGCTGGAGGATTGGCGCTATTGTGGAAGCAAGAAATTAACCTTTAAGTTCTCTCTACTTCTACAAACTGTATTGATACTTG
TATTATCTTTGAAGGGAAAATTTTTTTCGCTTCCTTTGTCTACGGCGACACTAACAGACCTCAACGAAAAGAACTATGGGA
TCAGTTGATTGATCTGAACACGGCCCGAGAAGCTCCCTGGTTTTTAACCGGCGATTTCAACGACCTACTTAGAAACGCAGA
GAAAGATGGTGGGGCAATTCGCCATGAAGGATCCTTCACAGATTTGAGAACCTTCTTTTCCGAAGGTGATTTATTTGACCT
CCAATACTCGGGAGATTTCTTGTCATGGAGAGGGAAGCGTGGTGATGATCTGGTTCGCTGTAGGCTAGATCGTGCAGTTGC
GAATAGTGACTGGGCTGAGTTGTTTCCGACTGCAAGTTCCCAATACCTTGCCTTTGAAGGATCTGACCACAAACCCTTATT
ATCCTTTTTTGAGCCAGAAAAGAAGAAGAAACGTGGAATGTTCCGCTATGATCGACGGCTTAAGAACAATCCAGAAGTTAA
AGAACTAGTGGCCAAGACGTGGAAGAATAGATCATTTAGAACTGTCAATGACAGGATCTTTGCTATACGATCAGTCCTAAT
CGAATGGAGCAGACAGCAAACTCTAAACAGTAGAGCCCGTATAGAGGAAAAGAAGCACCAGCTGGAACAAGCTCTAACGGA
CCCTGTAAACGACACAGAGCTGATAGCGAAAGTATCGAAGGAACTCGACGAAGCCTATGCAGCAGAGGAGAGTTACTGGCA
ACAAAGGAGTAGGCAGCTGTGGCTCAGCCTAGGAGACAGAAACACCGGATACTTTCACGCAGTGTCAAAGAACCGGAAACG
AGTTAATGCCCTCTCGGTCATTGAGAAGGCAGAAGGGGAAGCAGTCTACCAGGAGGACCAAATTGGGAGGGTTATTGTCGA
GTACTTCCAGCGACTATTTACCTCTATGGGTGGAGACAGAGAAGATACAGTGCACTACGCCCTCTCCCCAATGATCTCGGC
AGAGACAAATGAAGAGCTAATCCGCATACCATCTGCATTGGAGATTAAAGAGGCAGCATTCTCTGTTCACGTAGACAAAGC
GCCGGGGCCTGACGGTTTCTCGGCCAGCTTCTTCCACACAAACTGGGAAAATATAGGAGCAGATATAGTCAAGGAAATCCA
GGAGTTCTTTGTGTCGGATAAGCTGCCTGACAAGATCAATGAAACCCATATCCGGCTCATCCCGAAGATTCAAAGCCCAAA
GACAGTT**GCGGAATACAGGCCCATCGCTCTCTGCAATGTCTACTACAAGATCATCTCCAAGATCTTGACCAAGAGACTGCA**
GCCACTGTTATTGAACATCATCTCGGAGAACCAGTCCGCCTTTGTCCCGGGCCGGGCAATATCGGACAATGTCTTCATCAC
TCATGAAGTCCTTCACTATCTCAAGACGTCCAAGGCTGAGAAGAGAGTATCTATGGCGGTTAAAACCGACATGAGCAAAGC
CTACGATAGGCTCGAATGGGACTTCATCAGATTGGTATTTCAACGCCTCGGTTTCCACCCGAAGTGGATCAACTGGATTAT
GCAGTGTGTCTCTACTGTTACTTACTCCTTCCTCATTAACGGCTCGCCTAGAGGAAGAGTCACACCGAGTAGAGGTATCCG
TCAAGGAGACCCTCTCTCACCATACATCTTTATCTTGTGTAGTGAGGTCCTCTCGGGTCTATGTAACAAAGCGCAAGAGGA
AGGAACCCTTAAAGGGGTCCGTGTAGCACGAGGGTGTCCTCGTCTCAACCATCTCCTCTTCGCCGACGACACAATGTTTTT
CCTTAGAGCAAGCAAGGAAAATGGCGAGGCTTTATGCCGATTACTGAAACGATATGAGGAGGCTTCCGGGCAGTCAATCAA
CACAGAAAAATCTTCGATTAACTTCTCTCGGCACGCACCGACAGCTCTTAAATCAGCTATCAAGGATGCTCTTTCTATCCA

GAAAGAGGGAGGTACCGGCAAATACCTCGGACTCCCAGAGCTATTTGGAAGGAAGAAGCAAGATCTTTTCTCATCTATTGT
GGATCGGATAAAGCAAAAAGCATGTGGCTGGTCAAACAGGTTCTTATCTACTGCGGGAAAGATGACTATGCTCACCAGTGT
ATTATCTCCCATTCCATCGCACGCCATGTCCTGTTTCCAGTTACCGATATCTCTGTGCAAGAGAATTCAATCAACGTTGAC
TCGATTCTGGTGGGACACAAACATAGGAGATAAGAAGATGGCTTGGATTGCTTGGTCTAAGCTGGTGCAACCGAAAGATAG
TGGAGGTTTAAACTTCAGAGATATACAGAGTTTTAATGAAGCTTTCCTCGCAAAGCTGAGCTGGAGAATCATCAACCATCC
CGACAGCCTACTTGGAAGAGTTCTACTTGGAAAATATTGCAGTGAGGAGAGCTTCTTAGAATGCTCAGGAAAAACTGCTGT
CTCGCATGGCTAGCGTGGAGTTTTGATTGGACGCGATATCATTGTCAACTCCGCATGATGGGAAGTAGGCAATGGCTCTAG
CATCAACATCTGGGAAAAACCATGGCTCAGCTGCTCCACACAGCTGAGACCCATGGGACCACCACCGAGAGAATTCTCACA
TCTCACAGTGTCTGACCTCATGCTCCCGGACCGAAATGAATGGGATATTGGGATGATCCAGCGTGTGCTCCCATTTGAGGA
ACAGAGAATTTTAGCCATTAAGCCAAGCTTAACCGGGGCTCCGGACAATCTCTCATGGCTGAGCACTGACACAGGAGATTA
CTCTACCAAAACCGGCTATGCTGCAGTCCTCTCTTCTCGAACCGCAGAGGATACAGTAAGCACGGAGGATGCATCCTACGA
CTGGAAGAAGAATGTTTGGAAAATCCAAACAGCACCAAAGATCAAGCTGTTCATCTGGAAAGCGCTTCATGGGGCGCTACC
TGTGGGCGAAGCTCTCAAAGCCCGGGGAATCAACACCAATGGACAGTGCAAACGATGTAATCTACCTGAATCTATTGATCA
TCTGCTTTTTCATTGCGACTTTGCTAGACAGGTTTGGGAATCAGCCCCTGTCTCCCCAAGCATTGAATACAGTGGATCCAT
AGATTTAAGGAGCAACTGGAGCAGCTTCTGCTCAAGGAAAAACTTGCCACCGACTGGAATTTCTACAGGAGCCCTTGCCCC
TTGGATCATTTGGCAACTATGGCTAGCTCGTAACAAGTTAGTTTTTGAAGGAAAGATCATCACGGTGGAGGAAACAATATC
CAGAGCAACCGCTTGCGCCCGAGAATGGATATCTAGCCAAATTCAGCTGTCTGCAACAAAACAAGCACTACCACCCAGACC
CCCGCTTCATGACTGCGTTTTGGTGAGAACCGATGCTGTCTGGAACGAGAATTTAAAAATTGCAGGGCTAGGATGGACAAC
TAATCGAGAGGAGAGTCTCCCAATCCTCGACCACTGCGCAGCATGTTGAGTCCCCTCTTGCAGCAGAAGGACTGGCGATGA
GGGAAGCCCTCCTAAAGTGCAGAGAAATTGGCCTCCCCAAGCTTAGATGTGAGTCGGATTCGGCGATCCTGATCAAAGCTA
TTAACTTGCGCTCCCCACTAGTTGGTTTGTATGGCATTTTAGCTGACATTTATTCTATTGCCTCTTCTTTTGAATCCATCT
CGTTCACCTGGATCTCGCGTGAGAGAAACAGTGTAGCGGATGGACTGGCGAAGAATGTTTTATCTTCGGAGCTGGCCATTA
TGGCAGCAACAAACCCTGTTTCGAACGTTTGAATTAATATAAGTAGTGTTTCAAAAAAAAAAAAAATTGTTTCGACCTTG

A 690 bp *BoNA-LINE1* from *Brassica oleracea* (EU642504.1) in 5′-3′ orientation

GCTTATAGGGCTGGGCATGAATACTCGTTACTTACCTTATATCCGCTACTTGAACCGTACTCGCCTTGTTTTTTCAAATAC
TTGTCGTTGCCAAGTCAAATATCAAATCGTTGAGTTTTGCTACTTACAAGTACAAGACAAATAATAAGTATCACAAAAAAA
GTTACGACTCGCCTTGATATTACTCGTTACTTGTTTATGACTAAAAATGATAACTAAACCATAAAAACCAAAGCCCTAAAT
AAATATTTCTTTGNTGTAAGAAGTAAACAATACTTTCTTATTGAATTCCATCTTTCTTTTTTTTATATTAGGGCAAATATT
TTATTTAAAATACGTCTCGTGTTTTTACCAAGTCAAGTAAATAAAACAAACTTTTATAACCTTATCTCGTAAAATATTATC
TATCAAGTAATTTTTAGAAACTTGAAAATATATCCAAGTAATTAATAAATATTTGTAAGTAACAAGTAATCGAACAAGACA
ACTAACTTGCAAGTAACATATTTTTGGATAAGTACTCGAAATAACAAATACTCGTTTCTGATAAGTCAAGTATAAGTCGAA
AAATTCATTACTCGTTCAAAGCAAACCAAGAAATATACGTAAAAAAGCAAGTATCCGCCTTGCGTCCACCAGCTTCTACAT
CGTAATATTTAGACCCAAAAAGAAAAAAAAAAAAGCTTCTA

A 914 bp *BrNA-LINE4* in *Brassica rapa* (AC189298.1) in 5′-3′ orientation

AAACTACTTTTTGGATGTTCTGGGGGAGAGACGTTTAATAGATTTGGGCATTGGAAGAGATGCCACCGTGGGGGATGCTCT
TCAGACTTCTCGAAGAAGAACGAGGCATCGCTCAAGTTTGCTGAAAGATATCGAAAGGGCTCTGGACTCCTTTAGAGATCA
TCTAAACACAGCTGAAGATGATGTTGATATGTGGAGACGTTCATCTGGATATAAACCAAAATTCTATACTAGTGAGACTTG
GTTACTTCTTTGTGAGGCGAAGCCGCAGTGCCTATGGTCCAAAGGTATATGGTTCTCGCAGGCTACTCCTAAGTTTGCTTT
CATGGCTTGGCTTGCAACTAGAGGAAGAATATCTACAATGGATCGAGTCTCTCAATGGAGTCAAGGCATTGACACAATCTG
TGTATTATGCAAGAACACTCCTGAATCCCGGAACCATCTTTTCTTTGAGTGCTCTTTCTCAAGGCAAGTGTGGGAGCACCT
AGTAAAAGGCATTTTGCAGAACTCATACACTACAGAATGGTCAGGTATTGCGGAATTGCTTATTGATCAGCATTTGGAGAA
ATACAGGCTGTTCTGCACTCGCTATGCTTTCCAAGCCGCTGTCTATGCGATATGGAGAGAACGCAACAGACGTCGACATGG
AGAATCACCAATACCCCCCAAGCTCTGATGAAGATGGTCGATAAAGGCATGAGAAACAAACTTAGCTTGATGAAACTACAT
GGAGGAAAGGGTTTTGAAGGCATTTTGCAGTTTTGGTTTGCCACAAGAGTGTAAATAAAGTGCAGATTAGGAGTAGTAAAG
AGTAGTTGATTTCCATTTATAAGCATATGATGCACTTGTTGTAAAGAGATCTTTTTGATGAATAAAATTTAACATTCTTTC
AAAAAAAAAAAAACTACTTTTTG

***BoSINE1-1*: A 216 bp SINE in *Brassica oleracea* (EU642504.1)**

AAATCAGAACAACCGGGCCTCGTGGTCTGGTGGTATAGGAACCTCAGCTGAGGTGCCCGCCATCACGAGTTCGAGTCCCGG
CCACAGCGGATTTAACATGGTTTCCGTTTGGCCGCCAGGACCCTCCAGTTCCGGTTGGACGCGGTGGGATAGTCGGACTAA
GGTCCGGATACCTGGATTATCAAAAAAAAAAAAAAAAAAAAAATCAGAACCAACC

***BoSINE2-1*: A 219 bp SINE in *Brassica oleracea* (EU642504.1)**

AAGAATATAAAAGAATAAACCAGTGGCCTTGGGCTAGTGGTACTTGAGGGGAAAACCCTCAATACGGTACCCGCAGTTCGA
GTTCCGCTGGCCACCCGGACTAGGGTTAAATCCCAAGAATACGTGGAGTGCCGTGGGCCTCGCGGGAATAGTCGGTTGACC
ACGGTCGCCGGAAACCCGCGGTTACCTTAAAAAAAAAAGAATATAAAAGAATAATAA

***BoSINE3-1*: 272 bp SINE in *Brassica oleracea* (EU642504.1)**

GATCCAATGATTTAACCACATCGTGGTGGTCTAGTGGTTTCCACTAGAGGAATAGTTGCCCTGTTGGTCCAGGTCAGGGAT
CGATTCCCCTTTAGTGCGAAATTATTAGCTCCACATGTGGCCACGCGGATATGGGTCCATGTTTACGGCCCATTTAAATAC
CCGGGAGAGGATCCATCCGTGGGTTGCACCTCCCACCCGGGAGTTAGGTCTGTGTTTTTAATAGACCCGGGTTTAACCTTT
TTTCGTCAAAAAAAAAGATCCAATGATTT

***BoSINE4-1*: 446 bp SINE in *Brassica oleracea* (EU579455.1)**

TTTGCTTCATCTCAAGTGTTGCATGCGGCATCTGCATGTCCAGTAACTTAAATTTGCTTCATCTCAAAAAGCTTGCAGTAC
AAGCGGTTGTTTTTCACCTTTGGAAACAGAGGAACAAGTTAATTCACAACAATGTCTTTGTGCCTCCTTCAATTGTTTTCA
AAGCCATTGATAGGGAGCTGAGGAACATCATCTCCTCTAGGAGATCCATCAAGCAATTCAAACGTCTGATGGTTTTGTGGA
TTCGATAAGCTGTTGTGGTTTAAGCTTTCCAATGGTTTGGATGTTGTTGGTCCATCTTGTTTTTATTCTTTTTTGCCAAGA
TGTATCAAACCTAAGATCTTGTCTTTGTAAAATCTTTTCTTCATTTATATGATATTTACAATTTAG**CAAAAAAAATTTGCT**
TCATCTCAAGTGTTGCATGCGTCATCTGCATGTCCAGT

***BoSINE5-1*: 225 bp SINE in *Brassica oleracea* (AC240089.1)**

AGTTTTGTGGTAGTGGATAGGCTTTAGATACAATCACATTCTGGATTCATCTTCCTACCGCGGAAACTACGGAAACTAATT
CACATTATTCTATTCACCACGGCAGCGGATACATCCATATGAAAACTGGGAAAGAGTGTCGTAGACTACACTTTCTACCCG
AAAGTTAAGTTTGTGTCTTTAATATATTCGGGTTTAACCCTTTCCTTTTT**CAAAAAAAAAGTT**

***BoSINE6-1*: 335 bp SINE in *Brassica oleracea* (AC240089.1)**

CTACTGACGCGACTTAATCCCTTCTCTACAACATTTGGAGGCAGCGAAACTCAGCCGTGCATCTTTCACGATTCACTCCTT
CTCACACGACCTTCAACAGCATAGACAGGGACATCCGCAACTCAATCAGTGCGAGAAGACATAGGAGGAAGTTCATTCCTC
TTATGCAACTTTGGATTAGATAAACTTGTTTCTCTAATCTAACTCTCTATGCATACCTTGTTTTTTCTTTTTTTTTGCCAA
GGTATCACCTATTAGGTTTTTGTAACCATAAACTTTTCATTATATGAATATTAACAGTTTAG**CAAAAAAAAAAAAAAAAAA**
CTATTGACGCG

***BoSINE7-1*: 401 bp SINE in *Brassica oleracea* (AC240089.1)**

AATACTCAGGAGAAAGGTATCAGGAGGCTGGAGGTGGAAGTTGATTCCAAAATGGCGGTGGGTTTTCTTACGACAGGGATT
AAAGATATTCATCCACTGTCTTTCCTGGTACGATTGTGCCATGGCTTCTTGACGAGGGACTGGTTGGTCCGAATTGTTCAC
GTGTATAGGAAAGCTAACCGATTAGCTGATGGGTTAGCTAACCTTGCGTTTTCTCTTCCTTTTGGGTTTCATCGTCTTGAT
GTTGTTCCGCGTGGGTTTGTTAGTTTGATACGAGAGAATGTTGATGGACTTGATGGACTGTTACAGCCACAACAAGTTTGA
TTGTAATTTGAAGTTTTAGTTTAAATAAAATTGGGAGGCTTCCTCCCATCTTCCTAC**CAAAAAAAAAAAAATACTCA**

***BoSINE8-1*: 484 bp SINE in *Brassica oleracea* (AC240089.1)**

GATAATATCATAATTCGGGTCATGGAAAGATAATGATATATTTTCAGGACAGGGTTGGTATAGTACCTTACCGGGTTTCGA
TGGCTTATTAGGGTCACAGAATGTAAGGGCATGTATTTCACCACTACATTCAGAGATAGATGCGCTGATTTGGACAATGGA
AAGTATGAGGAATTTAAGACAGTTTCAGGTTACGTTTGCAACGGATTGTTCTCAATTGGTGAAGATAGTTTCAAAACCAGA
AGAATGACCAGCATTTGAAAGTTATCTGAAAGATATTAAAGTCCTACAAAGGAGTTTCTACAACTCAGAGATCGTTCATGT
ACCTCAGATGGAGAATAAAAGGGCGGATAGCTTAGCACCCAGTGCTAAGAAACAACCGTCTTTCGTCGTTCATATAGATAC
AGAGTTGCCGGTTTGGTTTACAGAGTCAAGTTGAGTCTGTGAATGTATTGTTGT**CAAAAAAAAAAAGATAATATCATAA**

***BoSINE9-1*: 524 bp SINE in *Brassica oleracea* (EU642504.1)**

ATCAAAGATGTTCAAAGGCGGAGATGATCTCGTGTTTCGATTGATCTCTTTCGCGTGTTGACAGTGCGGGTAGGATGATTT
CTCTAGTGGTTGGAAGTTTGGTTGTGTCGGGGGAGATCTCGATGGTTTGAGAAAGACCTCCGACATGATGTAAAAGCAAGG
AAAATAGGTTGTTTCGGCGGAGGCTCTGCGGGTATCTGAGCTCCGGCGAAACGAGCGGTCCGGCGATCTCTTCCGGTTTCT
TCCGGAGTCGATGAAGGGCAAGTTGAGGCAGAGAAACGCATGCCGGTCTGAATGGTGTAGATCATCCCACGTGTCTCTTCT
TCTGCTTCCTTGTTTCTCTCATGGGCCTTGATCGTCCGGTTTTGTAAGGTTGGGCCGAGAGGCCGTCTAAGTTGTTTGGGG
CTTCGGCCGTATGTTGTAGCCCGTTTATGTTTGGACCTTTGGTGTTTCTGTTGTATCTGGGCTTGGTCCATTTGCTTTTAA
TAATATAAACCTGAT**GGAAAAAAAAAAATCAAAGATGT**

***BrSINE10-1*: 376 bp SINE in *Brassica rapa* (AC189298.1)**

GTAGGGCCTATTTGTGGAAGCAGCGTAGCCTATTGGTTAAGGTTTAAAGGCTTCTACACCCAGGTCTGGGCTTCGAATCCC
AGACTATGCAATTTATTGCAGATTACAGGAAATCCAGGTTTCAAGTTTCGGAGAGAGCGGTTTATTAAACAATTATGCAGA
CTACGGAGGAAAGGCTTGCAAAGGATCTTCAACATGGTGCAAGTAAATCTGGTCAAACGTGGATCTTCATAGGACGGCTCA
GGTGATGCAGTTAGGCGTAGGTCTTCATAAGGCAGGTAGTATTGTCGGTTGTCAAATCGTATATGTAATCTTTCCTATATC
ATAATTGTAAGATCATAATAAATCAG**CGTTAAAAAAAAAGTAGGGCCTATTT**

CACTA AND HARBINGER DNA TRANSPOSONS: CHARACTERIZATION AND IMPACT ON *BRASSICA* GENOMES

This section covers sequences of reference CACTA and Harbinger elements characterized in present study. The details about the elements are listed in table 5.1 & 5.4 respectively. The TSDs of elements are shown in red, while TIRs are shown in blue colours.

***BoCACTA1:* 9399 bp CACTA in *Brassica oleracea* (EU642504.1) from 20580-29972**

TAG*CACTACAAGAAAAACA*CCGGTATTCCGACGACAAATATCGTCGGTATGTCCTCAGAATAACGGTATTCCGAGGACATAC
CGACGAGAATGGTCGTCGGAAATTTCTTGTCGGAAAGTAAAATTTTCTCGGAAAATTCCGAGAAAATTCCGAGGAAATACC
GAGAATTAAAGATTCCGAGGACATTCCGAAGAAACATTTCCTAGGACTGTTCGTCGTATCTCCTCGGATATTTCCGATGGA
ACGGTCCTCGGAAACATTCCGAGAGAAAAGAGGTATCGGAATATTCCGACGAACTTTGGCCGTCGGAATATTCCGAGAAAC
CTTTGCTGTCGGAATATTCCGAGGAAGTGTATCTGTCGGAATATTCCGAGGGCATATATTCCGAGGAGATATGCCCTCGGA
ATATTCCGAGGAACTTCGGCCGTCGGAATATTCCGAGAAACCTTTGCCGTCGAAATATTCCGAGGAAGTGTATCCCTCGGG
ATATTCCGACGACAACTTCCTCGGAATATTTCCGACGGATACACTTCCTCGGAATATTCCCAAAAATATTTTTTTTTTAAT
TAAAAATTTATTTTTTTTAATTGAAATTCGAAAATATAAAATTAAAATTGAAATAGAAAACATATTAGATAATATTCAAAG
TTGTACATACCAAAATAAAACATTCAGAGTTTTTTTTTTAAAAAAAGAAAGAAACTACGGGAACTGCTCGTTCGGGTACAT
CCTCTTCATCATCTCCATCATTTGCTCGTTCTGTTTCCTCGTTGCCTCCCATCCCGCCTCTTGATCCACCATCTTTTGCTC
CAACGCAGCTATGCGGTCATCTTTGTCCTTCAACTGGTCCATAAGCACTTCTGGATCAACAAAGGGCTGTGGTGCAGAAGG
AGGAACCGACTGGGCTCGACGACCCAAACCGACCAAACGTCCCTTCTTCTTTGGAACCGACTGAAATAGAAAATAAACAAA
TTTAAATAATAGAAAAGATGATAAAATGAAAATAAAGAAATAAATGAATTGAACTTTTTTATAAAGAACATACCGATTCAA
CGATTTCGTTGATTCGAATCCAGGACAAGTTGGTCGAAGCCGTCGAAGTGGCGTCCAGGTCGGTTTGAAGCTGAGATAGAC
GGGTCTCTTCGTCTTCCTTCTGAGTTTCGATCAGGCTGAGCACATCCCTCACAACACCATCATCAATCTCTCCGGTGTGCT
TGTTGGTATACGCCGTCTTCATTAGGACGAAATCATCAACCGGCTCGCCATCATTTTCATCCGCCTTGAAAAAATAAAAGT
TAAACAAACATTAGAAATTAGAAGAAATGCATAAAAAATAAAATAAAATACTTAAATAATTAAAAAAAAGATTGAACTTAC
CAAGCGATCCCCCAGAGTGGCAATAGTCTGGGCACCCAAATTGTGCTTGTACATGCCCTTCCCGCTACGATCGCTCTTGCG
GTTGGCGGAGTTGGTCACGGAGGTAGCTTTCGTCTCGTCCTTATCCCAATGCACACATAACTCCTCCCAGACCGTGCTGTT
GATCGACTTCGGGACCTTTTAATAAAAAAAATATAAAATAGTTTAATAAATCCAAAAATTGTTTAATAAATTCAAAAAAAC
CTGGTTGATGAGCCACTTCTGCTTCCACTCGTAGATCTGCTTCCCATAGTTGTCCATAACTTTATGGACGAAGGCGTCACG
GATAAAGTTCCTGTGATCGGAATTCCAGGTGAACTCTTGCTGAAAAAAAACACAATTTGTATAAATTTTATATTAAAGATT
AAAAATATAAGTTAAAAAATTAGAATACTTACCGCAAACTGACGAAGCCACATCTCCTGGTCCTCGGCAAGGAAGTGAGTG
AAAGTCTGATATCCTTTGCTGAGGTTCGAGTACATCATATTGTTGATCCATGCGCTGATCCCGTTCCCGGATCGGTTGAAC
CTAATATAAGAAACAAATTATTAATAATGAATCAAATTTTAATGAAAGGAAAAAAAAAACTTTTAATTACCATGTTTGACG
TCGTCCCTTTGGACACGGAGTGAGATAGGGAAGATGCTCACGACCGGGCTGGTAAACTAACTGGGCAACAATCATCACTCC
CGGAACGACTGGAGGAGCGGGAGGGGGTGCAGCAACGGGAGGGGGTGCAGCAGCTGGAGCAAAAACCGGAGCGAGAAATGA
AGAATCCCGACAGTGGCTGCTCGATCCCCGAGAGTGGCTGGACGAACCCTGAGAGTGGCTCCCCGTACCACTACGACCTCG
TGGTGAGGCACGACGAGATCGAGACCGGGTCTGATCACCAGTAGACCTGTAAATTACAAAAAACATATTTAAAAATAGTTT
TAATAAATATTAATTGAACAAATATTTAAAATTAGTTATAATAAATATATAAATTAAAAAAAAATATTTAAAGATAGTTTT
AATAAATATTAATTGAAAAACATATTTAAAATAGTTTTTATAAATACAAAAAATAATAAAATAGTGTTTATAAATAAAAAA
AATGTTTTATAAATACAAAAATAGTTTTAATAAATAGAAATATTTTTATAAATATGAAATCATCTTTTATAAATCCAAAAA
TCGAATTTATATACAAAAAACGTTTTGTAAAATCCAAAAAATCGAATTTATATACAAAAAACGTTTTGTAAAATCCAAAAA
TCGAATTTATATACAAAAATCGATTTTATAAATACAAAAATAATAAAAAATTTATAAAAAAGTTCTAAATCAATTCAACAC
AACAAATTACAACCAAATCACAATCCTAACCAATCACCCTAACAAAAATCTATCAAAAAACTTCTAAAATCATCAAATCTA
CATAAAAACCTAACAAATAGAACCTAAGAGAGTGGGGGATAGGATCCTTACATGATTTATGTAGGTTTAGCGAGGATTCGC
CAGAGAGATCGTCGTGAGAGAAGGGGGAGATCGCCGGAGAGGAAGAAGAGAGAAATGGGGAAGAAGAAGTGTTTCGTCTTT
ATAAAACCTAGGGTCCGACGGAAACTATCCGTCGGAATTTCCTCTGTATTTTCAATTTCAATATTCGCGAAATATTTGGCG
GCTGGTTTGCCCGGTTAAATGAAACTATACCGAGGAAATTCCGACGGACACTTAAATATCTGTCGGAATTACCTCGGTATG
TTCATTAATTCGAGGAAAAAAGAATATCCTATGCATGTATTTCTATTGGCTTATATTGTTCCTCGGAATTTCCTCGGACTA
TTCCGAAAAAATTCTGAGGAAAACATGTTTGGGGTTTCAAAACATCAAATTGTTTTGCCTATATCATTTCTTATACAAATG
CAATGCATACCATTGAGGAATTTTTGTATAGATGATCATAAACTATGAAATAACAAAATTTCAAAACGAATCGTAAGTATT
CCTTTTACCGTTCATTAAAGTGTATAAGTGTTTCTCTTATGTTGTGGGGATTTCGTTCATACAATCGGAAAAATGTTTATT
ATAGGGTAAGGAACAAATTTTTGACTTCATAATTAGTCTAAGACACTTAATAAGGGTTATATAAGTGTTATTCAAACCGCA
AAACGTTGTTTTCTGTTTAAAAACCCTACTTCCTCGGAAAAGCCTCGGAATATTCCGAGGAAATTCCGAGGAACACTTGGA
TTTGCTCGGAATTTCCTCGGAATATTCTGAGGCTTTTCCAAGGAAACAGAAACCCTAAAAGAAAAGTCCTAAATCGTCGGA
AAAGTCGCGGACTATTCCGAGGAAATTCCGAGGGAATCCCTAAATCCTATTATCCCTCGGAATTTCCTCGGTATTCTTATA
TTAAAAAAAAAGACGATGCTACTCTGTTTCCGAGTTATCATCACCAGATGAATCTGAATCTGGATCTTGGTGAAACTCTCC
AATCTCTGGTTTATCCTCTACGTGAACGACAGTTTCCTCTCCAAAGTCGGTTAAATCGACTACAAGGCCAACTGCAACTAA
ATCTTCTGCTGCACTTAAGTTACCGGATGTGCTTGGTTGTAGTGGGTCTTCCAGATGAGAACTTCCCTGAATTCGGCCTCT
CGGGTTGAGTGATGTAACAATGACCCATGGATCATCTCTGTTCCTAACCCGGGGGTACTTGATATAACAAACCTGTAACAT
TTAAAAAAATTATTTAATTTGTAAATAAGTGTTATTAATACACATGATGATGAATCATTATGAATAATTGATGTATGTATA
TTTAATTACCTGAACGGCCTGAGAAGCAAGAATGAAAGGATCATAATATTGCAACTTTCGCCTCGAATGTACTGATGTAAC
ACCAATGCATCTGTTCTCACACCTCGATCTGGAGTGTTGTCGTACCAATCACAATAGAAAACAGTACAGCGCAATCCAACC
ATGTCTGGATACTTAATTTCCATAATCTCTCGTATGTTTCCGTAGTATACATCATCTCCAGATGCAGAACAAACGCCAGCA
ACATAAGTCGTACTCGAACGTCTCCTTTTTTGAGTTGTGAATGCATATCCTTGAGTACAAAATCTCGGATATGACTTCACA
ACATACTCTGCTCCACGCACCATCTCGCGTATCCAATCGTCAAATGTTTTACCTCTGGTCATACCAGCAGACACCTATTAA
TAGCACATATATATGTTATATCAATATATGTGAATTAGTATAAATATGTGATAAATAATATTTTAATTTGTTTAAAACACT
CACATAAGTAAGCATCCATCCAGAAATTTCTCTCTGTTTCAGTTCTTCAAGTTCTTCCTCTGTGGCGTATCTATATTCGAA
CCGCTTTTATGCCATGAAAATCATGTACATATGTTGACAATAATATAATTAATTAAGATTTAAATTTCAACTTGTTAAAAT

AAAAATTTGTATGCTAATTTATTTACCTCTCATATTGAAGAACATCTTCACAGTTGGTGAGCAAATATGTTTGCAAATGAC
TGCGCTCCTGCTCAGTAAGTCGACGGTCCTTTGGTTTTCCGCTAAGTCGTCCAACATCTGTGAAGATGTCTGGAACCGTAA
CATGATATGTTGCCCGTTCGCCTCTATCATCATGCCGAACAGGTCTTCTTTTTTTGGTCTGCACTTCTGCTGGAAAGTAGA
ACTCGACAAAGTTTGAAGTTTCTGAATTTATCATCTGTGCGATTATAGAACCTTCCACCCTACTATAATTTTTCACCATCT
TCTTCAAATGGTACATATACCGCTCATACAAATACATCCATCTATACTGCACAGGACCACCAAGTTCCAATTCTCCTGCCA
GGTTAATAACAAGATGCTCCATAACATCTAAAAATGATGGAGGAAATATCTTCTCAAGGTTGCACTGAATCACGGCTATGT
TAGTCTTCAAATTTTCAATACCTTCAAGAGTCACTGATCTCGTGCATAAATCGCGGAAGAAAGCACTAATCCCTGCAATTG
CTTCATGAACATTCCGTGGTAATAGTTCCTTGAAGGCAAATGGAAGGAGGCCATGCATCATTACATGGCAATCGTGACTCT
TCAAGCCGATAAACTTTCCTTCCTTTCTGTCGACACAGTTACGCAAATTAGATGCGTAACCATCTGGAAATTTGGTAATAG
GTTTTTCTCTGCCTATTGCAAATGTTGTATTACAGGGGGGTGAATGTCGAACCAGTCCTAGTGATGTGAATCAGTGAGAAT
ACAAGTCATTTCTTAAGCTAAGCAGATTCAATAGTGGTGAGTGTGTGACAAGTAACAGTAGCAAACAGAGCAATACAACAA
AGTAAAAAATCAATTTAGACAAGTGAATCCATGGGTATTGGGAATTGACTTCAAGTAACTAAGATCCAATCTAGGTGACAA
GCTTTCAATCAAAGTAATCTCTTAAGTCTAAACACAATTCTAGACAAGTCCTATGTCTAGGTAAATGCCCATTTGCTTAGA
AATCATTAAACATCAAATGTCTTTGGCTTAATTCAATCAAGCAATCTTTAAGTTCAAGTTCAATAACTATCTAGCAATTTT
AACATCAAGTGTCCTTGGCTAATCTCACTAGAGCTTAGTTGAGTTGATTCAAACACTTCATCTAATCATGTCTGATGAGAA
GTGTTTAGAAATCAGGTTTAGAGTGATCAAGACTAAACAAGCATTAAAAATACTCAACAAGCAAGTTCATACAAGGATCTA
ATACAATAACATCATAGATCTTCACTAAGTTACTCTAATCTCCCTAACCCATGAATTCTTAAGGAAACTACTCACTAATCT
CCATGAAAACACTTAATCTCATAATAGATTGAAGCATATTCAAGTAGATACAACAGAGAATAAAGATAAACAAGGATTAAA
CAAGCCAAACGAAATTAAAACTCAAGAACAGATGATGAACAATTGAAGAACAGCTCAAGAACACTAAGAACAATGATATTT
CAAAGAGAGAGAGTTTTGACAAGTTAAATTATTACCAAGTAAAATAATGATGTTTCTTGCCCACTCTTTTGATAAAAAGAA
AGCAAATATTATACAGAAAAGTAGGAAAAATCATGCAAAGGAAAAGTGGGTGGAAGAAAGTGGGAAGTTGGTTAATCCCAT
CAACTTTCCAGTGGTCCCCGGCAAGGGTTGATACACCTGTAGTAAATCGATTATAACTTATCCAATACATCTCCAAATGAC
TTGAAACCACTTCCATTGGAAAGCTAACTCAATTTCCAGTGTCTCCACAAATTTTGGGCTTCAGAAAAGATCTTTAAGGCT
TTCATCCGTGTTCATCTTTCACTCTCTGAAATTCCATGATTCGGGAGCGACCTCGCTGTGTCGCTCTAGTAAGGTCGCTCC
AGTAGGGTGATCAATACGCGAGCGACTCTTTCATGTCGCTCTAGTAAGGTCGCTCCAGTAGGGTGATCAGAGCGACTTGGT
GGTGTTGCTCCGGACTGGTCGCTCCGGGTGATCAATACGCGAGCGACTCTTTCCTGTCGCTCTAGTAAGGTCGCTCCAGTA
GGGTGATCAGAGCGACTTGGTGGTGTCGCTCCGGACTGGTCGCTCCCATGCCTTGCTCGCCCAATGACGACTCTAAACACT
CCTTTTTGAGCTCCAAATGCCTCCAAGTGTCTCCAAGAACTCCATGTGGTCCTCCCATACCTGATAAGGACTCATGTATGC
AAAATGCAACCTAAACATGGCTAAATCCTAATCTATATGATCAAAATGCACATGGATGAATGGATAAAACAATAGAAATAT
GCAAGATATCAAAATTCCATATCGTTTGAAATCCAATCAAAGAACGCATCTTTTCCCTCTGCATCAAGTCGGTATATGGGA
AAAGGAGCCCTACCATTCTTATCAACATGAAGTTCTAAACGAGCACATATATCGACTAAATCCAGTCTTGACTTCAAATTA
TCCTTTGTTTTACCTTGAACGTTAAGGATCGTGTTCATGAGATTGTCAAAAGTTCTTCTCAATATGCATGACATCTAAATT
ATGCCTCAGCAGATGATCCTTCCAGTATGGCAGATCCCAGAAAATACTCTTTTTGTGCCAGTTATGTAGCTCTCCAACCGC
ACTGACTCGAATGTTTTCATGTCCACCTACATCTGGTGTCCTTTCTGCATCAAAATACCTAAACTGCTTCAACAAATCTTT
CCCACTAACTT**CCTCAGGTGGACCATCAAACACCTGCTTGTTCTTCGTAAACAAAGTCTTACTCCTGCGGTATGGATGATC**
AGGTGGTAGAAATCGTCTGTGACAGTCAAACCAACACGAAACATGAAAGTACTGCCACATAAGTACTGCCTTCATTTGAAA
GTTTTTTTTACATGAAACATCGTATGTTTCAGCACCTTGAGCTCATAGTTGTTGCAACTCATATATTAGTGGCCGAAGAAA
CACATCAAGTGATCTCTTAGGATGCTTTGGTCCGGGAACGAGAATGGAGAGAAACAAAAACTCTCGTCGCAAGCACAAGTT
TGGCGGTAAGTTGTATGGTGTAAGAATGACTGGCCATAGAGAATATTGTCTTTCACTCTTGCCAAACGGGCTGAAACCATC
AGTACATAATCCAAGGTAGACATTTCTTCTCTCATACGCAAAGTCGGGATATTTTGATTGGAAATGCTTCCACGCTTTTGC
ATCTGAAGGATGTCTAATCTCACCATTTGTTGAATGCTCTGCATGCCATCTCATTGGTTGCACTGTGCGTTCATACTGATA
CAACCTCTGCAACCTTTTCGTCAAAGGCAAATACCACATGCTTTTAAATGGCACTGGAACTCTTCCACTCGTATCTTTATA
ACGAGCATTTCCACAAAATTTGCATGTAATCCGCTCTTCATCCGCCCTCCAATAAATCATGCAGCTATCGTCGCATACATC
TATTACCTGATACGATAAACCAAGACCAGCTACTAGTTTCTGAACCTCGTAGTATGAACCAGGGGCTACATTATCCTCAGG
TAGAATACCTTTTACAAAATCAGCAATCGCATCCACACAGTCTTCAACCAAATTATAATCTGTTTTAATGTCCATCAGTCT
TGTTGCAGATGATAAAGCTGAATGACCATGTCTTCAACCTTCGTACAAGGGTTGCTTTCCAGCATCCAACATATCATAAAA
TCTCATAGCTTCTGCATTGGGTAAATCTTCTCCTCTAAAATGATCATTTACCATCTGCTCAGTACCTACACCATAATCTAC
ATCCATTCTAATTGGTTCTTCTAACCTAACCGCAGGCTGAGGTTCGCTAGTACTACCATATTCATAATCAGTTTCTCCATG
ATGATACCAAATTTTGTAACTTCGTGTAAACTCATTCAACTATAGATGAGTCCAAACATCTCACTCTTTAATAATCTTTCT
ATTTTTACAATTAGAGCAAGGACATCTTAACATACCCGTTTCTGCTTCCGGTTGTCGTTGAACTAACCCCATGAATTCGAA
TATACCTCGTTGGTATTCTTCCGTAAGCAATCTCGTGTTCGGATCCAAATGAGGTCGATCGATCCAAAAACGATAATAATT
TGAAGAAGACATGTTTTTTTCAATCAAATTCGTGTGTAAATATAGTATGTGAGAGGATGAAGAAATGGAGTGAATGAAGAG
GTTGGGGGGTGCGGGTATTTATAGATGAAATGCTGCCGACAGTACCGAGGAAATTTCGACGAAATACTGACGGCAACGCCT
CTTTCCTCGGAATTTCCTCGGAATTTTTAAAATGCCCAACGGCTCTCCAACGGCTATAATATATTCTCGGATATTCGTCGG
TGTTTTCCGAGGAATATGCCTTCTTCGGTATTTCATCGGTATATTCCGAGGAAACCCAAATTTTGGGTTTTCTCGGAATTT
CCTCGGAAATTCGTCGGAATATTTCGAGGATCTCATTTTCCGTCAGAATGTCCGTCAGAATTACGA***TGTTTTCTTGTAGTG***
TAG

***BrCACTA12*: in *Brassica rapa* (AC189341.2) from 99395-107196 bp in 3′-5′ orientation**

GAT*CACTACAAGAAAACA*TCGGGATACTGAGGGAAAAAATCGTCGGTATGTCGTCGGAATAACGCTATTCCGAGGACATAC
CGACGAAACAAGTCCTCGGAAATATCTCCTCGGAATTTCATACTTCCTCGGAATTCTGTCGGAAAATTCCGACGGAATTCC
GAGGAAACAAAATTCCGAGGAAACTCCGAGGACACGAAGTTCGTCGGAAGATTCCTCGGAAAATACCGAGGGAATTCCGAT
GGTCCAATCCTCGGAAGTTTCGACGAAATAATCCTCGGAATGTTTATCGGAAAATACCGAGGAACAAATGTTCCTCGGACT
TTTCCGAGGAACTTTCTTCCCTCGGAAATTCCGAGGGAATGGAGTTCCTCGGAAAATTACGAGGAACTTTTTCCCTCGGAA
AATTCCGAGGAACTTAAAGTTCGTCGGAAATTGCCGAGGGGAGAAAGTTCCTCGGAATTTTCCGAGGAACTCCATTCCCTC
GGAATTTTGAAAAAATTTATTTTTTTAAAAAATTTATTTTTTTTAAAAATTAATTTTATTTTTAAAAATTAAAATTTTTAA
AAGTTAAAATTAAAATTGAATAAAACATAAAATAAAACATAGTAGATAATATTCAAATTTAAATAAAACATTCAGAGTTTT
TGGAAAAAAAAAAAAAACTACGGGTCTTGAATTTTCGGGAACGTCTCGTTCGGGTACATCCTCTTCATCATCTCCATCATCT

GCTCGTTCAGCCTCTTCTGGGTCTCATAGCCCGCCTGTTGAGCTGCCATCTGGGTCTCCAACGCAGATATCCGATCATCCT
TGTCCTTCAGCTGAGCCGTAAGAACTTCTGGATCAACATACGCATGTGGTGCAGAAGAAGGAGCAGCCGACCGGGAGCGAC
GACCCAAACCGACCAAAACGTCCCTTTTTCTTTGGAACCGACTGAAATATAAATAACCAAGTTTAAATAATTTAAATTGAT
GATAAAATAAAAATCAAGAAATAATTAAATTGAACTTTAAAAAAAAAAAACTTACCGATTCAACGATTTCGTTGATTCGAA
CCCGAGACAAGTTGGTCGAGGCCGTCGAATCGTCATCATCGGTTTGAAGCTGAGACACTTTGTCATACACCTGAGTTTGGA
CCAGGTCGACCACGTCCCTCACAAGCCCATCATCAATCTGGCCGGTCTTCTTGTTGGTATACGCCCTTTTCATAAGGGCGA
GATCATCAACTGGGTGGCCTTCATTTTCTTCCGAAAAAAAATAAATTAGACAAACATTAGAAATTTGAAAATGCAAAAAAA
ATAAAATTCTGAAACTTAAATAATTGAAGAAAAAGCGGTTGAGCTTACCATGCGATCCGCCTGAGTGGCAATAGATTGAGC
ACCCAAGTTATGCTTGTAGATGCCCTTTCCTTTACGGTCGCTCCTGCGGTTGATGGAGTTGGTGGAAGAAGTTTCTTTCGT
CTCTTCTTTATCCCAATGCGCACACAACTCCTTCCAGACCGTGTCGTTCATCGACTTTGGGACCTTTTAATTTAAAAAAAA
ACATTTAATAATTTTAAAAATAGTTTAATAAATTAAAAATTGTTAATAAATTAAAAACTACCTTGTTTACTTCCCACTTCT
TCTTCCACTCGTACATCTGCTTCCCATAGTTGTCCATAACTTTATGGACAAAATGGTGATAGATAGAGAGCGTATCATCGG
AATTCCAGTTGAATTCTTGCTGAAATAGTTTAAAAATAGTTTTTAAGCACAAAAATATAATAGTTTAATAATTTCAAAAAA
TAAGTTTTAATAATATTTAAAATGTCTAATAAATATAGAAAATAGTTTAATAATTACAAAAAATAAGTTTTAATAATATTT
AAAATGTCTAATAAATATAGAAAATAGTTTAATAATTACAAAAATAGTTTTAATAATATAAAAAATATAATTTAAAAATTA
GAATACTTACCGCAAACTGACGAAACCACAGATGCTGCTTCTCGACAGGGAAGTGAGTGAAAGTCGGATGTCCGCTGTCGA
GGGCCGAGTACATCATACGGTTGATCCATGCGCTGATCCCGTTCCCGGATCGGTTGAACCTAATAAAAAAAATAAATTATT
AATAAAAAAATAGTTTAAATAAAAAATAGTTTAAAAAATAAAAGTTTAAATTACCATGTTTGACCATGTCCATGTGGATAC
TCAGTGAGATAGGGAAGATGGTCACGACCGGGCTGTCGAACCAACTCCGCAACCGTCATCACTCCCGGAGGACCGGGTGCA
GCAGCGGGACCAGGAGCGGGTGCAGCAGCGGGAGCGGGAGCGGGAGCAGCAGGAGCGAATAATGGAGAGGGAGATGTATGG
TATGAGCTGTGGGGCGAAACGGAATCCTGAACATGGCTGGAAGAACCCCGAGACTGGCTCCCCATACCACCCCGGCTACGA
CGCTGGCGAGGCCGGATCTGATCATCATTAGACCTGTAAATTAAAAAAAAAAACATATTTAAAAATTAGTTTAAAAATTTT
AAAAATAGTTTAAATAAATCCCAAAAACATAATAATTACAAAAATAGTTTTCATAATATTAAAAATGTTTAATAAATATAG
AAATTTATATTTTACATATAAAATACATAAAACGTTTTATAACCAAAAAAAAATGTTTTTAAATATTATATAAATGATTTT
TAATCTTAAAAAAAGTTTTATAGATTTTAAAAATGTTTAATAAATATATAAATGAATTTAAAATGCAAAAAATAGATTTTT
TATACAAAAAACGTTTTCAATATATATATATAATTTCAAATTCGATTTTTATAACCACAAAATTAATAAAAACTAAAAAAA
AAATTCAAATCAAGTGAAATACATAATCAAACTCGATTTTACACAAATCTACCATTCACCCTAACAAAATCTATAAATCTC
ATTCCAAAATCATCAAATCTACTTAAAAACCATACAAATCTAACCTAAAAGAGTGGGATAGGGTTCTTACATGATTATGGG
GGAGGATTAGGGGAGATTCGCCGGAAAAACGCGAGAGAGAACGGTGGAGTCGCCGGAAATCGCGAGAGGAGAGGCTGTGCG
GCGCGGAGAGAAAGAGAGAAATGGGGAAGAAGATCGGGCTCGACCTAATATTATGAGGCGTCCGACGGAAACTATCCGTCG
GAATTTCCTCGGAAATATTTTATATTTCCGTCGGAATTTCCTCGGAATTCTTTGATTCAATTTTCCCGAAATATTTGGCGG
ATTGGTTTACCCGGTTAAATGAAAATATTCCGACGAACCAGGATTCCTCGGAATACCCTCGGTAAACTCCGAGGAATTTCT
GAGGAACCAGGGTTTGGGGTTTTAAAACATCGATTATTTTCGCCATATTTCATTTCTTATACAATTATAATGCATACCATT
GAGGATTCTTTGTATAGATGATCAAAAACCATGAAATAACACAATTTCAAAAATAATTGTAAGTATTCCCTTTACCGTTTA
TTAAAGTGTATAAGTGTTTCTCTTATGTTGTGTGGTTTTCGTTCATGCAATCGTAAAAGTGTTTGTTATTGGATAAAACAC
CAAAGTTCATAGTTCCAAACATCATAATCTGGTTATGACACTTAATGAAGGTTATATACGTGTTATTCAATCCGTAAAACG
TTGTTTTCGATTTAAAACCCCAAGTTCCTCGGAATTTCCTCGGAATATACCGAGGAGTTTCCGAGGAGATCCTTATGTTCG
TCGGCATTTCCTCGGAATATGCCGAGGAGATTCCGAGGAAAAAGAATCTATAATCAAATCCCCAATTCCTCGGAATTTCCT
CGGAATTTTCCGAGGAAATCCCGACGAACATGAGGATCTCCTCGGAATTCCCTCGGAAAATTCCGAGGAAATTCCGAGGAA
CTTTGGGTTTTCAATCGAGAAGATCTATTCCGAGGAAATTCCGAGGAAAACCTACCTGTCCTCGGAATTTTCCTTTAAAAA
AAAAAAAAAAAAGGTCGACGCTAGTCTGTTTCCGAGTCGTCATCACCAGATGAATCTGAATCTGGATCTTGGTGAAACTCT
CCAATCACTGGTTCATCATCTACGTGAACGGCGGCTTCTTCTCCGAAAACGGTTAAATCGACTACAAGGCCAACTCCACCT
AAATCTTCTGCTGCACTTAAGTTGCCGGATGTGCTTGGTTGTAGTGGGTCTTCCAGCTCAGAACTTCCCTGAACTCGGCCT
CTCGGGTTGAGTCTTGTAACAGTAACCCATGGATCATCTCTGTTCCTTACCCGGGGGTACTTGATATAACAAACCTGATCG
GCCTGAGAAGCAAGAATGAAAGGATCATAATATTGCAGCTTCCGTCTTGAATTTACTGATGTAACACCAAATGCATCTGTT
CTCACACCTCGATCTGGAGTGTTGTCGTGCCAGTCACAATAGAAAACAGTACAGCGCAATCCAACCATGCCCAAATACTTG
ATTTCCAAAATCTCATTTATGTGTCCGTAGTATACATCATCTCCTGATGCAGAACAAACACCAGCATCGTAAGTCGTACTC
GAACGTCTCCTCTTCTGAGTTGTGAATGCATATCCTCGAGTACAAAATCTCGGATATGACTTCACAACAAAGTTTGGTCCA
ACGACCATCTCGCGTATCCAATCGTCAAATGTTTCACCTCTGGCCAAACCATCACTCACATAAGTAAACATCCATCCACCA
AATTCTCTCTGCTTCATTTCTTCTAGTTCGTCCTCTGTGGCGTATCTATATTCTAACCGCTTTTCTGCCATGAAAATCCTT
TCATATTGAAGAACATCTTCGCAGTTGGTGAGTAAATATGTTTGCAAATGACTGCGCTCCTTATCAGTAAGTCGACGGTCC
TTTGGTTTTCCGCTAAGTCGTCCAACGTCTGTGAAAATGTCTGGAACCTTCCAATGATATGTTGCCCGTTCGCCTCTATCA
TCATGCCGAGCAGGTCTTCTGTTTTTGGTCTGAACTTCTGTTGGAAAGTAGTACTCGGCAAAGTTTGAAGTTTCTTCATTG
ATCATCTGTGCGACTATAGACCCTTCCACCCTACTTAAATTTTTCACCATCTTCTTCAAATGGAACATATACCGCTCATAC
AGATACATCCATCTATACTGCACAGGACCACCAAGTTCCAATTCTCTTACCAGGTGAATAACAAGATGCTCCATAACATCA
AAAAATGAGGGAGGAAATATCTTCTCAAGGTTGCACTGAATCACAGCTATGTTAGTCTTCAAATTTTCAATACCTTCAAGA
GTCACTGATCTTGTGCATAAATCGCGGAAGAAACCACTTATCCCTGCAATTGCTTCATGAACATTTCGTGGTAATAGTTCC
TTGAAGGCAAACGGAAGGAGACGCTGCATCATTACATGGCAATCATGGCTTTTCAAGCCAGTAAACTTTCCTTCCTTTCTG
TCGATACAGTTACGCAAATTAGATGCGTAACCGTCTGGAAATTCCACATCGTTTGAAATCCAATCAAAGAACGCATCCTTT
CCCGCTGCATTAAGTCGGTATATGGGAAAAGGAGCCCTACCACTCTCATCAACATGAAGTTCTGAACGAGCACATATATCG
ACTAAATCGAGTCTTGACTTCAAATTATCCTTTGTTTTACCTTGAACATTAAGGATCGTGTTCATGAGATTGTCAAAAAAG
TTCTTCTCGATATGCATGACATCTAAATTATGCCTTAGTAGATGATCCTCCCAGTATGGCAGATCCCAAAAAATACTACTT
TTTTTGTGCCAGTTATGTTGGTTTCCAACACCATCTACCGGAAAATGCTCATGTCCACCGACGTCTGGCGTCCTTTCTGCA
CCAAAATCTCTAAGTTGTGTCTTCAAATCTTTCCCACGAATTTCCGGAGGTGGACTGTCAAACACCCTCTTGTTCTTCGTA
AACAAATTCCTACTTCTACGATATGGATGATCTGGTGGTAGAAATCTCCTGTGACAGTCAAACCAACACGTTTTCCTTCCG
TGTTTTAGTTGGAAAGCATCAGTGTTATCTTGACAATATGGACATGATAGCCTTCCATGCGTTGTCCATCCAGATAACATA
CCATATGCTGGAAAATCACTTATTGTCCACATTAGTACTGCCCGCATTTGAAAGTTTTCTTTATACGAAACATCATATGTT
TCAGCACCTTGAGTCCATAGTTGTTGCAACTCATATATTAGTGGCTGAAGAAACACATCAAGTGATCTCTTAGGATGCTCT
GGTCCGGGAACGAGAATCGAGAGAAACAAAAACTCTCGTCGCAAGCACAAGTTTGGGGGTAGGTTGTATGGTGTAAGAATG

ACTGGCCATAGAGAATACTGTCTTCCACTCTTGCCAAACGGGCTGAAACCATCAATACATAATCCAAGGTAGACATTTCTT
CTCTCATACGCAAAGTCGGGATACTTTGATTGGAAATGCTTCCACGCTTTTGCATCTGAAGGATGTCTGATCTCACCATCT
GTTGAGTGCTCCACATGCCATCTCATTGGTTTCGCTGTGCGTTCAGACAGATACAGCCTCTGCAACCTTTCCGTCAAAGGC
AAATACCACATCCTTTTATATGGTACTGGAACTCTTCCACTCGTATCTTTATAACGAGGCTTCCCACAAAATTTGCATGAA
ACCCGCTGTTCATCCGCCCTCCAATAAATCATGCAGTTGTCTCTGCATACATCTATTACCTGATACGATAAACCAAGACCA
GCTACGAGTTTCTGAACCTCGTAGTATGAGCCAGGAGCTACATTATCTTCGGGTAGAATACCTTTTACATAATCAGCAATC
GCATCCACACAGTCTTCAGCCAAATTATAATCCGTCTTAATGCCCATCAATCTTGTAGCAGATGATAAAGCTGAATGACCA
TCTCTGCAACCTTCGTACAATGGTTGCTTTCCAGCATCCAACATATCATAAAATCTCCTAGCTTCGGCATTGGGTAAATCT
TCTCCTCTAAAATGATCATTTACCATCTGCTCAGTACCTACACCGTAATCTACATCCGTTCTAATTGGATCTTCTAATCTA
ACCGCTGGCTGAGGTTCGCTAGTACTACCAAGTTCATAATCAGTTTCTCCATGATGATACCAAATTTTGTAACTTCGTGTA
AATCCACTCAAATATAGATGAGTCCAAACATCCCATTCTTTAATAACTTTTTTATTTTTACAATTAGAGCAAGGACATCTT
AACATACCTGTTTTTGCTTCCGGTTGTCGGTGAACTAACCCCATGAATTCGGTTATACCTTGTTGGTATTCTTCCGTAAGC
AATCTCGTGTTCGGATCCAAATGAGGTCGATCGATCCAAGAACGAAAATAATTTGAAGAAGACATGTTTTTTATGAATCAA
ATTCGTGTGTAAAGAAAGTGAGAGGGAGGATGAAGATATGGAGTGAATGAAGAGGAAGAGGGGTGCTTGTATTTATAGTTG
AAATCCTGCCGACGGTCCGAGGAAATTCCGACGGAATTCCGATGCGAACGGCTAGTTCGTCGGAATTTCCTCGGAATTTTT
AAAATCCCCCAACGGCTCTCTAACGTCTATAATATTTCCTCGGAATTCATCGGTTTTTTCCGAGGTATACTAGTTTCCTCG
GTATTCCGTCGGAATATTCCGACGAGAAGAATTTTCCTCGGAATTCCGTCGGAATATTCCGACGGAATACCGAGGAAGAAA
AATTTTGTGTTTCCTCGGAATTGCCTCGGAAATGGCTCGGTATATTCCGAGGAATTCATTTTCCGTCGGAACGTCCGTCAG
AATACCGC***TGTTTTCTTGTAGTG***GAT

***BrCACTA35*: A 3029 bp CACTA in *Brassica rapa* (AC232476.1) from 93851-96879 bp in 5′-3′**

TTG***CACTACAAGAAAACA***GCGGTATTCTGACGGACGTTCCGACGGAAAATGAATTCCTCGGAATATACCGAGGCATTTCCG
AGGCAATTCCGAGGAAACACAAAATTTTGCTTCCTCGGTATTCCGTCGGAATATTCCGACGGAATTCCGAGGAAAATTCAT
CTCGTCGGAATATTCCGACGGAATACCGAGGAAACTAGTATTCCTCGGAAAAAACCGATGAATTCTGAGGAAATATTATAG
ACGTTAGAGAGCCGTTGGAGAGCCGTTGGGGGATTTTAAAAATTCCGAGGAAATTCCGACGAACTAGCCGTTTGCATCGGA
ATTCCGTCGGAATTTCCTCGGACCGTCGGCAGGATTTCAACTATAAATACAAGCACCCCTCTTCCTCTTCATTCACTCCAT
ATCTTCATCCTCCCTCTCACTTTCTTTACACACGAATTTGATTCATAAAAAACATGTCTTCTTCAAATTATTTTCGTTCTT
GGATCGATCGACCTCATTTGGATCCGAACACGAGATTGCTTACGGAAGAATACCAACAAGGTATAACCGAATTCATGGGGT
TAGTTCACCGACAACCGGAAGCAAAAACAGGTATGTTAAGATGTCCTTGCTCGAATTGTAAAAATAAAAAAGTTATTAAAG
AATGGGATGTTTGGACTCATCTATATTTGAGTGGGTTTACACGAAGTTACAAAATTTGGTATCATCATGGAGAAACTGATT
ATGAACTTGGTAGTACTAGCGAACCTCAGCCAGCGGTTAGATTAGAAGATCCAATTAGAACGGATGTAGATTACGGTGTAG
GTACTGAGCAGATGGTAAATGATCATTTTAGAGGGGAAGATTTACCCAATGCCGAAGCTAGGAGATTTTATGATATGTTGG
ATGCTGGAAAGCAACCATTGTACGAAGGTTGCAGAGATGGTCATTCAGCTTTATCATCTGCTACAAGATTGATGGGCATTA
AGACGGATTATAATTTGGCTGAAGACTGTGTGGATGCGATTGCTGATTATGTAAAAGGTATTCTACCCGAAGATAATGTAG
CTCCTGGCTCATACTACGAGGTTCAGAAACTCGTAGCTGGTCTTGGTTTATCGTATCAGGTAATAGATGTATGCAGAGACA
ACTGCATGATTTATTGGAGGGCGGATGAACAGCGGGTTTCATGCAAATTTTGTGGGAAGCCTCGTTATAAAGATACGAGTG
GAAGAGTTCCAGTACCATATAAAAGGATGTGGTATTTGCCTTTGACGGAAAGGTTGCAGAGGCTGTATCTGTCTGAACGCA
CAGCGAAACCAATGAGATGG**CATGCGGAGCACTCAACAGATGGTGAGATCAGACATCCTTCAGATGCAAAAGCGTGGAAGC**
ATTTCCAATCAAAGTATCCCGACTTTGCGTATGAGAGAAGAAATGTCTACCTTGGATTATGTACTGATGGTTTCAGCCCGT
TTGGCAAGAGTGGAAGACAGTATTCTCTATGGCCAGTCATTCTTACACCATACAACCTACCCCCAAACTTGTGCTTGCGAC
GAGAGTTTTTGTTTCTCTCGATTCTCGTTCCTGGACCAGAGCATCCTAAGAGATCACTTGATGTGTTTCTTCAGCCACTAA
TATATGAGTTGCAACAACTATGGACTCAAGGTGCTGAAACATACGATGTTTCGTATAAAGAAAACTTTCAAATGCGGGCAG
TACTAATGTGGACAATAAGTGATTTTCCAGCATATGGTATGTTATCTGGATGGACAACGCATGGAAGGCTATCATGTCCAT
ATTGTCAAGATAACACTGATGCTTTCCAACTAAAACACGGAAGGAAAACGTGTTGGTTTGACTGTCACAGGAGATTTCTAC
CACCAGATCATCCATATCGTAGAAGTAGGAATTTGTTTACGAAGAACAAGAGGGTGTTTGACAGTCCACCTCCGGAAATTC
GTGGGAAAGATTTGAAGACACAACTTAGAGATTTTGGTGCAGAAAGGACGCCAGATTGATGATGGTCTTGTGAGGGACGTG
GTCGACCTGGTCCAAACTCAGGTGTATGACGAAGTGTCTCAGCTTCAAACCGATGATGACGATTCGACGGCCTCGACCAAC
TTGTCTCGGGTTCGAATCAACGAAATCGTTGAATCGGTAAGTTTTTTTTTTTTAAAAGTTCAATTTAATTATTTCTTGATT
TTTATTTTATCATCAATTTAAATTATCTAAACTTGGTTATTTATATTTCAGTCGGTTCCAAAGAAAAAGGGACGTTTGGTC
GGTTTGGGTCGTCGCTCCCGGTCGGCTGCTCCTTCTTCTGCACCACATGCGTATGTTGATCCAGAAGTTCTTACGGCTCAG
CTGAAGGACAAGGATGATCGGATATCTGCGTTGGAGACCCAGATAGCAGCTCAACAGGCGGGCTATGAGACCCAGAAGAGG
CTGAACGAGCAGATGATGGAGATGATGAAGAGGATGTACCCGAACGAGACGTTCCCGAACATTCAAGACCCGTAGTTTTTT
TTTTTTCCAAAAACTCTGAATGTTTTATTTAAATTTGAATATTATCTACTATGTTTTATTTTATGTTTTATTCAATTTTAA
TTTTAAATTTTAAAAATTTTAATTTTTAAAAATAAAATTAATTTTAAAAAAAAAATAAATTTTAAAAAAAAATAAATTTTTT
CAAAATTCCGAGGGAATGGAGTTCCTCGGAAAAGTCCGAGGAACATTTGTTCCTCGGTATTTTCCGATAATCATTCCGAGG
AATATTTCGTCGAAACTTCCGAGGATTGGACCATCGGAATTCCCTCGGTATTTTCCGAGGAACCTTCCGACGAACTTCGTG
TCCTCGGAGTTTCCTCGGAATTTTGTTTCCTCGGAATTCCGTCGGAATTTTCCGACGGAATTCCGAGGAAGTATGAAATTC
CGAGGAGATGTTTCCGATGACTTGTTTCGTCGGTATGTCCTCGGAATAGCGTTATTCCGACGACATACCGACGATTTTTTC
CCTCGGTATCCCGA***TGTTTTCTTGTAGTG***TTG

***Bo-N-CACTA1:* 3265 bp non-autonomous CACTA in *Brassica oleracea* (AC240092.1)**

GTT***CACTACAAGAAAACA***ACGGTATTCTGACGGACATTCCGACGGAAAATGAAATCCTCGGAATTTCCCGAGGAATTTCCG
AGGATATTCCGAGGAAACCCAAAATTTGGTTTCCTCGGAATATACCGACGGAATTCCGAGGAAATATCAATCCGTCGGAAT
ATTCTTATGAAATACCGAGGAAAAATGTGTTCCTCGGAAAAAACCGATGAATTCCGAGGAAATATTATAGCCGTTGGAGAG
CCGTTGGGGGATTTTACAAAATTCCGAGGAAATTCCGACGATGATGACTCGGAAACAGACTACAATTGATTTTTTTTTAAA
GAAATACCGAGGAAATTCCGAGGAACACTTGATATACTCGATCGATCGATGTGTTTTTGAACATATATCCATCGATCGATC
CGTTTATAAAAAAACGTTCGAGATTTTGTTCATTCGATCTATTGGATATCCACATCGATCGATCGGATAATCTAATCGATC

GATCACATTGTTACGCTTTGGTACATCTTAGTCATCGATCGATGCGTTTTTGGACATATATCCATCGATCGATCCATTTAT
AAAAAAACGTTCGAGATTTTGTTCATTCGATCTATTGGATATCCACATCGATCGATCGGATAATCTAATCGATCGATCACA
TTGTTACGCTTTGGTCCATCTTAGTGATCGATCGATGCGTTTTTGGACATATATCCATCGATCTATCCGTTTATAAAAAAA
CGTTCGAGATTTTGTTCATTCGATCTATTGGATATCCACATCGATCGATGGGATAATCTAATCGATCGATCACATTGTTAC
GCTTTGGTCCATCTTAGTGATCGATCGATCCATTTTGGGATCGATCGATCACTTTGTTACCCCAAACCCTGTTTCCTCGGA
AAAGTCTCGGAATATTCCGAGGAAATTCCGAGGAACATTTGATATACTCGATCAATCGATGCGTTTTTGAACATATATCCA
TCGATCGATCCCTTTATAAAAAAAATGATGTTTTGAAATCCCAAACACTAGGGAAGAAAACGGTTTCCTCGGAATTTCCTC
GGAATATTCCGAGGAAATTCCGAGGAAATAGGGTTTTAAAACCGAAAACAACGTTTTGCGGTTTGAATAACACCTCTATAA
CCCTTATTAAGTGTCTTACGTTCATTATGAAGTCAAAAATTTGTTCCTTACCCTATAATAAACACTTTTTCGATTGTATGA
ACGAAATCCCACAACATAAGAGAAACACTTATACACTTTAATGAACGGTAAAGTGAATACTTTCAATTCGTTTTGAAATTT
GTTATTTCATGGTATATGCTCATCTATACAAAGAATCATCAATGGTATGCATTACAATTGTATAAGAAATGAAATACGGCA
AAAAAATTGATGTTTTGAAACCCCAAACACTAGTTCCTCAATGAACGGTAAAGGGAATACTTTCAATTCTTTCAATTCGTT
TTGAAATTTGTTATTTCATGGTATATGCTCATCTATACAAAGAATCCTCAATGGTATGCATTACAATTGTATAAGAAATGA
AATACGGCAAAAAAATTGATGTTTTGAAACCCCAAACACTAGTTCCTCGGTATTTCTTCGGAATATTTTCATTTTACCGGG
CAAATATTTCGCGAAAATTGAAATTAGAATTCCGACGGAATTCCGACGGAATTCCGACGGATAATGTCCGTCGGACCCTAG
GTTTTATAACCACGAGCCCCTTCTTCTTCCCCATTTCTCTCTTCTTCCTCTGCGTGAATCCTCTCTTCTCTCCGGCGATTT
CCCCCTGAATTCCGACGATATCTCCGGCGATCTCCCCCTTCTCTTACACAAATCATGTAAGGACCCTATCCCACTCTCTTA
GGTTCTATTTGTTAGGTTTTTTGTGTAGTTTTGATAGATTTTTGTTAGGGTGATTGGTTAGGATTGTGATTTGGTTGTATA
ATAGGTTTAGAATTGTGATTTGGTTGAATAATTTGTTTTGTTGAATTGATTTAGAATTTTTTTATAATTTTTTTATTTTTT
TGTATTTATAAAATCGATTTTTGTATATAAAATCGATTTTTGTATATAAAATTGATTTTTGTATTTTACAAAACGATTTTT
GTATATAAATTTGATTTTTTGGATTTTACAAAATGTTTTTTGTATATAAATTTGATTTTTTGGATTTTACAAAACATTTTT
AATATCTATAAAACTTTTTTTGTGATTAAAAACTATTATTTGAGATTTGTTTTTTAAAAAAAAATATATATTTATATTATA
TAAATTTTTTATTTATTAAAACTATTTTTTGTTTATTAGAACTATTTTTATATATTTATTAAACGTTTTTAATATCTATAT
ATCTTTTTAGAATGATTAAAAACTATTATTTGGGATTTTTTTATAAAAAAAAAAAATATATATTTCTATATTTATTAAAT
ATATTTTTTTTTTAATTTACAGGTCTCATGATGATCAGACCCGGCCTCGACAGCGTCGTGGTCGTGGTGGTACGGGGTCAA
ACATGGTAAGGATGATCGCATATCTTTGTTGGAGACCCAGATGGCGGCTCAACAGGCGGGCTATGAGGCACAGAGGAGGGC
TGAACCAGCAAATGATGGAGATGATGCAGAGGATGTACCCGAACGAGGTGTTCCCGGACGTGCCAGACCCGTAGTTTTTTT
TTTCTAAAAATTCGGAATGTTTTATTTTTATTTGTGAAACTTTGAATATTAATTAATATGATTTCAATTTTAATTTTAATT
TTATATTTTCGAATTTAAATTTCAAAAATTTTATTTTTTTAAAAAAATTAATATTTTTTACATTCCGAGGAAATGAAATAC
CGAGGAAATTCCGAGGAACACTTGATATACTCGATCGATCGATGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCAAAAACGCTTCGATCG
ACGAAGTTCTGAGGAAATATCCCGACGACGTTCTCCCTCGGTATATTCCGAGGAGCTTTCCGCCGAACTAGTGGTCCTCGG
AATGTCCTCGGAAATTTGTTTCCTCGGAATTCCGTCGGAAAATTCCGAGGGATTTCCGAGGAAAGAAGAAATTCCGAGGAG
TTATTTCTGAGGATTTGTTTCGTCGGTATGTCGTCGGAATATCGTTATTCTGACGACGTACCGATGATTTTTTCCCTCAGT
ATGTCGT***TGTTTTCTTGTAGTGGTT***

Br-N-CACTA6: **1288 bp non-autonomous CACTA in *Brassica rapa* (AC241034.1)**

TGG*CACTACAAGAAAA*AGCACCGCATGCCGACAGCCATCTTTGTCGGTAAGTAATAGGAACAAGCTCGAACCGATGGTCTT
CTGAGAAAACTCAATTAGTCGGAGTTTGGGTCTCGGAAATGGATTGATGTCGGAGCTGTGTCGGAAATTTCCGACGAGATA
TATCCTCGGAAATTTTGTCGGAATCGTTTGTCGGAATACACCGAGACCGTTTCGATGAAATGTGAAGCCACAAACTCGTCG
AAGCATGCTCGGAAGTTGGTTTGTCAGAAGTTAATCGGTATTTACCGAGGCCTTTCCGATGATACATGACATTGAATATAC
CGAGGAGAAATCGTTACTACAGTATTTCCGGTGAATCTGTTTTACACTTCTGGGAAATTGTTTGAAGACCACCGTAATACA
CTAATACATAAAGATGTTTTTTCTGGAAATTTAATAGATAAAGACGTTCTTAACAAAGGAACAAATATACACCATTTGATT
ATTATAAAGATATACACGGGATATATATTGCTTACAAGAAAAATACAGACTAAATAACTCATAAAAACTAAAAAGTATAAA
GGTAAATGTATTGAAACACTAAAACTAAATTTTAAAATATTAAAAATATATCTTCTTCATCTTCGTCTTCTCCACACCTTC
TGTAAACTCAACTTCTTTGTTGACAAGAAGAATCTCTTGATGTGCATACTCAAATAGCTTCTTAGTGTCTTTAGCCTTATC
TAATCTCAATTGTAAAATCGATTTTCTGCAAACACTCGTATATATGTTTTCAAATTTCTCTAGTTCAGATCTACACATGAA
GACACACATGAACGCTTATATAAAGAAATACTGAAAAAGCCATGGTTGGTTTTGCTTCGAATGGCGTCACATGTACGCTAG
AAGCTATTATGACAATTTTCCGTCAAATCTTCCGAGATCAATTTCTCCTCGGTAACTGCCGAGAAAGAGCTCCGACAATTT
ATCGTCTACTTTTCCGAAGCCATACCCAGAAAATAATTGATTGTCGGAAATTGATTGTTCCGAGAAATTACCGTCAACATT
TTCTGAGAACCTTCCCACAAAACCTGTTCCTCAGAAATTTTTTTGGTTTCCGACAAATTGCCGACTATGCATCTCGAGAAA
ACAGCGAGAAAATAGATTCATCGGTAATTTCCGAGATGATCCCGACATAATTCATATGTTATCGGAATTTTATCACAAACA
TGTCGTTACCTTTCCCACAAAAGAAGTTCTCGGTAAATTTCATCGGAGTGGCCATTG***TTTTCTTGTAGTG***TGG

BoHARB1: **3843 bp Harbinger in *Brassica oleracea* (AC240081.1)**

TAA*CAATAGGTCTGTTCGTTTGGTGCCCGCAGATTCCTGCGGCTG*AGTCTGCGGCTGCGCAGCTTGTTGTTCGTTTCGAAG
ACGCAGGACATTGCCGTATCCGTAACCGCAGAAAATAGATTTAGTTCTGTTCGTTTGGTTGCCGCAGACGCAATTTTCATT
TATTTCATATTTTAAATTTATGAATTTTATGAATTTGGAAAATTGTATTTTAAATAAATAAAATGTAGTTTAAATTTGTAA
ATTGTATTTTAAATAAATAAAATGTAGATTTAAATTTGTAAAATTGTATTTTAAATAAATTTAACAACATTTTTTTATCAT
AAACAAAAGCTAAATAATACATCAAGTCTTAAATAAATAAAACAAATATTTCAAATATCACTCTTCCAAACTCGCCTTCTT
GTGGTCTTAATGTATCCTCCGCAAACTTCAGAATAGCACTCAAAACATCATCAAGCTTCCTTTGGATCGTATCCAATGATC
TTTGATACCTTGCAGCAATATCCCGCTTAGTCTTATCTTGACCAACAGTCTCGAGAAACATCGCAACAGATTCCTCAAGGT
AGACATGAGGATACATGACAAAAGGCAGTTTCATACAAAGATCAGACCAAAAGCAAAAGTTCCATACTAGTTCTAAGTTCC
ATACTAGTCATAAGTTCCATAATAGTCCTGAGCTCCATACTAGTCTTAAGTTCCATAATAGTCCTAAGCTTCATAGTCACG
TTTTATCTTGGTCCTAAGCTCAAGAAACATGATCTTAGCTTCATCTGTCTTCATTGTTATAAACGCCCTTCTGCTCCCTTC
ACTGTCTATAAGATGTTCCATTGCTGTTTCATATAATGGAGACCACTCTCTCACTCCAGGCAATGTGTACAGAATCTCTAA
CACATTCTCAACACTAGAAGCTTCTTGACGCTCCAACATCCGTTCGGCTATCCTATTCTTCACCATAAGAGCTTCAGTACG

TTTGTCACAAGCTTCTACAACCATATCCTTCTCCTTACGCTTTCTTTTTCCTCTCGAACTTCCAGAATGAGTCTGAGTTTG
GCGCTGGGTTTGTATCTGAGCTTGTGGCTAAGTTTCTGTCTCTGCTGCTGGCTGAGAATCGGCTTCATCACCATCATCTAC
ACCCACTCTAGAATTCAAGCTTGCTTCTCCATGTTGAGCACTCCATCCTTCTGCTCCAGTTACTACTACAGCACAAAACTC
CTCTTCAAACAGATCCATATTAAATTCCTCTTTCCATATGGATCTTCTAATCCCAGCACACTCCTATATAACAAAAAAAAA
GAGTAAGTCAATGAGGAATCATGAAGCAATCATGAAGCAAATATAAGAAAAAAAAATACACAGAAGAGATAAAAGAAACAG
ACATTTTCGCGCTCACTCCACCAATCATCTGACATATCAACCCTTCCCATGCTGTCAAACCCAAGACCTATCTTATTCTTA
ATCAACTTCCTAAATTTGGTGTACTTCTTTCTACTTGTGTCGTACTTGTTCTTGAACACAGGCCAATCGGGATACACTTTC
TTAAACCTCTCGTTTCCCTATTTTATTCATTGATGTCGTCCTATTTCCTTTTCTTCTCTCTTCCGCGTAGAGTTCGAAAAA
ATACCTAACTTCTTCAGGAGTCCAAGTCTAAATCATAGAAACAAATAAACTCAAGTTCTATACTGACCTAACGACAGTTCA
GTTCCATACACTACTCAAGTTCTATACTGACCTAAAGCAAGAAAGTTCCATACATATAACAAGTAAGACAGTTCCACACGA
CAAGACAGTTCCATACATATAACAAACAAGACTAGTGTGAGAGAGCCTTACATCTTTTGTCATCTTAACAAGTGAGAGGAA
CAGCGGTAGAGAGTGAGCTTGAGATGAGAGAGCCTGCAAAAAAAGACAGACAAGCTCAGATTGATTGGTGAACGTATAAACA
TTAATAGTCAGAACATCAAATACAATAAGCAAAGCCAAAGGCTGTGAAACGTATAAACATTAGATTGATTGGCTATAACAA
GCAACAAACGCTAGCAAGCTACTACTTGCCGATCACAGACAAGCTAGCAAACGTATGGGTTTGTAACTAAGAGAGAAACCG
AGTAGAGTCAATAACCAGAACACAGAAGATCTTATACAAACCCATCAGAGAAACCCATAACACAAACCTAAGAAAGAGACA
ATTGAGCTTAAGATGTACAATGCTTACAAGAGAGAGAGCTGGAGAGGCGTCATAGAAGAGAGAGAGCTGGAGAGTGCGAGC
TTGAGAAGAGTACCTGCAAAAAAGACAGAGAGATGGAGAAAGAGTGTTAATCTCAAAGAGCTTGGTCCGAAAAAAAAAATA
ACAGACCTTTAGAACCAGAGCATCAGACTCAAACCAGATCACTCAAGTTTTATTCCTCAAGTGAAGTAACTCAGGAAACAA
ACCTTACTAAAGAGATTCTGAGCTTGAGAGGGTTTGAACTTGAGAGAGCCTGAGCTTGGAGAGGGTTTGAACTTGAGAGAG
CCTGCGACCTGCAAAAAAAAATAACAGACCTTTAGAAGCTGGCAAATGACAGTACATGTTTTGTGACTTAAAACCACTTGGA
AAGAGAGCTCAAATGCTCAAATATACATGCAAGGACGTATTGGTGATGTAATTCATCTAATTAATCACAGACTCACTGATG
AGAGTAAAACAAATGCATCACAAATATACACAGAACGTCATAACAATACGTAAGAGATTAAAAAGAAGAAGAAGGCAATTT
TTGGCAATCTATACGATCAACTAGTGTAGTCACAGAACATAACTAATAAATCAAAGACATCATACCTTGAGAGAGTGAGCT
TGAGAAGGTCTCGAGAGAGTGACTTGAGAGGAGCAACGGCTTTACACAAACCTAAGCATCAAAAACCTAATGCTATCATCA
CTTCATCAACCCAAAAACCCATAAGACTCAAATCAATATGTGGTATCAGAGACAGTTACCTCGGAAAAGAGAGAGTTCGAG
AGAGAGAGATCGTGAGAGAGAGCTCGGAGAAGAGAGAGCTCGTGAGAGAGAGCTCGGAGAAGAGAGAGCTCGTGAGAGAGA
GCTCGTGAGAGAGAGCTCAGCTTTACACAAACCTAAGCATAAACAAGACATGATGGGTTAGTGTCAGAGATCAAGAGAGAG
CTTGAGATGGGTTTGTTCAATAACAGAGCAAAGAAAAAATCTGAAACTTTCAAGTCAAATACTAGTCCAAATCTTAACGAA
TTGATCCAAAGTTTTCAAAAGGATGAAATAATTACCTCAAAGAGAGAGAGAGAGCTCGAGGTCGGAGATTGAGAGACCGAG
AGAACTCGGAGATGGGTTAGTGTCAGAGATCAAGAGATAGCTTGAGATGGGTTTGTGTGATGGATACAGAGCTCGAATCCT
CGCATTGCACCATCACTTCATCTCACACTTGCAGATGGAGAGAGAGCTTGAGATGGAGAGAGGACTTGAGATGGATAGAGG
GCTCGTGATGGAGAGAGGGCTCGAGATTGAGAGAGATCGAGAGAGAGATCGAGAGATGGAGAAAGAGATTGAGAGAGATGG
GAGAGAGTTGCGTCGATGGGAGAAGACCTGCGGCTGCGTCAATTAGATTTAGGTTTTTTCGTAAATAATTAATTAGGTTAA
GGATATAATTGTAAATATGTTAAATAGACGCACAATTTACGTCGCGACCGCGAGATTTTACGGCGTCAAACGCAGGTCTCG
GCGTCAGACGCAGCGGTTAGACGCATGTCTCGGCGTCAATCAAACGAACAAAATTTTTAAAAATAGTCGCCG***CAGCCGCAA***
GTACCTGCGGCAACCAAACGAACAGCCTTAATGTAA

BrHARB4: 3527 bp Harbinger from *Brassica rapa* (AC189588.2)

TAA***TTAATGGTTGCTTTA***GTGTTTGGTTACGGTTGTGTCACTGTAATAAATAGAATTAATGGTGTAGTTTGAGTTCTTTGA
AATTGAATTGAATTGTTGTTAGTTGTTAGTTGTTTGTTGTTAGATATATGTCTAGTCATTGCAATTGAAGTGTTCTTTCCA
GCTTTTTTAAAAAGATCTTTCATCCATTTGCTTGATCTTTTTATAAGCTGTGTCGTTAGTGCCTTCTCCAACAGATGAATG
CGCATTCTCGTTCCTTCTCCTCTTTCGCATTCTCCTCTGTCTCCTCTTTCTCCTCTGTCTTCTGTGTCTCCACTCTCTCCT
CTTTCTCCTCTGTCTCCACTCTCTCCTCTTTCTCCTCTTTCTCCTCTCTCTCCTCTTTCTCCTCTTTCTCCTCTTTCTCCT
CTTTCTCCTCTGTCTCCACTCTCTCCTCTGTCTCCTCTGTCTCCTCTCTCCTTGACTTCGAGGATGGATTTCAATCCATTT
CAGGACTCGGCTAATTTTGTTGATCTACTCCATAGTCAACAAAATGTTGTCTTTGGCAGTCAAGGACGTGTTCCACTCTCT
TCAACGCAAGTGCCGCCTTCTGGCAGTCAAGGACGTGTTCCACTCTCTTCAACGCAAGTGCCGCCTTTTGGCACTCAAGCC
GCAGAGCTTCCAGCAGAGCGTAAGGAAAGAAGGACGTGGATAGTCACAGAGGACATAGTGCTTATTAGCAGCTGGCTCAAC
ACGAGCAAAGACCCTGTCGTAGGGAATGAGCAGAAGTCAGCAACTTTCTGGACAAGAGTTGCAGCATACTTCTCGGCGAGT
CCAAAACTTGCTGGATGTGAAAAACGCGACGGGAATCAGTGCAAGCAACGTTGGCACAAGCTGAATGAAGCCGTTTGCAAG
TTTTCTGGGGCGTATGAGGCAGCCACAAGAGAGAAAACCAGTGGCATGAATGACAACGACGTCCTAAAACGTGCCCATGAA
ATCTACTTCAACAACCACCAAAAGAAGTTTGTCCTTGAGCATGCGTGGAACGAGCTTCGCAACGACCAGAAGTGGTGTGAT
CTCGCTACATCTAAAACTGAAACAAGCTCCAAAAGGAGGAAGTTCGCGGATGGTTCACATTCAGGAGCATGCTCTCACGTC
AATGAATGCAATGCTGGTGGAGAAGGAACATCTCGTCCCCCTGGTGTTAAGGCTGCAAAAGCTGGAGGAAAGAAGCCACAT
GTCGCGGGGGAGGATGTGTGTGATTATCAGCTCATGTGGAGCATCAAGAAGGATGACTTGGCAATGAAGCAACAGCTCTCC
AAGATGAGAGTACTTGAGAAGCTTCTTGCCAAGGAAAATCTAGAGGATTATGAAGAAGATCTCAAGAAAAAGATCAGTTTA
GAGTTAATGTAACTCTTGGAATAATGTTAACCTTATGTATGTTTCATGTTCTTGGCTTGTAGTTGTTTCATTTTATGTCTT
CCTTATGTTTCATGTTTCATGTTATCGGCTTTAGAATGTAGTTGTTTCTAGTTTCATGTTCTCGGCTTTAGAATCTAATTT
GTAATGCTTTGCTCTCTTATTTTATAAGCAACTTAATGTTTTGGTGTTTTGTTGAGTTATCTCTTGATGTGGACAATTGTT
TTATTTGTTTCAGGTAACAGTAAATAAGAAGCTACATTGGATCATGTCACGGGTTGCATACTATCAGTTCAGAGTGTTGTC
TCGGCGTGTGTTGTCTCGGCTTTGTAGGATGTCACGGTTTTAATCTTTTGTATCAGTGTGTCACGGGTTTCAGTTTCAGCT
TTTGTATCAGTGTGTCACGGTTTTAATCTTCTGTATCAGTGTGTCACGGGTTGTTCAGTGAATCATATACTTTGTATTTTC
ATGTGCATCACACAGCCTCATTTTTTATATATATATCAATGTCTCTCTACTTCCAACATCACGCAAACTCATCTACTCTCT
ATTTCCTCTCTTTTGCTAACACTACCAAACGCACAGTCTCTCTATTCTCTCTATTTCCTCTCTTTCCCATGTTTCTCACTT
TGAACACAATTTAAAAAAAAACTCTTAAATCACTTCTATGGCTTCTTCTTCAAATCCAAACACTTTCGATGAATCATGTGA
TGATACATTTGATGACTTCTTTGATCAAAAATTTGATGAAAAATTTGATGAAAAATTTGATCAATTTTTTGAGCAAGCGTT
TGAGAATTTAACTACTCGAGAGGCTCCAAAAAAAAAAAGAAAACCGAAAGTTTATATCGAGAGAAATCGTGAAGAAGGGCA
TATTCGTTTATGGAATGATTATTTCAGTGATAATCCAACATATCCTGATAATTTATTCCGACGACGTTTTCGAATGAACAA
GCCATTGTTCTTGTACATTGTTGATCGACTCTCCAACGAAGTTCCCTATTTTCGGGAAACAAAAGATGGTCTCGGAAGGAT

TAGTCTCTCTCCTCTTCAAAAGTGTACCGCAGCCATTCGTGTCTTGGCGTATGGTTCTGCAGCTGATGCGGTCGACGAATA
CCTCCGGCTCGGTGAAACAACCACTCGCTTATGTGTGGAAAATTTTGTGGAAGGAATAATATATTTGTTCGGT**AATGAGTA**
CTTAAGAAGACCAACACCAGCTGATCTTCAACGTCTACTTGATGTTGGAGAATATCGTGGATTTCCCGGGATGATAGGAAG
CATCGATTGTATGCATTGGGAGTGGAAGAATTGTCCAACCGCTTGGAAAGGGCAATATTCTCGTGGTTCGGGTAAACCAAC
AATCGTTTTAGAGGCGGTTGCTTCATACGATCTCTGGATATGGCATGCATTTTTCGGACCTCCAGGTACATTAAATGATAT
CAACGTTCTTGATCGTTCACCTGTTTTTGATGACATAATAAAAGGTGAAGCTCCGAATGTCACTTTCTCTGTCAATGGAAG
AGAGTATCATATGGCTTACTATCTTACGGATGGTATTTATCCGAAATGGGCAACTTTTATTCAATCTATTCCTCTACCACA
AGGGCCACAAGCGGTCTTATTTGCTCAACATCAAGAAGCAGCCCGAAAAGATGTCGAGCGGGCTTTTGGAGTCCTGCAAGC
TCGCTTTGCCATTGTTAAAAATCCAGCGTTGTTTTGGGATAAAGTCAAAATTGGTAAGATTATGAGAGCATGTATCATACT
CCATAATATGATAGTAGAAAACGAAAGAGATGGATACTCTCAAAATGATGTTTCAGAATTTCAGCAAGGAGAAGACAACAG
AAATTCACATGTCGATCTCACGTATTCTACGGATATCCCTTCAAATATCGCAAACCAGATGGGGGTTCGGATAAGAATTCG
TGATAAACAAATGCATCAACAACTGAAAAATGATTTGGTTCAACATTTATGGCATAAATTTGGACCTGGTGAAGACAACAA
CTGAGCTTGGAAGCTTCTTTTAAATAATTCTTGTTTAATTTAGTAATCTTTGTTTTTATGTTTTAATTATAAAATTTCTAT
TAATAAAAAAATTTAAATTTTATTT***TAAAGAAACCCTTAA***TAA

Bo-N-HARB1: **1199 bp non-autonomous Harbinger in *Brassica oleracea* (EU642504.1)**

TTA*GAGAATCTCCAAAAGAAACTCTAT*AACTCTAAATATAGAGTTTTTTGCTCTCCAAAAAGAACTTTAAATTTGAAGTTT
TGAAAAGTGAAACTCTAAATATAGAGTTTCACTCTTCAAAACTTCAAATTTGAAGTTTTATCTTTTTATTTGCATTTTGGT
CCTTACAATTATACATCATATTTATAATTCTTAAATATTTTTTTGTTTATTGTTTTATTCCTTAAAACTTTTATATCTCAT
AAATATTTCAAATTTGTTTTATAAATTTAAGTTTTACACATAAAATTAAATAAAAATTTTAAAACAGATTTATAATATTTT
AAAAGTAGAATTAAACAACAAGAATATTACAAAAAACCATAATAAAAACTTATTAAAAAGACACATGAAGACATAATATAC
TCAATTTAATATTAAAACAACACTAATAGTCTGGTAAATTTGCTCCAGAACCTCCTAAATCTCCAAAATGCTACCCAAACA
AATTTCGTGTAACCGAAGATGATTGTTGTTGTTCCTCTTGACGCCTTTTTCGTACGATTCTTTCTTGTTCAGATCGAATGT
ATTCACGAATATTAACATCATCGATAGAAGCTAAGTTTTTTAGCAATATTTTATTTTCCTCTTTTACTTCTTTGAAAGCTA
ACTTTTTTTGCTTCATTTTGCAGTCGTAATTCAATCATCTCATGACCTTTTTGCCTGCTTGTTTTACTTTCGTTTAAAAGA
TCAAGGATTTTTTCGTTTGATGATATCAAACTATCAGCGGGTATATTATGAGGATATCCAATTTCAATAATTTTTGGCTCG
TAATCTACAAATAAAAAATAAAGAGAATTAATTCTTACTTCGAAATGCACTAGTTGATCATATATGGGAGCATTACGGAAA
TAATTTTATGGAATAATGTAGTATTTGCTTGCAGTTTAATATTTAATTATGTACTTTTATTTATAATTTTATATCTTAGTG
TAAAATTTTTTTAATTAATATTTCTGTAATATTTATATATATGTACTAGTTATTTTAGAAGTTTTATGAATTTACATCAAC
TATGACAAATATAAGGATCATAGTGTAAAATATAAATAATTTTGAAGTTAGGTTTGAAGTTTTACTTTTGGAGAAGAACAC
ATTGAAATTTTAAAAATGGAGTTTTAGAAACTTCAAA***ATAGAGTTCTTTTTTTGGAGATACTC*****TTA**

***Br-N-HARB2*: 819 bp non-autonomous Harbinger in *Brassica rapa* (AC189298.1)**

TAC*AATATGGTGAATTGAAATAGAAT*CTTGGAACATAAAAAGAAGCTAAATATACTTTTGGGATTCTGCCAAAAGATGATC
CCATTTTAATCATATAACTAGATTTTGACCTGCGCTTTGAAAGCGCGGGTTTATTTTTGTTTTTTTTTTCAATTGACAAAT
ATTTAGTAAATGTCACATTTTCATATATTTGTGTTTTATTTTATAAAAGACTTAAAAATTTTATTTTATTTATCGTATTTC
ATTTTAAATGACTATTTATGTTAAAAAAATTAAACTTTATTTTTTTAATGAATTAAGTTGGTATAACTCTGATAAATTAAT
TTTATTATGGGGTTAATATTTTAATTAAAAAATTATATACTTTTAATAAAGATTTATCCTTTTCAATAAAAAAAAGCAATT
ATTTTTATGAATGCTTAAATTATATTAAGAAAAGAGAAAAAATAATAATTAAGAATAGTTGAAAAAAAATTATTTGAACTT
GGACTCAATGGCCCAAAGGAAAAAAAAAGTGAGAATTGAATCTGATTTTTTAATAGGCCCAAATGGCCCAAGAGAGATTTG
ATTTGGGATGGATCTAAAAATAATGACCCAATATAGATTTGTTATTAATATTACTTAATTACCCTTAATGAAACATGCAAT
GTTAGTGAAGGAAACATGCCCCTAAGGTAATTATGACAATAGGATCCTGCTTTAATAGTATAGATAGGTAGGTTAAGGTAA
TATTTTTTTCTTTACAGTATTACATCAAATATATAATTATTGGGGAGATACGTTTCTTGTAAGT***ATTTTTCTATAAGTTTTC***
GATATTTAC

CHAPTER 6
NON-AUTONOMOUS DNA TRANSPOSONS & NOVEL INSERTIONS IN *BRASSICA* CROPS: DIVERSE AND ABUNDANT

The non-autonomous hAT sequences are given below each represented by TSDs (red) and TIRs (blue). The details of each element are given in table 6.1

***BrN-hAT1*: 670 bp non-autonomous hAT in *Brassica rapa* (AC189298.1)from 1727-2396 bp**

AGTATTTT*CTAGGCCTGGGCATT*TTACCCGGACCCGAAGATCCGACCCGAAACCGATCCGAAAAAACCCGGTTCGGGTAAG
AACCGACCCGATCAAATTACCCTATCGGGTCTTGTTGTAGAGGACCCGCGGGTCTTGGACCCGACCCGACCCGAACCCGAA
ACCCGATGGGTACCCGAATTAATAATGTAAATTAAATAAAATGCATCAAGGGAGAATCGAAGACATGTTATTTGTCTAGAG
AATATGGGAGTCAACCACTAGGAGACAACGAATTGTTGTTGTATTTCGCATAGTTTAGTATATAACTATATATACACATTT
TAATATAATAAAAACACAAATTTTGAATAAAATTTTGAATATGTTCGGATATATAAATATTTTCAGATATTTTTTTGTATT
TTTAGGTCTTTTTTTAGATCGAACCCGAACCAAACCCGACCCGAAACCGAACCGAATCCGAACCGATAAATTCTAATTACCC
GAATGGGTCTAACTATCTAAGACCCGACCCGACCCGGACCCGGCACGACCCGAACCGACCCGGAACCGAAAGTTTGAAATT
ACCCTATCGGGTCCTAAACTCTTAGACCCGAAAGATCCGGACCCGAAAGGACCCGACCCGAATCCGATCCGACGATCCG***AA***
TGCCTAGGCTTAGTATTTT

***BrN-hAT2-1*: 716 bp non-autonomous hAT in *Brassica rapa* (AC189298.1) from 32998-33712 bp**

GTGTGGACTCTATGTTACATGGAAACGGAAGCGGGTACGTGGAAGCGGAAGCGTATGGAAGCGCAGAAGCGAGATTTTTAA
AAAAATTAGGAAGCGGATACGTGTTGGAAGCGTATCCATATATATATATATATATGTAAGACTAAAAATTAAAAATTGG
AATATATATACATATGTAAGATTATTTTTAAAAATCTAAAACTAAAATTTATATGATTTAAATTTAAAATAAAGCATTTAT
TCATTTATAATAATTTTGAAATAACTTCATATTTAAACTGTGAAAATACACATAAATTAAATAAAATAAATAATATTATAT
TTTTGTAATTCTTTATAATTAATTGACATAATATATTTGAATATATGATTTATCTTTAATAAAAACCTCAATGCATAAAGA
TAATTTATAGTATTAGTTTTAATATTTGTATATTTCTATTCTCTCTATTCAATACTATTAAAATTTGGATTATATATAAAT
TAAAAACTATAATTTTATATTCTTATTGATATAAGACATTATGTTTAAAAAAAAATGGAAGCGTGATTCCAAAACGGAATC
GTAAGTTTCCAATATGTTTTTAAAGATAATATTTTAGAAGCGTTTTGGAAGCGAGATTCCGTAAGCTTCCACAAGGTTCCG
ATTCCGATTCCGGTTCCGAAGCGGGAAGCGGACGTCCGATGAAGCTTCCGTGCAACGTAGGTGTGGAG

***BrN-hAT3*: 620 bp non-autonomous hAT in *Brassica rapa* (AC155341.2) from 46457-47076 bp**

TACAAATTTGGCTTTGATTGGTTAACATGTAGTACACTTGATTTAAGTTGACTTGAGCATATAAATCAAAAATGTTACCAAT
CATGAATTAGTTTTAAGATTTTGAAGTTGTGTTTGAGTTATAATTAATTATTATTGAGTTACAATTTTGACTTATAGCCTT
TGTAAAAATGATAACTCAAATCTTAAATCTACATCAACCAAAATATCATGTATTAAGATTTAAGTTATTGAAAATTTTGGG
TTTGTCAATTAACTACCCACTACCTTTTGAATATTAACAATTTTAACAATTTACAATACATATAAGCACTTAGCCCTACGT
TTAGTTATATTATTTAATATAATAATAAAATATTATTTATATTTATGAATTTTATTTTTTCACTATTCTTATTAAAATATT
ATTTATATTTTAAAAAATAATATATACTGTAACTTATACATCAAAAATGTTACCAATCAATAATATTGATAAAATAGTAAC
TCATATTTTTACTGCGAACTCACTACATAAATCTACAGTTTATACCACATAGTTTTTACTGCAAACTACATCAAATCATAC
ATTTTATTGTAACTCAACTCTAATACTACATGTCACCAGTCGAAGCCTTTAAAATT

***BrN-hAT4*: 786 bp non-autonomous hAT *Brassica rapa* (AC155341.2) from 105234-106019**

ATATTAGCTAGGGGTGGGAAAAAACCGAACCGAAACGAACCAACCGAACCAAAGTCTATTTCAAACTATTCGGTTGAAGAT
TTTCCCAATCCGGATAGTTTGGTTTTAAACCGAACAGAACCGAGAAACCGATGTGTTTTTATAATTATTTTAATTAAAAAT
ATTAGTAATACCGATATATTAAAATTCTAATACCATACCACATATTTTCCATTTTTCATTTATTAACATATATTGTTCTAA
CTGAATATGTAATCATTAAAATATTTTATTTAGAAAAATAGAAAAGTTTATATCTTTATATTCTTATATATATGTAAACAT
GGATGTTAGATCTAAATCTAAAACCTAAAGCAAATTTTATTAAAACAAAAGTTCTTCTCCACAACATCACATAGAAAATTA
TTCATTGCAAGCTTTTTAATAATGATATAAGAGACGTTACTAATGATGTTGGTTTTTTTTTAGTTAACTATCATTTGTTTT
AATTCTTATATATCTTTTGTGATTTTTAAAAGTTTTTGTTTCAGTTTGATATTGAATCTAGTTCACATAGTTTATGTTTAC
ATAATTTATACTAAAAACATAATTTCTCTCAAAAAAATTAAACTAAGAAGAACCTAACTAAGAAAAACCTAAATAAACTGA
TCCAAAAATTAAATCCAACCGAACCTAATCCAAACCAAACCAAACTAACAATGGTTTATTTCAATTGGAAAAATACTATAA
CCAAACTAACCAAACCGAACCGAGAAAATAACTGAAGTGCCCACCCCTAATATTAGC

***BrN-hAT5*: 979 bp non-autonomous hAT *Brassica rapa* (AC155344.1) from 34806-35784**

GATAGACACAGTGCCGGTCCGGACTTGAATGATGCCTAAAGCGCATTAAAAAATAATGCCCTTTAACTATATTTATTTATT
TTATATTTTAAGTTTGTTAAGAAAAATTATGCATCTTTGATTTAGTTTCAAAAGAAAATTATGCAAGATATTATCAAATTT
TATTTTCTTTAAAAAAATAATTTACATAAAGATAAAAGTATTTTTGAATTGGTTATTTTCATTTTCTAATTTTTTTTTATT
AATTTTGAATAATTGTTTTTTAGAAAGTATGTACTCATTCACACATGATATAATATAGTTTATTTTTATAGATTTTAGTTT
AAAAATACTAAAAAAGAATAAAATATCCAGTGGATAACAATAATTAAATTAGCAAAGACTGATTTTTTTTAACAAAATAAG
ATATTGAAAACTAAAACAAACAGATCGTTAAAACATAATTAACAACTATTAGTTTATCCAGTTTTATTTGAAATTGTATAC
TTTATATAACTAAGTAAAAGTTTAATATAACTTTTGTCAAAAAAAAAAGTTTAATATAACTATGTACAAGCCCGTATTTTT
TCGAAATTAATGACCGATATCATTCAGAAAATATATATTGTATACTTTATATAATATAAAGTAAAAATAATTGAGTGGAAA
AATGGAAATATAATTAAAAAAAAACAAACTATCTGAAATGGAAGAAATTAAGATTTTTTAAAAATCCATCGTATCTTAAAG
ATCACATGTCATTTGTTATCATTACCCTAAGAAAAGAATTTGTCTTGTCATATCATTACAAAAATAAAAATAAAGAAAAAA
TAGGCAAGAACTTGTTCGAACAGAACAACTGATGGACTAAATTCCAGGAACAAAAACCACTATACCAAGGGAGTAAGGTTG
TACTTGGTGCCCCTAACCTTTTGTTAAACTTTGGCGCCCAAAGCCTTTGCTTTATGGACTTTAGCCCAGGGCCGGGCCTGG
ATAGACA

***BoN-hAT6*: 402 bp non-autonomous hAT in *Brassica oleracea* (AC155344.1) from 88339-88740 bp**

AATTGGAGCAGTGTTTTTAAAACCGGACCGGGAGCTGAACCGGAAATTTTTTGGGTCACGGTTTAATATGGTCTGACCGGG
TCAAACCTGGTTCAATAATCTGGTTTAATATACTTTTAGTATATAAATTTAAAAATAATGTTAATAAATATGATACATAAT
TAAAATATAAATAACTTTTAAACATAAAACATTATAAATATGAACTAGTTTTATGTTTATATGGATGATTTATAGATTTTT
GATAGTTTTCGTAGTTTAATAACTTTAAGATCTAATCCGGTTACGTGATCGGTTCATGGTTGAACCAATTATTAGATCCAA
CCTGTTTATATACCCGGTTCATGGTTGAACCCGGTCTAACCATCGGGTCGGTCCGGTTTTAACAACACTGAATTGGAG

***BoN-hAT7*: 701 bp non-autonomous hAT in *Brassica oleracea* (AC240081.1) from 4835-5532 bp**

TTTCGGTAGGGCTGGGCAAAAAAACTCGTATCCGAAAAACCGAACCGAACCATATCCGAAAAGTAGTATCGAACCCGAACC
GAAATTGATTAAATATCCAAATAGGTTCAGAATTTTGGTATTTAGAGAACCGAAACCGAACCCGATCCAAAATGAAGTATT
TCGGGTACCCGATTATATCCGAAATTGATTTATATACCTATATATATATTAATTATTTTTAGATTTAATATATATTAAAAA
CATCAAAATATATAACATACTTTTAAGTTGTCCAAAATACTTATAAATATATACAAATAGTCAAAAGTACATGTCTAAAAT
AGCTAAAGTATACTCAAAACACCAACAATACTTAAAAATATCTAAGGATTCCCAATCCAAATATTCAAACCAAACCAATTT
ATATGTTAATTTTAGGTACTTTGACATATGCTATTCAAATTTATATGTAATATATTGTTTTGTTTATAGATTTTGAGAAAT
TTAAAGTATATAATGAATTTTAAAAAAAATTCAAAAAAAAATTAAAATAATTTAAATGGGTTATCCGAACCCGAACTGAAT

CCGCAAAGATCCGAACCGAACCCGAACCAAAATTTAGAAATACCCGATTGGGGCTAAAATATTTGAACCCGAAAATCCGAA
ACCCAAATAGATCTGAACCGAATGGATATCCGAACT***CCCACCCCCTA***GTTTCCA

***BoN-hAT8*: 998 bp non-autonomous hAT in *Brassica oleracea* (AC240081.1) from 65232-66229 bp**

GGAATAC***TATAATTTTTATT***TAATATATATAAATTATTTTATGTATTAAATATTTTTACATATTATAAAATAATAAATATA
TATTAAATAATTAAAAGTCAGTAGCTATTAAATATATAATTAAATTTGTACGAACATATAAATCAATTTTATTAATCCAAA
AAATATTTTTTTTATATTTGATAGGATATGTTATTAAATTTAAATGATACTAACATAGATAATATATTTTAGTATATTTTT
AATATTAATGTCTATTAAATGATGATTTTACTCATATAGTTTTTTTGATCATTTGTATCTTTTATAGAAAAAAATTTAAAT
TACTGGTAACAAAATTTTCATTGTGAGATTAATAGTTTTAGTAATTTATAATTTAAAAAAAAAGATAAATTGTCAATGATC
GTTCAAAACTTTTATCAAAAAAACTGTTCAAAGTAAATTTTGAAACTAAAATATTGTATTTTATATGGTTTATAGTTTAAT
TTAAAACGATATATATAATAATCTTAATAATTAATTAAATTAGACTTTTTACTTATATAATTTTTGTAATCATTTGTATTT
GTCATAACAGAAATTTTAAACCATGAATCATTAAATTTGAATGTGAGACTTTTAACAGTTTTAGTAATTTATAGCTGTTTG
TAAAAATTCAAAATATAACATATATTATATGGTTATTGTGGTTGTTTAATTTATTTAATAGTTTAAAATTAAACAAATATG
ATAGAAGATACACTATTTTTTATCAAATCTTTATTATTCAAAATCATTAATTACCATATATACTTTAGCCACATTAGGCAA
TTCCGTAAATTTTATTTAAAGAAATAATAAAGTATATTAATGATGAATTTATTGTTAGTTTAATAAAAAGCTTATTATATA
ATTAGATGGACCAACATATTTTTCTAATGATTATAAGAATCATTCTGGTGATGACACGTAACTACAAAAAGAATTATAATT
TTTCT***CAGATAATATATA***G***GGAATAC***

***BoN-hAT9*: 588 bp non-autonomous hAT in *Brassica oleracea* (AC240081.1) from 66529-67116 bp**

CCTAGTGTTAGGCCTGGG***AC***GGATCGGGTATCCGGACAATTTTAAGATATCCGGATCCGGATCCTTATCCGGCGGATTCAT
AATTTTACTATCCTTATCCGGATCCGGGGTTCGCGGATATCCGGGTGTCGGATATCCTTCTAAAAATTATAATATCCGGCG
GATATTCGGATCCGGATTTGGATCCTTAAAATAAATAAAAAAATAATATTAATATATATAAAATATTAACAATAATTTAAA
AATAAAAATATATAGAATGTTTTTAATTATTTCTATGTATAATATTACAAAATTTACATAAAATTTATATATACTATTATA
AAAAATGAAAATATATTAAATAAAATTAGTTTTTATATATAGATATTACTATTTTTGAAATAGTTATTAATAAAATTTACG
GATCCGGATATCCGGACTAAAAAATCAAGATATCCGGATCCGGATTCGGCTTTGACGGATCCAACATTTTACTATCCGGAT
CCGGATTCGGTCTCTCCGGATATCCGGATTTTCGGATCGGATCCGGATCGGATAACGGATCGAATCCGGATCTCGGATAAA
A***GT***T***CCAGGCCTA***CCTAGTGT

***BoN-hAT10:* 570 bp non-autonomous hAT in *Brassica oleracea* (AC240081.1) from 71911-72481 bp**

CTAATAAC***CCC***G***GTTC***GG***AAAAC***CCCCAGGCGCTAGTCGGGCGATAACCCCAGGTCTAGCGAGTTACCCAAAAATCAGGGA
GTATTCGGGGATTACGCGGAGCCTAGTTTTAAATTTAATTTAATATATTTATAATATATTTAGCTAATTTTAAACATGATG
ACACAAATTCTTAGACATAATTATATAATTTTCTATATATAATTTTAAACAGTCTCTTTTTGATTCGTAAATGGTGGGTTT
GGTACGAAAAAATCCCTAAATATAGTATGTATGTGTTTGTTTTGGGTTTGATAGCGAATTATAATTATTTAGTAATGAATA
ATTACATAAAATCTGTCAATGATGAGTAAAAAATAATCAACCGTGTTTTAACTTACACGGACGGAAAAAAAAAACTTAGGC
GGTCAACGTGGGATTAGGCGGGAATTAGGTGGTCAACGCGGACATTAGGCGGTCTTACGCGGATCAACGCGGGTCAACATC
TGGCTACACCGATTTGGGCATAATCGCCGCGGTCAACCTCTGAACAACGCCTAGACGGCCGC***GTTTT***TG***AAC***A***GGG***CTAAT
AAC

***BoN-hAT11-1*: 724 bp non-autonomous hAT in *Brassica oleracea* (AC240081.1) from 87115-87838 bp**

TAAAAATG***TAG***GG***GTGGG***TGTTCGGATACCCATTCGGGTTCGGATCGGGTATTTCGGATTTTTGGGTATTTCGGTACAGGG
ATATAGAACCCGTTCAGGTATTTCTATACTTCGGATCGGGTTCAGGTATTTTTAGTTCGGGTAGTTCGGGTTCAGTTATTT
TGAATCGGGTTTTGGATATTAAGATTCTAAAAGAAAAATAAAATAAAATTTTCATTTTTAAGTTTCTTGTATTTAAAAGTA
TAGATTTCACTTAACTGATTTATTTTTATTTTTTATATATTGAATGATTAATAGATTTGGAGTACCATCTCAAAAACGAAT
AGATATTAATTTGGCTATTGTTTTTAAACTTTGGACGTACTTTTTGTTAATACAAAAACAAAAAGTTTGATATACATTTTA
AATGAATATCAAATCATTTTCTCCATAATTATAGATATACTATATGATTTTGAAGTATGTGTAACATTAATATAAATATTT
TAAATAAAATGAGAGATGTAAACTAGAAATATAAAGGTAAATATACATATGTTCGGTTATCTTCGGATATCCATTCGGGTT
AGGATATTAACCGTTCGGGTTCGTATATCCAATCTCTCCTAACTTAATACATGTTCGGGTATTTTGCTACTTCGGTTCAGA
TTTCGATTCAGATTTTTCGGGTCGGGTTCGGATGCGGCTTCGGATATCGGGTAAAGTG***CCCAC***GT***CTA***TACAAATG

***BoN-hAT12:* 924 bp non-autonomous *Brassica oleracea* (AC240089.1) from 8543-9466 bp**

CCTACTCT***TAGGGCCGTTCAATATGG***TAGAACCGTACCGAACTGAACCGAAATAGACAATATGGTTTGGTTTTGGTATATA
CCATATAAACCGAATGAATATAATTTTATAAAAACCGTAGGATTTGGATATGGTTTGGTATATAACCGATTAAATCGAATA
AACCGAACAAAACCGATTAAAAGTAGAAACATGTAAATATGTATCTATTTTATAACAAACATGAAAATCTATTTGTTACAT
AAGTTAAATTTGTGTTAATAACTATTACCATAATTTTAAAGTAATTTGTAAAACACTTAAACTATAATTAAATAACAATAC
ATCGCGATTCAGACATCTTATTTTCTAAGTCTATCTTTTGATCTTTTTGCTTTATTTTAGTCTTCACTAAATTAATATGAA
GATTATAAATTTGATGGACAATAATTAATGAAAAATTGTCACAACTTTTTTCTTATCTATAAACAAACAGAGTTTTGTGTT
CAATTAAAAAAGCATGACTTTAATGAACACTAAATATGTAAGAGTGAAAAAACTTTTCTTTCATGTTTCTGTTTTGTTTCA
TATTTTTATTTTCAAAGTTTCAAGCTTTGATTTTAGTTATAGATTTGATTATTTTATTTGATGGTAGAAGCATTTTTACTT
TTTTGTTCATTTATTTGAACATGTAATATATTTTTAATAAATGACTGTATTCACAATATGACTCTAAAATTCATATAATAT
GATCTCAAACTAAATAATTATATTTTTTGGTATAAAACCGAATAAACCAAAAACCGACGGTATATAAACCGAACCGAACCA
AAGTAAATATGGATTTAGAATGGTAGTTATATTTTACTAACCGAAATACCGAAAACCCAAAAAAACCGAACCTAAACCGAA
CCGATAT***CCGGATTGAACACCCCTA***CCTACTCT

***BohAT13*: 629 bp non-autonomous hAT in *Brassica oleracea* (AC240089.1) from 24325-24980**

GCTTAGA*GCATCTCCAA*TGTATTACTCTATTTTCTATTCTAAAATAAAGTAACTCCAAAATGAAGTTGAGTTTTGCTCCAA
TGTATTACTCCGTTTTCTACTCCAAAATAGAATATTTTTAATATATTATGTTTTACGTTAATAAAAACTTATTAATATCAT
CTTTCATTTATACCATTTGCAAAATAGTCAATTTTCATATGTATGATCTTTTATATTTTATATCATATATCTTTTGTATGG
TGCTTATCTTTTATTTTTTATATGTATGGTCTTTTATGTTTTATATTATATATATTTTGTTGCATAAAATAGTTCATTAGA
TGGATATTTATATAATTAAGAAATATTAATAGTAATTTTTATAATATATATAGTAAAATAATATTTATATATTTATTTATA
ATAATATTTTATTAAAAACAAAATATTTAAATATGGAAGACCATTTTATAAATAAAAAATTTCAACTCCATTTTGGAGTAA
TGAGTTGGTTTACTCCATATTTGGAGTAATCATATCTATTACTCCATTTTGGAGTGGGTTTTGGAGTGGGATTATAGATGA
TTTTACTGCAAAATGAAGTTTGGAGTGAGTTTTGGAGTAGGG*TTGGAGATGC*GCTTACTA*GA*

***BoN-hAT15*: 595 bp non-autonomous hAT from *Brassica oleracea* (AC149635) from 52416-53003**

TTTGTAAC*CAGTGTTCTAAAA*TACGGTCTAAGTGCCCGCCTAGGCGGCGTTTAGCCGCTATACGGTAATTTACCGTACTAT
TTTCATGTTATACAGGGTTTTATGCGGATTTCTATAATTCGAATGAATAATAGCGTTTAGTCGGATATTATGCGGTCTACG
CGGACGTCTATGCGGATTTTAAACATTAATATTCAAAAAAATAAACTATTAATATTATTATTATATTTTATTAATGTTATT
ATTTATATTATTTTTACTTATTTCATGTATTTATATGATTGTAGATATTAGTATAATTGTTGCAAAATCATTTTATATTTA
TCATCTTTATATATTTAAAATGGCTTTAATTTTTATAATTTAAATCTATTTATATATGTTAAAATATATACTTTTACATAA
AATTGGATTTAAAATATATTTATATTTAGTTTTTTTCAAAATAATGTATATATATTTATCTTATACTTATATTTCATGTAA
GTATATATCCGTATATACCGCGTTTAGGCATCCGCCTATACGGCTATGCGCTATACAGTCATCTAACGCGTAATAAACGCC
TAGCGCG*TTTTAAAACACTG*TTTGTAAC

CHAPTER 7
POPULATION DYNAMICS OF MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS (MITEs) IN *BRASSICA*

The sequences below are selected from one member (reference) from each family of Stowaway, Tourist and Mutator-like MITEs. The TSDs are shown in red, while blue colour indicate the TIRs

***BrSTOW1-1:* 580 bp Stowaway MITE in *Brassica rapa* (AC155344.1)**

TA*TACCTTTCTGTTCCTAAATATAAGATGTTT*AGCAGGGCCGGTTCAATACTGCTAGAGGCCCTAGGGCAAAAAAAATTTT
TTTACTCTTTTAAATTTATATAGAGATAATGTTTAAAATATTTTTTAAATTTAATCATTAATATTTTGTATATTTTGTAAT
GAAAATAAAACTATATACATATCAAATAATTTTGGGCCCTTTTCAATTTTATTTATTGTATTTATAATTTTTTTATATATA
AAAATATAGATTTATAAAAAATTAGACCCCTTATTTATCATTACACGGGGGAACACGGACTCGGATCCGGGGCGGTCGCAC
CGCTTGTCCCCCCTTATAAGCCGGCTCTGATTTTAGTTAAAAATACACATTATTAAGAAAAAATAATTTTTGTCTAGAAAA
CAACATTAAAACTATAAATTAATACTAATATTTAACCAATTACAAAATAAAGTATTAAATATGATTAGTTGTACAGTTTTT
AATAAAATAAAAGTTATCTAAAAAAATAAAAACATCTTATATATAAACAATAAAATTATTTA*AAACATCCTACATTTAGGA*
*ACAGAGAGAGT*TA

***BoSTOW2-1*: 448 bp Stowaway MITE in *Brassica oleracea* (AC240081.1)**

TA*GGCGCTAGTCG*GGCGGTCGGGTTGGACCTAGCACCTAAAGAGAAAATCGGGGATTAATCGGAAATTATGCGGGGCGGAA
TTTTTAGATGGTTTACTATGTTATAAAACATGTTAATCTTTAATTGTGTATAACATTAATACATTTTCATGTTTAAGATTG
TATAAAACACACAAATAGAATATATAAACTTAATATAGTGTAATTTTCATCAAAATTATGAATATAAATGATATTTATAAA
TTTTTAGATCAAATAAATAAATAAAAATATTATTAAAAATAAAAAATAAAAAATAAAAAATAAAAAATAAATAAAAAAATA
GATTAGGCGGCCGCCTAGGCGGCTAGGCGGTCATTTAGGCGGTTTAGGCGGAAAAAAATCGGATATCCGATTTTTTAAACC
GATTTGGCATAAATCGGGGCGGAAAAGTGA*CGCGTAACGCC*TA

***BoSTOW3-1:* 237 bp Stowaway MITE in *Brassica oleracea* (EU642504.1)**

TA*AGAGCATCTTTACCG*GGGTAGCTTAGGAAGAAGGTGCTTAGAAAAAATTAGTTTAATAGTCGGTGGGCTCCACCCAAAA
TGTATAGCCGCTGCTTGATGCTCCCAGATAAGCGTCGAATTATGTGGATTGTTTGCACTGTTCGCGGGCCTCACCGACACG
TGGCGGTCCGCGATTGGTTTTTTTTTAATCAGATAAAAAAAATAAAAAAATAAGCATG*CGTTAATGATGCTCT*TA

***BoSTOW4-1:* 227 bp Stowaway MITE in *Brassica oleracea* (AC240089.1)**

TA*CTGTTTCCGTTTTACAAAGATATACTTTTT*AGTATTTTTATACATATTAAGAAAACACATTAAACTACCATAATAAATA
TATTGTTTTCTATAATTTTCGATAATTTTTAACCGTCAATTAATTTTATTGAAATTTACAATTTTTTCATAGAAACACACA
AAAATACATCTTTGTGAAACAATTTTTTTTTTA*AAAAAGTCTATCTTAATTAAACGGAAAGAG*TA

***BoSTOW5-1: 2*43 bp Stowaway MITE in *Brassica oleracea* (AC240081.1)**

TA*TTTCTTCCGTTTC*GA*TTTA*GTTGTCGTTGTAAGAAAAAAATTTCGTTTCAGAATAAATGTTGTTTTAGAGTTTTAATAT
AAAATTTATTAATAAGATTCTCCATTTTTTTATTCTATTAATTGAAATATGGTAAAGTGTATTGGTAATTGTGTTTTTATT
TTGAAAATATATAAAATCATATGTTTTATTAATCTATGTGCATAAACCTAGAAAAATAAT*TAAA*AT*GAAACGGAA*A*GAGTA*

BrTOUR1-1: **413 bp Tourist MITE in *Brassica rapa* (AC155344.1)**

TTAGGGGGTGTTAGTGGGATATGGATTTGTAGTGATTGTTAAAATTCTCAAATTCTATTGTTATTGGTTGGTGGATTCTAA
GATTCTTACTAAAATCTAGTGTTATTGGTTTGATGATTGTTTAATGACTTACAAAATCTCTTGTTATTCAAAAAGTTTGGT
TTTTAATGATTCTAAGTTTTCATCAAATCTAGAGTTATTGGGAATTGAATTCTAGATAATTTTACTCATAAAACAAAACTT
CAAAAATCTTGCATATTTCCTCTAGATTCTTAAAATCCTTGAATTAGAACATTTTCATAAACTCTTAAAATATGAAATACT
CTTTATAACTCTTTACAAATTTAACAAATCTCTTGACTTTTAGAATCAACCAACTCTATAAAAATTCCACTCCCACTAACC
CCCCCTTA

BrTOUR2-1: **285 bp Tourist MITE in *Brassica rapa* (AC155344.1)**

TAAAGAGACACCCCCATTAGTGAACTTACCGGAGAATTCACACAGATTCCAAAAAAAAAGAAGTATTAAAATAAGGAATAG
TGATGAACTTGTCTCTCTCCACCTCTTCTTAGTGAACCCTGGTTCATCACTGTAGCGCGGGTTCCACGGCATGTGGCGGCC
CGTGATTGGTCCGAAAAAATAATTTTTTTTTTTTAAAACAAAAATCAAACCAAAAAAAAATATAATAAAAAAATACTTTGTG
AACCCTTTTTATTCATGGGGTTCACTAACGGGGGTGCTCTAA

BrTOUR3-1: **258 bp Tourist MITE in *Brassica rapa* (AC189298.1)**

TTAGGACATCTCCATCCCTACTCCATTTTTTTTTCTAAAATGGAGTAAAAGTGATTATGGAGTAAAGAATACTCCAACCCA
ACTCCATATCTCACTCCATAATGAAGTTTACTCCATAAATGGAGTAATATCTTTTTTGTTTGTTCATCACTCCATTATAGA
GTGAGAAATGGAGTAGGATTGGAGCAATTTTACTTCATTTTCACTTTTACTCCATTTTGGAGGAAAAAATGGTGTTTTACA
TTGGAGATGCTCTTA

BoTOUR4-1: **267 bp Tourist MITE in *Brassica oleracea* (AC240081.1)**

ATAATACTCACTCTGTTTCATAAATGTCATTCTAACTTTTTTTTATGTTACACAAAAAGTGTCACTTTATAATTTCAATAA
AAATTATACTTACTTTCAGCTGAAAATTAATTGCAAATTGCATTGGTTTCATAAATAATTTTATTTATCTCAAATACTATT
GGTCAGAAAAATATAATTAATAACAACTTACATATATTTTCGCTACTTTCTTAGTCTGTGTGAAAAGTGTTAAAATGACAT
TTATTCAGAAACGTAGGGAGTATCATA

BrMuMITE1-1: **551 bp Mutator-like MITE in *Brassica rapa* (CU984545.1)**

TATCCTATTGCGCAATTGTCAATAATAGCACGAAGTTTTTTTTTATGTCTCAAAATAGCACTAGAAGGAGAAAGTCACAAAA
ATGATATTCATTAAAGGGTAAAATATCTCTTATATCCTTGGTTTAAAATTAAATAAACAAACAAAAATAAATAAAAATAAA
TAAAAAAAATGAAAAAAAAGAAATTTTTTTTATAGTTTCAGATTATATGTTTTCAGATTCGATTTTTTTTTTATTTTTTTA
TTTTTTTCGAAATTTTTTTTTTTATTTTTTTTCAAATTTTCTTTTTATAATTTAAAAATACTTTTTGAAACTGTTTTTTTAA
TTTTTATTTTTTATTTTAGTATTTATTTTTTATAAAATTTTAAACCCTAATTCCTAAACCCCCACCCCTTAACTCTAAACC
CTAAGGTTTGGATTAATTAACCCAATGGATATAAGTGTATATTTACCTCTTTAATGAAACCTATTTTTGTGACTTTGAATC
TTGAGTGCTACTTTGGGAACAAAAACTTCGTTTGGTGCTATCCTAGTCTTTTTCTCTATCCTATT

BrMuMITE2-1: **905 bp Mutator-like MITE in *Brassica rapa* (CU984545.1)**

CTTTAGAAACAGAAAAATCGCATAAAAAATTTTGAAAGTGTCACTTATTAGCACTTTATAACTTGAAATTTTTTCACTAAC
ACTTTTAATTTTCAAAGTGACATTTTTATCATAAAAAACCTTCAAAGAAGAAAAATGACCAGCTGAATTGGTTAAAAACAA
TTTTTTTTTCAAAAAATAATTATTTAATTAATAAATAAATAAATAATTTAATAAAAATCGAAAAAATAACAAAAACACTCA
TAAATTAAAAAAATGTGAGAAAAATAAATAAAGATTTAGAAAATTAAATAAAAAATAACTCAAAAATCAAAAATAAATAAG
ATCAAATTAAAATACAGAAAAAATAATAAATTGAAAAAAATGAAAATTCAAAAAAAAAATTCAGTTTCAAATACAATGTCT
GAAATTATCATTGGAGATATGCCACCAACCGTCATTGGAGATATGGCGGAAATGAACGGTTTTTAGCATATCTCCAATGAT
AATTTCTGACATTATATTTGAAAATGAATGTTGATAAAAAAAATATTTGAAAATGAATTTATTTTTGAATTTTCATTTTTT
TGAAATTTATGATTTTTTTTTTATTTTTCTCTATTTTAATTTGATCTTATTTATTTTTTCTATTTTTTATTTATTTTTTAA
TCTTTATTTAGTGTTCTTATTTTTGGCTAAATTTGTTGGTTTTTATTAATTTTTTTTTGTTTATTTTATTAATTAAATAGT
TATTTTTTGGAAAAAAAAACTGTTTTTAACCTGTTCAACCGGTATTTTTCTTTTTTGAAGGTTTTTATGATAAAAATGTCA
TTTTGAAGGTTAAAAGTGTTGAAAAAACTTCAAAATTTAAAGTGTTACTAAATGACACTTTGAAGGTTTTTTATGAGATTT
TTCCCTTTTGAAAC

BrMuMITE3-1: **1586 bp in Mutator-like MITE in *Brassica rapa* (AC232530.1)**

CAAAAAAAACGAGAATTCGGCCAAAAAAAACCTGAACTTTGCACGAATTGCCAAAAAAAACATGAACTTTTGGGCTGACCA
AAAAAACACCAAACTTTCATTGACTTTAGAATTAATTAACAATGTTTTTGCTGACTTGCCAATTTAGCACGCCGTCAACAA
ATTTAACAGAAATATTTAACATCGTTTATTGTTGGCGTTAAGTAAAACGACGTCGTTTGCAAATGAGATGAAACGACGTCG
TTTTACATGATTTGAAATTAAAAATATGTAAACCCCTAGATTCGAACCCAGGTTGGTTGGTCAAATGACAAGGTATTTTAC
CACTGGGCTACTGGCACTTTCAATGTACTTACTAACATGTTTATTTTTATTTGATACATGTTAAATATAAAAAATTCTCTA
AAAACTTAATAAGATTTTTAAAATTCCAAAAAAATTAAAAAATAAAGAAGATTTTTATAAAATTAATTTAGAAATTAAAAT
TAAAAATAAAATTTTCTTTTTCTTTTTAAAAAATAAAAACTCTTCCAAAATTTTCTAAAAACTTAAAAAGATCTTTAAAAT
TCCAAAAAAATTAAAAAAATAAAGAAGATTTTTATAAAATTAATTTAGAAATTAAAATTAAAAATAAAATTTTCTTTTTCT
TTTTAAAAAATAAAAACTCTTCCAAAATTTTCTAAAAACTTAAAAAGATCTTTAAAATTCCAAAAAATTGAAAAAATAAG
AAATTTTTTATAAAATTAATTTTGAAATTAAAATAGAAAATAAAATTTTATTTTTTCTTTTCAAAAAATAAAAAATCTTCC
AAACTTTTCAATTTTTTTAATACATTTGCTAAAAGTTGAAATTAATTATACACACATAAAAAGTTGATAATTGTAACTTTA
ATTTTTTGCATTTTATTATAATTTTTAAAAAAATTGGATTTTTTTTAAATGAAAAATTTGGAAATTTTTTGATTTTTTAAT

TTCAAAATTAAAGATAAAGTTTTTATTTTTTAATTTCAATAAAATTTTAAAGATCTTTTTAAGTTTTTAGAAAATTTTGGA
AGAGTTTTTATTTTTTAAAAAGAAAAATAAAATTTTATTTTTTAATTTTAATTTCTAAATTAATTTTATAAAAATCTTCTT
TATTTTTTTAATTTTATTGGAATTTTAAAGATCTTATTAAGTTTTTAGAGAATTTTTTACATTTAATATGTACCAAATAAA
AGTAAACATGTTAGTAAGTACATTGAAAGTGTGAGTAGCCCAGTGGTAAAAGACCTTGCCATTTGACCAACCAACCTGGGT
TCGAATCCAGGGGTCTACTTATTTTTAATTTCAAATCCTGTAAAATGACGTCGTTTTCAAATGAGATGAAACGACGTCGTT
TCACTTAACGTCAACATTTAACGTCTAGTTAACACTGTGTTAACTGAGAAGTCACGGCGTGCTAAATTGGCAAGTCAATCC
TAAGTTTATTAATAAACTAAAAAAGTCAACAAAAGTAAATGATTTTTTTGGCCAGGCTCGAGTACAGGTTTTCTTTGGCAA
***TTCGTGAAAAGTTTGTGGTTTTTTTGGTCGAATTCTC*CAAAAAAAAA**

BoMuMITE4-1: **899 bp Mutator-like MITE in *Brassica oleracea* (AC149635.1)**

TATATATAT*GAGAAATTCTTGGGTTCACCCCCTAGGTGAACCTCTAGATTCACCAACCAATAGTGTTTGAGTATTTGATAT*
TTGATATCTTTTAAAAAAGGAAACAAAATTGAATTTCCAAATAAGATTATATTTTTGAAATAAAACAATAAAAATACATAA
AAATAGTTACAAAAAATAAATAAATAAATATTGATAAACTTTTAGCAAAATACTAAATCCTATACCCTAAATCCTAAACTC
CAAACTCTAAATTATAAACCTTAAATCTTGGATAAACCGTAAACCATTAGAAAATTTTAAATTCTAAATCATACATTAAAA
ACTAAATCTTAATAACACTAAACCCTAAACCCTAATCACTAAACCCTAAACCCTTGGATAAACCCTGAACCCTTGGATAAA
TCATAAACTCT*AAATCAAAAATATTTAAAATTAAACCCT*AGAGTTTATGATTTATCCAAGGGTTCAGAGTTTACCCAAGGG
TTTAGGGTTTACCCAAGGGTTTAGGGTTTACCCAAGGGTTTAGGGTTTAGTGATTAGGATTTAGGGTTTAGTGTTAGTAAA
ATTTAGTTTTTAATGTATGATTTAGGGTTTAAGATTTTCCAACGGTTTAGAGTTTATCCAAAGTTTAAGGTTTAACGTTTA
GGGTTTAGGGTTTAGGATTTAGGGTATATGGTTTAGTATTTTGCTGAAGATTTAACAATATTAATTAATTTATTTTTTGTA
GCTATTTTTATGTATTTTTATTATTTTATTTTAAAAATATAATCTAATTTGGATATTCAATTTTATTTCCTTTTTTAAAAG
***ATATCAAATATCAAATACTCAAACACTATTGGTTGGTGAACCTAAAGGTTCACCCTAGGGGGTGAACCCAAGAATTTCTC*T**
ATATATAT

***BrMuMITE5-1*: 1159 bp Mutator-like in *Brassica rapa* (AC155344.1)**

TTTATTagaAG***TAATAAACCCGGGTTCAAACGTCGGTATTTGTCAAATTTCGTGTTTTTTCCATTAATGAAGTGTGGGTGA***
AAAAACTTTTTACATGCTCGTCATCAACTTCAAAGATATATCGATCCCTCCATGGATCTAGCTAGACTCGATCTTATGTAC
AAAAGCTAATCGAACAAAATTTCAACTATCTATACACTTTTTTTTTTTTTTGCTGTTGTAACTTTCCATTCCAATATCAAG
CTCCACGTCACGTCAATATTTGGTAGTGAGCATCCCGCTCGTGAAAACATGTGCTGCTAGTAACATCCGCTGGATTGGATT
***TGGTCGCCGGACTGGATTGGCTGGCCGAACTGGATTGGG*TCGCCGGATTGGATTGGGTCGCCTG*ACTGGAT*TGGCTCGGCG**
GATTTAATTGGGTCGCCTGACCGGATTGGATCGCCTGACTGGATTGGGTGGCCTGATTGACAATCAACGGAAGTCGATCTT
CTGGTTTATTCTCAATTTTTGAGAGAAAAAACTGCATGTTTAATCGAAGAAAAGACTTGAATACGCACAGAAACACACAAG
TCGAGTGCACAAAGAAATGCCTGATCCTTTTCTGATTCTCGTTTATCATGAGCCCGTACCAGGCGCTTCGCAAAGCTCGAT
CACTTCTTCGAAGTCGTAAAGAAAGTCCAGAGTCAGTAGAAAGAAGATGAAACTGACTGTGGCGTGGATTAAACAAGGATC
AGTTATGAATCTTCAAGAAGGAGCCGAGGTAGTTCAAGCCGTAGTAACATGCAATGGCGCCCCCTCCTTTCG***CCCGATCCA***
GTTCGGCCAACCAATCCAGTCCGGCGACCCAATCCAATCCGGCTGATGATACTAGCGGCACATGTTTTCACGGAGCGAGGA
TACTCACTACAAAATAGTGACGTGGCAGTGGAGTTTGATACTGGAATGGAAAGTTACAACAGAAAAAAAATTGTATAGATA
GTTTAAATTTTGTTCGATTAGCTTTTGTATATAGGATCGAGTCTAGCTAGATCCATGGAGGGATCAATAATTCTTTGAGAT
TGATGACGAGCATGTAAAAAGTCTTTTCACCCCACACTTCATTAATAAAAATAACTCGAAATTTGACAAGTACTGAGTTTC
***GAACCCGGGTTTATTA*TTTcTTttc**

CHAPTER 8
THE LTR RETROTRANSPOSON LANDSCAPE IN *MUSA* GENOMES

In this section the LTR retrotransposons characterized from *Musa* genomes are represented. From a long list of retrotransposons identified in present study (Table 8.1), only one to two representatives from each superfamily are given below. The TSDs are shown in red, while TIRs are represented by blue colour. The red colour in internal regions is representing RT. The green bold sequences downstream to 5′ and upstream to and 3′ LTRs indicate PBS and PPT respectively.

MaGYP1: **4982 bp Gypsy retrotransposon in *Musa acuminata* (AC226032.1) from 65772-70752 bp.**

CCCGG*TGATAGGACCCTAGCAGGGTTGGGTCTTACTCAACTAAGAAGTAGCTAAAAAAACCCTGTGAGATTGTAATAAACC*
CCACCTAGCCACTTCTATCCTATAAAGAAGAGTGGCTAGGATTTATATTTGGGCCTTTGGGCTTCCTAGCCGTTGCCCCTT
TTATTGGGCTGGTTGGTTAGGGTTGAGGGCAAAAAAGGTAACTCACTTAGGAAGCAGCCCCTAAGTCTAGGGTTAGGAGGT
TATTTCTTAGAGAAGTATTAGGAGTTGTAAAGGAATATAAGTCTTGAATAGGAGTCCTATAAGGAGTTGGCTAAAAGTCCT
ATTAGGAGTTAGGGTTTAGAAGTCCTATAAATAGCCATGTATTCATCCTCTTTTCTTAAGCAATAGATGAATCTTTTCTAC
AGCCTTTGAGCGGTAATTTGGAGGAAGGAACCCCTATAGAGTTCTAAGGAGGTCGATCCCCTAAAGAGATCAACCCCAAGC
***TTAGAATCTGCAAAGGTTCTAACA*TCTAG**TATCAGAGCAGCGATCTT**ACTGCTTTTGCTGCCATCCCTACCACCCACCATC
AGCCAAATAATTTTTACAATTGCTCTTGACAACACCATCTCTATTGTCATCTTCTGCCGTAGATCAAATCAATCTACTACC
GTCACCTCCAATTGCATTAGAGATCTGGTATTCTCCTATCACTAATACCTTTTGCTGCTGTAATTTTTCATTTAAAATACC
AAGAAGAAGTCTTCTTATCTACCTCGTGCTGCGAATTCCTTTCGTCGCAAAAGTTATCATCTTTGCGAACGAAGGATTGCT
ACAAAAAATTTCAACCGCATATTGTTACCTTAACCAATCTATTTGCAATCCTTTTTGAATGAAATTTCTTATGCACTTCAT
ATATCATCTTGGCTTTCATTTAAGATTAATCTATTGAGGAATTCATCTATAAAAACATTCTTGTGTTGCTGTAAATTTTCA
TCATAAACCGTTGAAATTTTTCGATGAGAATTCTGCTATCGCAACTTGCACCAATTTCCTTGCAGTTATACTACTGAAATT
TTCTTGATTAAATTGGTCATCTTTGCTGTTATATAAAATGAGATTTTGCTGTCAAATTTTCTCCTCAAATCTCTCGGAGTT
TGCTGAAAATTTCATTGAGAAACTGTTGTTTCTTCAACGATAATTCTGCTCTTATAAGACAACATATATCTGTAGCAACTT

CCACTGATTTATATCAAACCTTTGCTACTATCTACCACAGATCTAAAACTTTTTTCCCTAATCTTCATCCACTACTATCAT
AGCCCTAAAACTCTATTAAATCATACCCTCAAACTGCACCTAAAACAACCATAAAACAAGCCTAAACCGCAGCCTACACCC
ACCAACTCTTTCGATCACCACTTCTCTTAACCTATACATGCCTTTAACCCGACAACAAAAGAAAGATCTTAATATTACAGA
TTTGGAGGCATATACTATGGCATCTAAGGAGGCAATTAATGATAAATTCAAAGCCTTCGAGGCACGAATGGAGGATAAGAT
TCAGACTCTTTTTATCGACCACCAAACCCGAAGAAATCACATCAAGGAGAGAGCTCTGACCAATCACATCAAGCCCGAAGA
GATGACTTCCAACAGAGGGGAGGCTCTATGACCGATCCCTACTATCCATGCATGAGGGTGGATTTCCCTATATGGGAAGAA
GGAGACCCAATTGGTTGGATCTCATGCACGGAGCAATATTTTTGGTGCCACAAAACCACGGATGCATCTATGGTGGAAATT
ATGGCTATACATCTTAAAGGGGATAATATAAAATGGTTTAACTGATTTGAACATACTCATGGAGTCCTCTCATGGTGACAA
TTCGAAGAAGGACTGTTGATCCGCTTCGGACCAACCGATTACGAGAACATTGACGGATAGCTAGCAAAGATATGAAAAACC
TCCACCATTCAGGAGTACCAAACCAGGTTTGAAAGGTTATCTAATCAAACTCATGATTGGTTTAAAAAATAGCTATTAAGG
ACCTTTATTGAGGGCTTGAAGCTAGAGATCCGGGGAGAAGTTAAAGCGCAACAACTGTACACGCTTATGGTAGCCATCTCT
TTCGCACGACTTCAAGAGGAGCGATTGAACCATGAAGTCCGGAGGACTAGAGTCGCTCCCCGACCAGCAATACCAAAGCCC
CTAGCCCCCCCTACTATTAACTAAGCCCCTGCACCAAAAAGGTTGACAAGAAAAGAGTTTCAGGAGCGATCTGCGAAGGGG
TTATGTTGGCATTACGATAAGCTGTGGAGCCACAAGCATCGCTGTAAAAAAGGGAGACTTCTTATGATTGAACTAATAGAA
GAAGAGGTCATTGAATATCCAAAAGAGAGCCTTAAACATAAAGAAGAAAATGTGGAAGAAGAGCCATAACTGACTGACTTT
ACGGTACACACACTAGCCGGCTACTCAACCCGCAAACGATGAAAGTAGGAGGCATTCTCAAACAACAATCGATCACTATTC
TTATTGACACAAGCAGCACTAATAACTTTCTAAATAGTAAGGTTACTGCTCGGATGACGCTCTACATCAAAGGTTACAGCA
AGTTCGACGTAAAGGTCACCGATGGAAGAATCCTAAAGTGCGACCAAAGATGTCCGTAGGTGAAACTATTACTGCAAGACC
AAGAAATTATCATCGATTTCTTCCTCCTACCAATTGATGATTATAAGGCCATGCTTAGCATAGAATGGCTGACGATGCTAG
GTGATGTCTCTTGAAATTTTTTTAAATTAATTATGAAATTCTATTGCAAAGGCAAACATATCATCCTACGTAGGAAGCGCG
AAATCAACGTAACCACCATTTCGACCTAACGAATAGAGAAAGTTTTACACAAGGTAAATGGTGGCTTTTTTATGCATCTCC
AGTAGCAACCATAGGGGAAAGAAAATATTATTGCAGATGCACTCTCACAACTACCTAAGCAAGCTAAGTTTTCAATCATCT
CCTTTCCTACCACCAGCTTCCTCGAGGACACTAGGGAAGAATGGAAGAAGGATCCAGAGATCAGCAACATCATAAAAAAAT
TAAAAGATCCAAGTGCAATAGCCCACTACACTTGGGATTCGAGGGATCTATGCTACAAAGGTCGCATTGTGCTTATACTTG
ACCCCCCTTGCATCAAAATCATCCTGCATGAAATGCACTCCACACCTTCAGTAGGGCACTCCGGATTCCTGAGAACCTACA
AAAGAGTGAAGAAAAATTTTTATTGGAGAGGGATGAAAAAAATTATTGCTGAATATGTGACATAATGTGATATATGTCAAC
AGTACAAGGGTGAAACAATGGCAAGTCCAGGGAAGCTACAACTACTACCCATACCGGACTTGGTGTGGACTGACATCTCC**A**
TGGACTTCATCGAAGGGCTGCCACCTTCCAAAGGTAAAAGCATAATTCTCATGGTGGTTGATCAACTTACAAAATATGCTC
ATTTTTGTGTTGTGCGAAATCCCTACACTACTACTAGTATTGCCTAGATTTTTATAGAAAATATTATTCAACTGCATGGGA
TGCCAAGGTCCATTGTAAGTGATCGTGAAAAGATCTTCACAAGTAAATTTTGGATCGAGTTATTTCAATTACAAGGCACCA
AAATCAAGATGAGTACAGCGTATCATCCACAAATTGATGGCCAAACAAAGGTAGTAAACCAGTGTTTAGAGACATACCTTT
GGTGTTTTGCTAGTGACCGACCAAAGGAGCGGGCGAAATAGCTTCCTTGGGCCAAATGGTAGTATAATATTACATATCATT
TATCTACAAAATGTGCCCCTTATGGAGCGCTATATGGTCGGCCGACCCTTGTGATTCCAAAGTATGCAATTGGCTCGGCCA
AGGTAGATCAGGTTGACCAAGAATTGATTGATAGGGACAAACTCCTACAACTGTTAAAGGATAATCTCTCTATCGCTCAAG
CCAAAATAAAGCAGTAAGCCAACACACGACGAAGCGAAAGAGAATTTTCAATAGGAGATTAGGTATCTTCGCCTACAGCCA
TACAAAGAACTCTCCATCAACACTCGAGCCTCCATGAAGCTATCCCCATATTTTTATGGGCCCTATCAGATCACAGAGCGT
ATTAGAGCCATGACGTACAGACTTAAATTGCTTGAGGATGCCAAAATTCACTTTGTCTTCCACGTGTCATGCTTGAAGCCC
AAGTTGGGATAACACGAGTCAAGCCAAATCCAACTGCCCAATACCACTGAGGATGGAGTTATTTAAGCCCAACCACAAGCC
ATCATCGACCGTAGGATCATGATACATCGTCGACATCCCTCTACTGAAGTACTAGTGCATTGGAATAATTTACCATTTGAA
GACGCCACATGGGAATCAT**ATGAGGAGCTGAAGA**TCCAGTTCTCTGAGTTCATGGAATCTCAGCTTTGAGGACAAGGCTGA
TTTGGAGAGGGCGAGCT***TGATAGGACCCTAGCAGGGTTGGGTCATACTCAACCAAGAAGTAGCTAATAAAACCCTGTGAGA***
TTGTAATAAACCCCACCTAGCCACCTCTAGCTTATAAAGAAGAGTGGCTAGGGTTTATATTTGGGCATTTGGGCTTCCTAG
CCGCCGTCCCTTTTACTGGGCTGGTTGGTTAGGGTTGAGGGCAAAAGGGGTAACTCACTTAGGAAGCAGCCCCTAGGGCTA
GGGTTAGGAGATTATTTCTTAGAGAAGCATTAGGAGTTGTAAAAGAATAGGAGTCTTGAGTAGGAGTCCTATTAGGAGTTG
GTTAGAAGTCCTATTATGAGTTAGGGTTTATAAGCCCTATAAATAGCCATATATTCATCCTCTTTTCTTAAACAATAGATG
AATCTTTTTTATAGCCTTTGAGCAATAATTTGGAGGAAGGAACCCCTATAGAGTTCCAAAGAGGCCAATCCCCTAAAGAGA
***TTAACCCTAAGCTTATAATCTGCAAGGATTCTAACA*CCCGG**

***MaGYP13*. 5418 bp Gypsy element *Musa acuminata* (AC226048.1) from 123405-128822 bp**

AAACT*TTTTTTTTTTGGTAAGTAAAAGTAAATTGATTATGTGTAAAGTTTCTACAAATAGCTTTATCTATACAAGTCATAG*
ACAAAGCCAAGTAACTCATAGAGTGCGAATGCGATAGTTATGACTACACATGCTGTAGACTATAACTACCTGTGGGTACAT
TATTTGGCATCATTCCCAAGATCTTTTGCTAATAGCAAAATTAAATTTGATGAAAATATCTTTTTCTCTTCGGGTGAAGTG
CTCGGTTTTGAGATCATGCTCCAAACAGGATCTTGAGTCTCTAATAATCTCTTTCAAGAGCTGAGCATTGTTTGTGTGTTG
ACATTCAAAGATTCTATTACATCTCTCCTTCCACATCCACCAAATAGCGCATCTGAAAGTGACCTTTGCCAGTTGTGTTAG
AGGCCCTTTTCTTGAGAAACATTGCATAAGGTCGCCTAGTTCTTGGTGCAAGTTTGGTTGCGGCAGTCTATTGAGTTCAAA
TCGCTGGAGAATCTCCTTCCATATCCATTTTGTATAATCACAAGCAAAATAGAGGTGGTCAACAGTTTCTTCTTGTTGCAA
ACAAAGGGAGCATCGGTTAACATGATAGATACCTCTTCTTTTCATATTGTCTATTGTAGGTAGTTTATTGAGAAGGGCCTG
CCATGTACATAAGGAGTGTCTTTGTGATCCCGGTCGATCCCATGTCCAACTGCTTGGGACCCACTTGTTATTTCTTTGCCT
AATTTGATTCCATGCAGATGTGGCACTAAATTTGCCATTGTCATGAGGCCAAATAAGTATATCTGAAGACTTGTTCCTGAT
TGGAATAGCTATGATAGTCGGCCAAAGTGATAACATTTCGGGAGAAATAGGATTAGGGAGACACCAAGTACCGTTAGCAAT
AAACTCGGAAACCTTCCAATCTTTGGGAGCCCCAAGATCTTTTCGTATTCTATCCCCATATAATTGAAATATACTCTTCCC
ATTCACCCAAGGATCGTACCACATATTAGTACTAGTTCCAGAGGAGATTGCATAAGAGATGTGTTTGATTAGCCAATTCCT
***TGCCTTTAAAATTC*C**TTTCCAA**GCTGCAGAGTGGTATGTT**CTTGTTGAAATTTCCCAGATTGATTCCTTTGAAAGATACCT
AGCAGAAACCCATTGAGACCATAAGGAACATTTGTTTTGCAAAAGATCCCACAATTGTTGTACCAAGCATGCTTGATTCCA
ATCTCTAGTATTTCTCAAGTTTAGCCCCCCTTCATCTTTCGGTTTGCAAATGCTATCCCAATTTACCATATGCATTGCCTT
ATCATTTGTGTCATTCCAGAAGAAGTTCATTAATAACCGTTCAATATCTTGGAGAAGTCCAGCTGGCAGAAGAAATGCATT
ACACCAATATATAGAGTAGGAGTGCAATACAGAACGGATAAGTTCGAGTCGTCCTGCTTTTGATAGAAGCCTATTTTTCCA

AGAAGAAATTCTATTCTTTATTTTTTGCAGGATAGGCTGACAGTGGGTTTTGTGAATGCCTGTGGATA**TGAGCGGGAGACC**
CAAATAAGGAACTGGCAGGCTTCCCTCACTAACCCCCAAAGTTTGTGCAATAAAAACTTTATCCTCAACGTAAGGGCTCAT
ATATACTTTACTTTTCGCATGATTTAATTGAAGTCCAGAAGCCATACCAAAGTCATTCATAATAGTAGCCAGGTTCTTTAC
TGAAGTTGGATCATTTTGGAGAAAGATAAGTAAATCATCTGCAAATGTTATATGTGATATATGAATAGAGCCAGCATTGGG
TACCTTAATGCTGCCATTTAAGACTGCATGTTCCAACATACAGCTAAGTCCTTCCATCGCAATAGTAAAGAGATACGGAGA
GAGTGGATCTCCTTGTCGGATGCCATTCGTGCTCCCAAAAAAACCAATAGGTGTGCCATTAAAGAGTATCGAAAACTTTGG
AGTTTCCAGGCAAGATTGAATCCAATTAACCCATTGCTGTGGGAAGTTCATATCGAGGAGCATCTTATAGATAAATTTCCT
ATTAACTGAATCAAAGGCTTTCCGTAAATCAGCTTTAAAACATATTTTTGTCCCTCTTGCTTTTGAATGCAAATCTTTGAC
CAAGTCATTAGCAAGCAAAATGTTATGATGTATACTTCTCCCTTTGATGAAAGCAGCTTGATTTGGACTAATGATCTTGTG
TATTACCTTTTGCATTCTATTTGCCAATACTTTGGAGATGATCTTGTAAATAAAGTTGCATAATGATATGGGTCGATAGTT
CTCTAGTGAATCAGCCCCTGAATTCTTCGGAATTAAAGAAATAAAAGTGCAATTCATCTCTCTAAGAATATGACCTGAAGA
GAAAAAATGTTGGCAAGCCTTGATCACATCATTTCCAATAATAGACCAAGCAGTTTGATAAAATTCAGCTGGAAATCCATC
TGGACCTGGGCTTTTGTGCTTTGGTGACTTAAAGACAACACATACAATTTCGTCGTTTGTAATTGGGGCATTGAGGATTGA
AATAGCTTCCGAATCAAGTTTTCTAGTCGGCTCCACATGAATATTATTAGTTTTCCCAGCTTGATTGAGTAAGGAATAAAA
AAATTGCTCTGTGTGTGCTTTAACCTCATCTAAATTCTCAATAAGATTATCATCTTTGTCTCTGCATTTGCTGGTACGGTT
AACAGCCCTTCGAGTAGCAATTGAAGCATAGAAAAATTTAGTATTGGAATCTCCTAATGTAAGCCAATGCTGTCGAGACTT
CTGCTTTGCAAAACTTTCCTCCTGTTTTAAGGCTTGCATGAAGTTATTTCTGGCTTCTGTCTCTTTGTAGATAATATTCTC
ATCATTTGGCTGAAGTTGCAGCTCATGTTGCAATGACATTAGTTCCTTCCTGCAAAATTGAACTGTGGTGATAATATTGCC
AAAGGTATTTGTATTCCACTCCTTTAGTGCTGCTTTACAGGCTTTGAGCTTTTGACATAAGATGTAAGGGGGGGATCCATG
CACATTAACATTCCAAGCAGATCTGATAACATCAAGAAATTGAGGATGAGTACTCCACATGTTAAAGAATCTAAACGGTTT
CTTGCCTCTTGGTGTCGCTTTGTCTGCGTGTATATACATTAAAGAATGGTCAGAAAACAAAGGAGCACCATACTCAAGAAG
TGAATCTGGATAAACTGTTAACCATTCATTATTAATCAAAGACCTATCAAGTCGAGCCATTATTTTTCTGTTAGCAGCACC
TTGATTACTCCAGGAGAGCCAATTACCCACAGATTTCATATCAAACAAAGCTGCAGTGTCAATGTAATCATTAAAACTTTG
AAGCTGGCTTTCTGAAAGTTGTCTTCCCCCCTCTTTCTTGTGTAGTTAGCAAAAACTTCGAAAGCAAATGAAAAAGAGGGG
GGAAGTTAGCAAAAACTTCGAAAGCAAATGCAAGATAACTTTCTTGTGTAGGCTCATGATTATTGCTTCCTATTGTAATGT
GCAGGTTGCAAGTCTTGCTTCTTGAGATATTCAAGATAGTGATGTTCAAGAAGACAACTTTTGCTTCTTGAAATAAGCAAG
TTAGCAATCTTTGCTACTTAAGATAGGCAAGCTAGCAATCAAGTTTTTGCATCTTCTTGACTTGTGCAAGCTAGCTATCTC
CCCCTTTGTCATTGTCAAAAAGAAAGGAAGAATACAGTTTTTTAGTTCTTTTTCCTTTTACAACGATTTCAAAATCATGAC
AAAGGATGAATAATCAAGTTATGAATGTCAAGACAATTTTTCATCATGAATATTTTATCATTGCATGCATCATTATCATAA
GTTGAAGTATTACATGCATAATATCATATCATCATTTATAATTATATTAATGTTTCAATATAGCAATTTTTTCTCAAAATG
TGTAACTTAGCACTCATAGCATTTTAAATCATGATTTACATCATATGCAATACATCATGTAGAAAGCATAAGTATTGCATG
CATCATTTTAGCAACAAATCGTAGCGTTTCATCACATGGCAAGATATCATCAAGTAGAATTACGATGCAAGTTCTAACTAG
CAAAAGCTTTCTCCCCTTTGTTTTTGTCGAAAAGAAGGGAGAAATCAATGACCAAAAATTTTTAATTATTATATAGGCCAT
ATTACAAAATCTTGTCATGATATCAAGAAAGTTCAAGAAGGACATGATCCATTCAACAAATCCTTTTAATAGGGCAAAGAA
ATTATATAACCAAGAATCATGATTAAAATATTTCAATATCATAGATCAAGTCATGATTCGTGATTCATCATAAAGCAAGTT
AATTTTCAGTCATCACATAAGGCATTTAAAGAATTTTTGA**AACATAGAAGAATGA**TT***TTTTTTTTTTGGTAAGTAAAAGT***
AAATTGATTATGTGTAAAGTTTCTACAAATAGCTTTATCTATACAAGTCATAGACAAAGCCAAGTAACTCATAGAGTGCGA
ATGCGATAGTTATGACTACACATGCTGTAGACTATAACTACCTGTGGGTACATTATTTGGCATCATTCCCAAGATCTTTTG
CTAATAGCAAAATTAAATTTGATGAAAATATCTTTTTCTCTTCGGGTGAAGTGCTCGGTTTTGAGATCATGCTCCAAACAG
GATCTTGAGTCTCTAATAATCTCTTTCAAGAGCTGAGCATTGTTTGTGTGTTGACATTCAAAGATTCTATTACATCTCTCC
TTCCACATCCACCAAATAGCGCATCTGAAAGTGACCTTTGCCAGTTGTGTTAGAGGCCCTTTTCTTGAGAAACATTGCATA
AGGTCGCTTAGTTCTTGGTGCAAGTTTGGTTGCGGCAGTCTATTGAGTTCAAATCGCTGGAGAATCTCCTTCCATATCCAT
TTTGTATAATCACAAGCAAAATAGAGGTGGTCAACAGTTTCTTCTTGTTGCAAACAAAGGGAGCAGCGGTTAACATGATAG
ATACCTCTTCTTTTCATATTGTCTATTGTAGGTAGTTTATTGAGAAGGGCCTGCCATGTACATAAGGAGTGTCTTTGTGAT
CCCGGTCGATCCCATGTCCAACTGCTTGGGACCCACTTGTTATTTCTTTGCCTAATTTGATTCCATGCGGATGTGGCACTA
AATTTGCCATTGTCATGAGGCCAAATAAGTATATCTGAAGACTTGTTCCTGATTGGAATAGCTATTATAGTCGGCCAAAGT
GATAACATTTCGGGAGAAATAGGATTAGGGAGACACCAAGTACCGTTAGCAATAAACTCGGAAACCTTCCAATCTTTGGGA
GCCCCAAGATCTTTTCGTATTCTGTCCCCATATAATTGAAATATACTCTTCCCATTCACCCAAGGATCGTACCACATATTA
GTACTAGTTCCAGAGGAGATTGCATAAGAGATGTGTTTGATTAGCCAATTCCTTGCCTTTAAAATCA**ATACT**

***MaCOP1*: A 5290 Copia in *Musa acuminata* (AC226035.1) from 79036-84325 bp**

CTGCA*TGTTGGGCTGATGGCCCATATTCAGCCCATGTGGGCTTTATCAGCCCACAGCCTACACCCTCTCTTAACCTAACCC*
TAATTAAGATTAGGGGGGTGTGGTGGCTGCGTTTTAGAGGCAGATTAAGGCTATAAAAAGGCAGCAACGAGGCAGATCTTT
GAGGACACGGGATTCCAAAGAGAAGAAGGAGATCAAGGCAGAAAAGGAAGAGAAAGAAAGGGAAGAAGACAAGGACAACGC
AGAGAGACTGTTCACAATCATCTAGCAGTGTTCTCATCTCAGGTTAGATCAAATCTACAGTAGCCTCTTGCTGTGATTACT
TGGGGAGGTTTTAGATATTGTGGGCAGTGACGTGATCCTTGTATCCCAGTTATTCTCTTGTGGTTGTTGCTAGGGTTTTGG
GCAAGAGATTGAGATTTGTATATTCATTATTCTCATAGTGGATTATCTCTAGTTTGCCCCGTGGTTTTTACCCTTCACATT
GAAGGGGTTTTCCACGTATATCTTGGTGTTATGTTTGATTGTGTTTCCATTTTATTCCGCTGCGTATTTTGGTCTTCTAGT
ATTTGTTCCTATACAAAGGTTATTCCTCTTTTTATCCACATCAACTGG**TATCAGAGCGGGGTTTTG**GTGATTTAATTTTTG
TATTTGAACATGGAGGCCAGTAATATTTCTCGCATGATTAGTTTAAATGGAAATAATTGGATGATATGGAAACCAAGAATG
GAAGATCTCTTGTATTGCAAAGATTTGTATGGACCTTTGCAGGGGGATAGTGCAAAACCTACAACTATGACAGATGATGAG
TGGAAGAGGTTAGATCGAAAAACAATTGGGTTTATTAGATAGTGGCTTGATGATAGTGTCTTTCACCATGTTTCTACTGAA
ATTTCTGCATATTCTCTTTGGAAAAAATTGGAAAGTCTCTATGAGAGAAAAACAGCTGGCAACAAAGCTTTTTTGATCAGA
AAACTTGTGAACCTAAAATATAGAGAGGGTGCTTCTATTGCTGAGCATTTGAATGAAATGCAGAGTATTACTAACCAGTTA
TCCTCTATGAAAATGTCTCTTGATGATGAGTTGCAGGCATTGTTACTTCTCAATTCATTACCAGAAAGTTGGGAGACACTG
GTGGTTTCCCTCAGTAATTCTGCGCCAGATGGTATTGTCACTATGAGTCAAGTAACAAGCAGTTTGTTGAATGAGGAGTTG
AGAAGAAAGAGTTCAGCAACATCTCAGAATGATTCACAGTCACTTATCTCAGAGAACAGAGGAAGGTCAAAGTTCAGAAGC
AGTTCACGTATGGGTAGGAGCAAGTCAAGATCAAGAAAAGATATTGTTTGCTATAACTGTGGTGAGAAAGGACATTACAAG

AACCAATGTAAGCAACCTAAGAAGAACAAGAAAAAGGGAAAAGAAGTGGAGTCTACAGAATCAAAGGATAATACTACAGCT
ACAGTGCAGGGTGGTGATTATTTGATTTTGTCTCCTTCTGATGATATTTTTTCTTGTGTGTGTCAGGATCTTGAGTGGGTG
ATTGACACAGGTGCTTCTTATCATGCTACACCTCGGAGGGAGTTTTTTGCTACATACAGGTCTGGAAACTTTGGTGTTGTC
AAGATGGGCAACTATGGCACAGCAGACATCATTGGCATGGGTGATATCCATTTAAAGACCAACCTTGGCTGCAAGTTGGTA
CTTAAGGATGTGAGACATGTGGTTGACTTGAGGCTGAATTTAATTTCAGTTGGAAGACTAGATGATGAAGACTATGAAAGC
AGATTTCACAGAGGGCAATGGAAACTCAGTAAGGGTTCTCTTGTTATAGCTAATGGAAAGAAATGTCATACTTTGTACAGG
TTGCAGGCTAAAGCTTATGGTGAGCAGTTAAATGCTACAGAGAAAGACTTCAGTATGGAGTTGTGGCATAGGCGATTGGGA
CACATGAGCGAGAAGGGGCTGCAAACTCTTTCCAAGAGAGAGGTATTACCAGATCTCAGAGGTATACATCTGAACCCTTGT
ATTGATTGTTTAGCTAGTAAACAACATAGAGTTTCATTTGCTAGTGCTGCTTTGTCTAGAAAAATGCATGCCTTAGACCGT
GTTTATACAGATGTATGTGGTCCTTTGAGGACAAAAACTCCTGGTGGATCTGTTGATGTTCTTGGTATAAGTGGTGCACTT
TATTTTGTCACTTTTATAGATGATTTTTCTAGGAAAGTTTGGGCCTATGCTTTGAAGACCAAAGATCAGGTTATTAATGTC
TTTAAAGAGTTTCATGCCAGGGTTGAAAGGGAGACAGAAAGGAAATTGAAATGCATAAGATCAGATAATGGTGGTGAGTAT
ACAGGATTGTTTAATGACTATTGCAGGTCACATGGAATCCAACATGAGATGACAGTTCCTGGTACACCTCAGCATAATGCA
ATTGCAGAGAGGATGAACCGCACCATCATGGAAAAGATCATATGTATGCTTTCACAGGCCAAGCTACCCAAAAGGTTTTGG
GATGAGGCTTTGAGGACTGCAGTTGATGTGATCAACTTATCACCATGTACAGCCCTAGATGGTGATGTTGCAGAGCATGTA
TGGTCAGGGAAAGATGTTTCCTACAGGCATTTGAAAGTGTTTGGTTGTCGTGCATTTGCACATGTTCCAGATAATGAGAGG
TCCAAGCTGGATGGTAAGTCTAAAGAATGTATTTTTCTTGGTTACTCACATGATCAGTTTGGTTACAGGCTTTGGGATCCA
GAAAAGCAGAAGGTGTTCAGGAGCAGAGATGTGATCTTCTTTGAGGATCAAACCTTTGAGGATTTGAAGAAGAAGGCACCA
GCCAAGACTTCTGCAGAAGGATTAGCAGATTGTGACCCAGTTACTCCTCCAGTATATCAGGGTGATGGGGGAGATATGCAG
GAAGATGGTGTAGAACCTGATATTGATCTACCTGTAGGACATGTTGAGCAAGAAGAAGTAGGAGAGCAAGTTCCCGCAGAA
CCTCAGTTGAGAAGATCTTCTAGACAACGTCAACCTTCCAGAAGATACTCTACAGATGAGTATGTGATGCTTACTGATGCA
GGTGAACCAGAGAGTTACCAGGAAGCAGTTGAGAGTGAGCAGAAAGAGAAGTGGTTAGCTGCTATGCAGGAAGAGATGGAT
GCTCTTCAGAAGAACCACACTTATGATTTGGTGCTGCTACCAAATGGAATGAAG**GCCTTGAAGAACAAGTGGGTTTTTAGG**
TTGAAGACTCAAGAATATTGTTCTCAACCAAAGTACAAAGCTAGATTGGTTGTGAAAGGCTTTGGTCAAAAGAAAGGTATT
GACTTTGAAGAGATATTTTCTCCTGTTGTTAAAATGTCTTCTATTCGTGTTGCTCTTGGTATTGCTGCTAGCCAGGACTTG
GAGGTTGAGCAGTTAGATGTGAAGACAGCTTTCCTTCATGGTGATTTGGAGGAGGAAATTTATATGGAGCAACCAGAAGGC
TTCAAAGTCAAAGGTAAAGATAATTTTGTCTGCAAGTTAAAGAAGAGCTTGTATGGGCTAAAGCAAGCTCGAAGACAGTGG
TACAAAAAGTTTGATTCATTTATGACAGAAAATGGATACAAAAGAACGGCTTCAGATCATTGTGTGTACATCAAATGGTTT
GGTGAGGATTTTATTATTCTCTTACTTTATGTTGATGACACGCTTATTCTTGGGAAAGATATGTCTAAAATTGACAGGTTG
AAGAAGGAAATGAGTGAGTCTTTTGCAATGAAGGACATGGGGCCAGCAAAGCAAATACTAGGCATGCAGATTTCTCGTGAC
AGGAAAAACAAGAAGATTTGGTTGTCACAGGAGAAATACATCGAGAAGGTATTGGAAAGATTCAGTATGAGCAATGCTAAG
CCAGTTGGTTCTCCTCTTGCAGGTCACTTCAAGTTGTGCTCAAAACAGAGTCCGTCAAGTGATGAGGAGAAGGAGAAAATG
CAAAAGGTTCCTTATGCTTCAGTAGTTGGAAGTTTAATGTATGCAATGGTTTGTACGAGGCCAGACATCGCATATGTAGTG
GGTGTTACTAGCAGATTTCTTGCAAATCCAGGCAAAGAGCACTGGGCAGTGGTGAAGTGGATTTTTAGATATCTCAGAGGG
AGCTCTAAGGTTTGTTTAAGCTTTGGAGGTGGACCACCTGTGTTGACAGGTTACACAGATGCAGATATGGCCAGAGATATA
GATACGAGGAAGTCTACTTCAGGTTATGTACTTACTTTTGCAGGGGGAGCTGTGTCATGGCAATCCAGGTTACAAAGGTGT
ATTGCTCTCTCCACCACAGAAACAGAATATATTGCTGCTACAGAGGTATGCAAAGAAATGTTATGGATGAAAGAATTCTTA
CAAGAATTGGGGCTGAAACAGGAAAATTATGTGGTGCATTGTGACAGCCAGAGTGCCATCCATTTGTGTAAGAACCCAATG
TTTCATTCCAAGTCAAAGCATATAGATGTCAGATACCACTGGATTCGAAATGTATTTGAAGAGAAGCAGTTGCAGCTTCAG
AAAATTCATACAGATGACAACGGAGCAGACATGTTGACGAAGACCTTACCAA**AAGAAAGACAGGAGA**TATGCCGACAGTTG
GTCGGCATGGCTTCACATTGAGGAGTCATGGGACAGCCTCCCTTATGGGCTGAAGGGGGAGGT***TGTTGGGCTGATGGCCCA***
TATTCAGCCCATGTGGGCTTTATCAGCCCACAGCCCACACCCTCTCTTAACCTAACCCTAATTAAGATTAGGGGGGTGTGG
TGGCTGCGTTTTAGAGGCAGATTAAGGCTATAAAAAGGCAGCAACGAGGCAGATCTTTGAGGACACGGGATTCCAAAGAGA
AGAAGGAGATCAAGGCAGAAAAGGAAGAGAAAGAAAGGGAAGAAGACAAGGACAACGCAGAGAGACTGTTCACAATCATCT
AGCATTGTTCTCATCTCAGGTTAGATCAAATCTACAGTAGACTCTTGCTGTGATTACTTGGGGAGGTTTTAGATATTGTGG
GCAGTGACGTGATCCTTGTATCCCAGTTATTCCCTTGTGGTTGTTGCTAGGGTTTTGGGCAAGAGATTGAGATTTGTATAT
TCATTATTCTCATAGTGGATTATCTCTAGTTTGCCCCGTGGTTTTTACCCTTCACATTGAAGGGGTTTTCCACGTATATCT
TGGTGTTTTGTTTGATTGTGTTTCCATTTTATTCCGCTGCGTATTTTGGTCTTCTAGTATTTGTTCCTATACAAAGGTTAT
TCCTCTTTTTATCCCCATCA**CTGCA**

MaCOP7: 5012 bp in Musa acuminata **(AC226041.1) from 2059-7070 bp**

ACTAA*TGTTGAAGAGAATAGAATATAGGAATTACTGAAGAAGCTGTTCTCGTATTGACATATTCTCTCCTATTTATACATG*
TTAGGATGGAGAATTTTTCTTAACAGAGTAGAAAGATTTTCTCATATAGTTAGAAAAATCTTATATACTATCATGCCCCCG
CAAGATGGTGCTCTTATCAAGGATACCAATCTTAGATCGATGTAAAGTAAAAAGTTTATGAACGAG**AGGCTTCGTGAGTGA**
GTCGGCTAATTGATCAGTCGTATGTACATGAGAAACTCATAGTTGACGATGGACAACTTGATCTTGCACAAAGTGGAAGTC
GATAGCAATATGTTTCATGCAGGATTAGAACACCGAATTAACATACAGATAGGTAGCTCTAACATTATCACAATATATTGT
AGGAGTAGAATTGATGTTGAGTTCCTTAAGTAGATTTGTAACCCAATTAAGTTCTGTAGTGACAATGGCGATGACACAGTA
TTCAGCTTCAGTTGTAGATCATGCGATTGTCTTTTGCTTCTTAGAACTCTAACTGATTGGATTAACACTAAGGAAGATAAT
ATACCCCAATGTGGATGTTCAATCATCAAAGTTACCTACCCAATTAGCATCAGCAAAGGCATAAAGAAGAAGTGGAGAGTG
TTTATGAAGAAAGAGACCATGATTGAGGGTCCCCTTAAGATATCGTAGGATTCGTTTGACCGTAGACCAATGCATAGTAGA
CGACCTATGCATAAACTATGATAATTTGTTGACTGCAAAGGAGATGTTTGGACGGGTGAAAGCTAAATACTATAAAGAGCC
AACAACTTGTCGATATTGAATGGGTTCCGTAGTAGGACTTCCATCAGATAATTTGAGAGAGCCACTAG**CAAAGAGAGGAGT**
TGTAACTGCATTTGCGTCCTACATGTTTATTTTTGATAATAAATCTTGAATGTACTTTCTTTGTGAGAGAAAGAGACTAAA
AGATGTGAAATATAGCTTCTACACCTAGAAAGTAGTTCAAGGGTCCTAGATCTTTAAGCGAGAATCGATCTGCCAACTATTT
GATGAAGGGTTGGATTTCTACGGGATTGTTGCATATGACAATGATATCATCCACATATGCCAGAATATATATTGTGTCACT
ATGGTGTTGTCGTAAAAATAACGAGGTATCAGATTTGGAGTTGAGAAAGCAATTGAAGTCAAAAAAGAGCCAAGTTCTATG
TACCAAGCTCTTGGAGTCTAACGAAGTCCATAAATAGTTTTTTGTAGTTTACAAACATGCATTGGATATTGAGAATGAGTG
AAACCAAGAGGTTGTTGCATAAAAACATCTTCAATTAGAGTTTCCTGTAAAAAGGCATTATTAACATCCAACTGTCATAAT

TGCCAGCCTGTAGTGGTAGCCAAACTCAGGATAAGTCGGATTGTAGTAGACTTAACAACTGGACTAAATATCTTAGTGAAA
TCAACTCCAAGTCTTTGATGAGGAATGGTCCGCATAGTATGATAAGTATGGTTTTGGAAAGAAAAGATAGACTCAATAAAA
ATAATGTAACGTAATATAAAGATTTTGTGAGTGTGGATATTATAGGAACGGAAAGTATTATGTTCAAGAGAGTAACTTATA
AAGACGCAAGGATTAGATCGTGATGTTAACTTATGAGAGGCATAGAGATGTAACCATGGATAACATAGACAATCGAATACT
CTAAGTTTACGAAGGTTTGAGGGTTTATGAAATAGTGTGTCAAAGGGGGACTGGTATTGTAGGATTGACATAGGCATTCTA
TTAATAAGGTATATGATAGTTTAAAAGACTGTTATCCAAAAGTTTGAAGGCATGGAGTTCTAATGTAGAAGTATGAGCCTA
GTTTCTACTATATATCGATGTTTATGTTTGGCGGAGCCAACTAGCCGAGAAGTATGCAGGGAAGACTTGAGGTGTTGGATA
TCACAAGTTGAGAGATAGGATGCTTGGGCTTGATATTCACCACCGCCATCAAAGTAAACTATTTTGATTGTGGACTGAAAG
TGATTTTTGACCAACTTTCGAAAGATGGGAAAGATAGTGGAAATGTCAGATTTATGGTGAAGAGGATATAACCATGTGAAT
TTAGTGAAGTGATCTATAAAAATGACATAAAATCTGAATTTATCAAAAGACGGGATTGGAGTAAGGCCCCAAATATCGGTA
TAAATAATTTCAAATGGTTTAGAGAAAGAAATGAAAGATGAGCCAAAGGGCTGTCGATGACTTTTATTACTAAGACATGCA
TCACAATGTATTATAGAACTATGTGATTTAAAAATAGGAATAGAGTAACGAGAAAGTAATTTCTGCTGAATAAGGGACGAG
GGATGACCAAGACGATGATACCACACATCAACTGGAGCCGCAATAGAAGAATGAACAGTGGGCTGGGTTATTTATGGAGCT
GACGACCATTCGTAAATATTGCATCTATTCGGGCCCTGGACCAAGGATTTCCCCATGCTCAAGTCCTTAACAAGAAAGGAA
TCAAGAAAGAATTCAATTGAAGTATTGTTATGTTTGCAAAACTGATAAATAAAAATGAGGTTGCGTTTAATATGAGGGGCA
CATAAAACATCATCGCGTGTAAATGTGATAGTATTAGAAGTAAGCATTGTAGAACCAATATGAGTTATAGGAAGTCCTTTA
CCGTCACCGATGATGATATCTTTATCGTCGCCATAGGTGTTGTGAAGAGACAAATTCTGAAGATCAACGGTGATGTGATGG
GAGGCGCTAGAGTCGACAATCCAGTTGGGCTAGCTAGTCATCGGAGTAGTTAGGAGATTTGCCTGAGGCTAGTGTGATGGA
GCATGGAGGCGAGGTTGAGACCGACAAACTTTAGCGGAGTGTCCAATTTTATCACACAACTGGCAAACGATCAGTCTTGTT
GGCTTGGTGGGGCAGGATGCCAAGATGTATGATTATTGGAGTTGTCATTTTATAAGAAGTTATGATTAGGACGATAAGAGG
GGTTTCGCATGGGATCCATAGGATCAAGATGTGTATAGGCCAAACCTTTGGAGATGTTTGGAGGGTACCAAGTGCCCTTCC
TCTTAGACTTGTGACTGACTTGAGCTATGATAGTCAGTCCGGGCAACTTGTCCTCATGCTTCAAATACGTCTTATAATCAG
TCAACTTGTCATAGAGGTCTTCCAATGATACTAGCGTGTCGCGTGCCCGAATTGCATCTGCCAACTCCTTGTAATCGTCTC
CAAGATCATTAAGGGTATGGATGATGATCTCTTCGTCACATAGAGAATGACTTATCAAAGCCAAATCATCGATAATAACTT
TAATATTTTATGGATAATTAGTGATAGTACTTCCCTCTTGCTTTGTCACCATTAGCTTGGATAGAAGATTGAGCTTATGAG
TACGCGAATGATTCACTAAGGTTGTTTACAGTTTAAATCACACTTCGGCAGTAGTCCCACATGAAGATATTAATGGAGTGA
TGGATCCAGCAATTGAGGCTTGAATAGCTTGGAGGATGAGATGATCTTGACGTAGCCACAATTTGTGAGCTAGATTTGGTA
CTGGACTGGGTTCTCCGAGGATGTTGATCATGGCTAGCGGACAACTGAAGGAGCCGTCAACGTAGCCTAGCAATTCATATC
CAAATAAGAGATTAGAAAATTAAGCACGCCATGATGCATAGTTGCCACCCTTTGATAACTTGAAAGGGATTAGCGTGACAG
CATTGATGGAGATAAGTCCTGTAGAAAAAGTACTATGAGTCCCTGTAAAAATAGTAACTTGAATATCAAAAGAAATAACAA
AGACATCCCAAACATTTTTTTTTGTTTGTTTTTTTTTTTTACTAAACAGTGATCTTTACGGAAGATAGAGGAGGGAGAGAG
GAGGCTACTGCTATAGCTGCTGTGGCTACTACGATAGAGAAGACTGCTAGAGGGCAGCACTAGTGGTGCACCAGTGAGAAA
CAGTGCAGCGTTGGTAGCGTTGGAGAGAGCTGTAGTGATGCTGGAGGAAGCTGCAACAATTAGGTAGAGAAATCTGCAGTG
ACTAGGTGGAGAAATCTACAGTAATTATGTGGAGAAATCTACAGCGATCAAGTGGAGAAATCTGCAACGATCAGGTGGATT
GCTGGGATGAGTTCTTTATTTGGTGGATGCTTTTATCTCTTTTCAGCTAGAAGGTTGCTGAGGATTAATGGTGAAGAGATG
TTTTATGCTGTCGAGAGCTGTTCTTGGGGCGGTTGATAATGAGGCAGCGAGCTGTGATCGATGCAGATCATAGCCTGATGA
AGATGATGCTGCCGCCGTGGTGGTTTGAGAGGCACCAGGCAGTGAGAAAGAAGTGATTCAGCAGCAACTGAAGGTCGTGGC
TGGGGTAGCAGCTAAGGCGTCACGCGCTATACACGCAGTAGTCAGGATCAGGGGCGGCGCTCGGAGAGAGAGCGGCTGAAG
GAGTTGGAGAGAGGAAGCCAATAAGCGCTAGAGGATGTTGGGGTTAGAGACAATACCGAAGGTCTAGTAGTCATTGGAGGA
CACCGGCGCAGTCGGCAGCGGTGGCGAAGGTGACCGATGAAGGGGACCAACGAAAGGGAAAGAAGGGCCGGAAGGGGAGAA
AGATGGGGTGGAAGATGCAGCGGAGGAAGGCTTTGGCAGTCCTCACCAGTCCTTTCGGCTTCCACACTAACACCCAGATGA
GCAGCGGAAGGCTTCGGTAGAGTTGCCTACATCCGTAGGGGAGAGCTGGTCGCAAGGAGAGAAGAGGGAAGGAGCAGGCA**G**
GGGTTGGAGAGGGATCCACCGTTAAGGGTGAGCAACCAGCCGCAAACAAACAGGCCTTCGTTGTCGTGGCTGCACCTGATA
CCA***TGTTAGAAGAGAATAGAAGATGAAATTACTGAAGAAGTTGTTCTCGCATTAACATATCCTCTCTTATTTATACAGGTT***
AGGAGGGAGGATTTCCCTCAATAGAGAAATTTACTCGTACAGTTAGAGAAATCTTATATGCTATTA**ACTAG**

***MACVI*: A pararetrovirus-like element in *Musa acuminata* (AC226046.1)**

CTCT*TGGTATCAGAGCTGTTGCTGGCTAGCCATTTTTAAGCCCGAATGGCATGGCTAACCCGAATGGTATCAAAGCTTGTT*
TTTATCTGTGCTATGGCTTTCTAAATTGCTTCAAGATAGTCATGTGGACAGGTTTTGAGGAGCATGGGTTAAATTTTTGTT
AAGTTCCTTCTATCATTTAGAATCAACTCGTGAAGCTTAAATAACCGTGGGGAAGTACTTTCTTTAATCCTGAGAAAGGTT
GTCAAAAGACGTAAGGTATCAAAATTGGGCAGGGAGACCCTAAGAAACCACTGATAAGGTGAGGCAACCTAAGACTAGAAA
AAGGAAAGGAAGACCAATGGTTAGGAATGCCTAGGATAGATGTCTGAAAGATGGGAAGAAGCTATACAACAATGGTATACT
AGCTCCCATACCTCAAAACTAGACTATCTTGACCTTGTTGAAAGTAGTAGCTCTACCCGAAAGGAACTAGCTCACAATCTT
ACTGTCATTTACAACATAACTTGCTTATCAAGTAAAGTACATCTAAAGAACTTCAAAATCCTCATAGAAAAAATTTAATTT
CTCTCGAAAGAGGTTAAGATACTGAGATCCTCTGTAAAAACTTTCTCTGCCTTGTTCTCCGAAAATAAACCTTTAACAAAG
CATGAAGTCAGGGATCTAGTTGAAGAAATATCCAAGCAACCAAAGCTGGTTGAAGAAGAAGCTTTAAAACTAACCCTAAAC
CTCGATCAAAAGCTTCAAAGAGTAGAACAACTCCTAACAAGGATAGAAAAACAAATTTTTGGATGAGCTACTTGTTCACCA
AAAACACCGAATCCTATAAAGAAGTACTCAAAGCTACAAAATCCATTAACTCTCCGTCCCTTGGATTCCTAAAAACCAGTG
ACTATCCCGGAACCCTAAGCCATCAATCAGCTGTAATCAAACAACACAACACTCAGCTACAACTACTCGTACAAATTGCTG
AGGATATAAAAGGGATACGTATAGAACTGCAAACCATATGAGAAATCCAGCAAGCAAAAGCCGGATCATCCCAGGGAATCC
CTGAAGATTTGATTATCAAATCTAAATTTAGGTCCTTCAGAAAGACCTAAAGAGGTGCAAGAAAGAATCTTAGTGTTCAAA
GATCCCATCCAGATTCTTAAGCAAGTCCAACAATGACAACCCAGAGTTCAACCCAAGAAGTCTCCACATCTGCCCCTTTAG
TTGAAGATCAAATCCGAGACTACAGGAGAAACCACAGAAGGCTATTCAATGCTCGCCAGGCAACAAGGCGAATGGGGCAAA
TGCTGCTAGGAGGACCGTCCGCTACCCAGCAAATCCTTGAGCAGCAGATAGACCCTCAGGCCCAACTGAGGCTATCCATGC
GAGAAAGGGCGACGATAGCACCTGCCGAGGTGCTATATCACTCCAGGCGAGATGATGCACACCATCGAGTTTATGTGCACC
GATCCGAAGAGGCAATGTTGGTCACCAATAATCAGGAAGACAGGGCTTTCATCATAGAGGAAAGCTATGACCGACTGCAAA
GGAGCCGCATGCAATACATTCACCTAGGCATACTGCAAGTCAGGATGCAAAGGCTTCTTCGGCAAGAAGAAGGAACACTAG
CACTATTAGTGTTCAGAGATAATAGATGGGCTGACGACAGGTCCATCATAGCCACCATGGAGGTAGACTTGACTCGAGGAA

GTCAGCTGGTCTACATCATTCCTGACATCATGATGACGATTAGGGACTTCTTCCGCAACATTCAAATTTCAATCCTCACTA
GAGGATATGACACATGGCGGAATGGTGAGGCAAACTTGCTAATCACCTGTGGAATAGTAGGGCGACTTTCAAATACCACGA
ATGTCGCCTTTGCATATGAAACATCAGGGGTGGTGGACTACCTAACAAGCCATGGTGTCCGGGCACTTCCCGGAAGGAGGT
ACTCCATAGCTGAGCTACATGGCAGAGACTGGGTCATTAGACCAACCTAGATTGCAATACCAATACAGCCAATAGAAGTAA
GGAGTCGCAACCTTATTGATGGCAGAATATCTATAAGCTTTGACAACTACAAAGCTGCATCTACATCCAATCGGATCAACT
ACAATACCGCTGATGATGAAACATTCAACGATGAAGAAGAAATCCGGAGCCACATAGTAGCTGTCAACATCCAATTATCTG
ATGACAGTGAAAATGAAGCTGAAGAATTACGTAAAAACCTGAATTCCTATTGTCAGGATATTAACGTTTCAGAAGGAGGTG
GGGAGATGCCATATCCCCAAAAATTTCAAAAGGAAGTAATTGCAGCAGGACTCGAGGAAGACCTAGCAATGGAATACCCCC
AACTTGCAAAGTTATCTCAACAGGTATATTCATCCTCTGTTGTATCAAATTACAGACCACCTGCAGATTCAACTATGGGAC
CAGCAAATTACCCCCAGTAGTGAATGTGGAATCCACAAGCTAGAGGCTCGCATATGAAGGCTACTCAAGACATCCAAGGTT
CAAATCAAAGGATTTCTCAGAAGCTTGGAACCTCCCATCAGCCTTCCAACAACAAGGGGCAATGTTTATAATTCCGTCCCA
ACTTGGGATGTTTAATGAAGTATTTATGAGATGGGAATCAATCACTAAAAACTTGGTTTCCCTACAAGGATTCACTGATCC
CCAAGCAAAAATGGAATTCATTGAAAACTTGCTTGGAGATGCAGAGAAATTAGCATGGATCCAATGGCAAATGGCGTACCC
AGAGGAATACCAACTACTAATGGCCAACGCAGACGGAACTGGAGGAACTCAAAATATCCTCTCACAACTAAGGACGATCTT
CATTCTGGAGGATCCGTTCCAGGGTTCAACTGCAGCACAAGAAGAAGCTTACAGAGACCTCGAAAGGTTATCCTGCACAAA
CCTCAAGTACATTATTCAATTTCTAAATGATTACATGAAGCTTGCATCTAAGACCGGAAGGTTGTTTACAAGCCCAGAACT
TTCCGAAAAACTATGGTCTAAAATGCCTGGAGAATTAGGAAAAAGAATCAAAGAGGCATTTAAGACGAAATACAGGGGAAA
TACTATAGGAATAATTCCAAGGATACTTTTCTCCTATAAGTATCTGGAAGCAGAATGCAAAGATGCTACCTTCAGAAGGGC
GTTAAAAGACTTATCCTTCTGCAGTGAGATCCCTTTCCCTGGATATTATAACAAGCCCGAGAGGAAGTATGGCGTGAGAAG
ATCAAAAACCTACAAGGGCAAACCCCACTCGTCTCACGCCAGAATTGAAAAGAGGAAGCACCTAATAAGAAACAAGAAGTG
CAAGTGTTACTTGTGCGGGGAAGAAGGACACTTTGCAAGAGAATGTCCCAATGACCGTAAAAACATCAAGAGAGTCACTAT
GTTTGAACAATTAGACCTTCCTTATGACTATGAAATTTTGTCAGTACAAGAAGGGGAAAACCAGAGCGACGCAATCTACTC
CATCTCTGAAGGAGAAGACGTAGAAGACCTACAACACGGTATTCACTCCTTCACCCACAAAATTTTTGCACTAATAGAAGA
TAGTAGAACTTGGTGGATAGGACCCGAGTCAGGTTACCGGGCCAGAGTTCAAGTCTCCCAAGCACAAGCTGAATGCAGACA
CATATGGGAAATCAACACCGAGTTACCAGCCAATTTGGAGAAGTGCAAATGCTGCAAACGGACATCACAAAGGAGGCACAG
GAGACACTGCCCCTTGTGTAAAATTACTTCATGCGGGATGTGCAGTATCTACTACTTTGATAAAAGAACCCCCGTAATGAC
TGAGGAACCACCAAGGTATGAACCAAGAAACTTGCCTCAGCAGCAACAGGATTATATTAATCACTATGAAGCAGAAATTAG
GAGGCTCGAGACTGCAGTAGAATCTGAAC**TGCAAAAGGCCAAGGAATTG**GAAGAGATGCACATCCAGGCCGTAACCACTGC
CCAAGAAAACCTTCGCCTACAACATGAAGTATGCGAATGGAAAAAGAAGTATGATCAGCTGGAAAAAGAGGCAATGGAGGT
CGATAGGTTAAGACTAGAAAGAGCTGATCTATTAAAGGAAATACAAAGATTGAAGGAAAATGAAATTGAGGAGGACAAGGA
AATCTTCGTTCTACTCGCAGACGAAACCCAGAAAGTCTTGTCTGTTGAAACTAAGGAAACAAGTGGTTCAAGGAAAGCCAA
AAATATGATGTTTAATCTAAAGGTACAAATCGAAATTTTAAACATCCCCCCTTTTGAAGTTAATGCTATATTAGATACAGG
AGCAACAACTTGCTGCATAGATGAAAATGCTGTACCAGATGCCGCAATGGAGGAAAATCCCTACATAGTACACTTCAGTGG
GATCACCTCTAAGACAATAGCAAATAAAAAGCTGAAAGGAGGTAAAATGACCATTGGGGATAACTCGTTCAGGATCCCATA
TACCTATGCCTTCGCGATGAAACTTGGGGATGACATCCAAATGATAATCGGATGTAATTTTATAAGGGCAATGCAGGGAGG
AGTAAGGATAGAAGGTAATATCGTTACTTTTTATAAGAACCTTACTACAATTAACACACTGCCCTACATCCCAGCAGCAAC
AGCCATAGAAGAATTGGATCTTGAGGAA**G**ATGTCTATGTTCAGATCCAAGAAGCAGTGTTTTTCTCCGCTGAAGCCCAACA
AAGTGACAATGCTATTAGGGCAAAATTTGGAAGCCTACTGGACCAACTAAAGGCCCAAGGATATATTGGGGAAGACCCCCT
TAAACACTGGGAGAAAAACAGGATTCAGTGCAAACTGGAAATTAAAAACCCTGATTTCGTGGTGGAAGACAAACCCCTGAA
GCATCTCACACCACAGGCTAAGGAGGCTTTTTCAAAGCATATAAGGGCCCTCTTGGAAATTAGAGTAATAAGGCCCAGCAA
GAGTAAACACTGAACGACTGCCATAATTGTTAACTCCGGAACAACAATCGAC**CCAATCACTGGAGTAGAAAATAAAGACAA**
AGAAAGAATGGTATTCAACTACAAGAGGCTGAACGACATCACGGAGAAAGACCAGTACAGCCTACCTGGAATCAACACCAT
TCTAAGAAAGGTCAGCAACAGTAAAATTTATTCAAAATTTGATCTTAAGTCTGGCTTTCATCAAGTTGCTATGCACCCAGA
CTCAATTGAATGGACTGCCTTTTGGGTCCCTGACGGGTTATATGAATGGCTAGCCATGCCATTCGGGCTTAAAAATGCACC
AGTAATATTCCAGAGAAAGATGGATGAATGCTTTAAAGGCACAGAGGAATTCATTGCAGTATACATCGATGATATACTAGT
CTTCTCTGAAAATGAAAATGACCATGCCCGACACCTAGCCCAAATGGTGGAAATTTGTCAAATAAACGATCTGGTATTAAG
CCCATCAAAAATGAAGATAGCAGTCAAGGAAATAGAATTTCTTGGGGTAATTTTAGGAAACTCAAAAATTAAGTTACAACC
CCACATCATCAAAAGGATCACTGAATATCAGGAGGAGGACCTTACCACCAAGAGAGGACTCAGATCATGGTTAGGTATACT
CAATTATGCTCGAAATTATATACCTAACTTGGGCAGACTTTTGGGACCACTCTATTCCAAGACAAGCCCAACTGGGGAGAA
AAGATTTAACGAGCAGGATTGGATGTTGATCAAAAATATCAAAAGCATGGTTAAAAATCTCCCAGATCTAGAGGTTCCTCT
AGAAGAATGCTTCATCATCCTTGAAACCGACGGATGCATGGAAGGGTGGGGAGGGGTCTGTAAATGGAAAAAACACAAAAA
TGACCCTCGAAACACTGAGAAAATCTGCGCTTATGCTAGTGGAAAATTCAGTCCTATCAAATCCACCATTGATGCAGAAAT
GTATGCGGTAATGAAGAATCAAGAATCCTTAAAGATATATTTCCTAGATAAAAAGGAAGTTATTATCAGGACAGACTGTCA
GGCAATTATTAGCTTCTTTAACAAGTCTGCTCAGAACAAACCTTCTAGAGTTCGGTGGATGGCTTTCGTAGACTATATAAC
AGGAAGTGGCGTGGAGATAAAATTTGAACACATTGAAGGGACTAGTAATATCCTTGCAGACTCTTTGTCCAGACTAATAAA
TATTCTAGTTGCAGGATGGCCAAGCGAACAAGTATTCCTGCTATTAGAAGCCACCTAGGAGGTTCAAGCACAACCAAACCC
GAAACAACAGCATCCCTCAACAAACTGTTAGTAACCTTGTCCAACAATATCAACAAAAGCTGGACGAGCTCAAATCAGAAA
TGCGAGGAGGAATCCCCCTCTGTAAAATCAAGGCAACCGAAGGAGAGCTCAGGACCATTCAAAGAGAAGCCGCCCGCCAAG
CCTGCGAAGCATTACAATGCTTCCACGACATCCACTCAGCAAAGGCAAAGGAATATCAGAGAAGGAGTGGGGGCAAGGACT
GTTGGTACAAAGACTGGCTACCAACTGTCCTCCAGCACCAGCAACAGCTGGAAAGAGCTCTTGCACTAACCAAAGATTCCG
CCCAGAGGATAAATAACTTCAGTCTTTAGAAACTCGGGGAGCCGTGCATAATTGGAGTCTAGCGTCGGCGCCACTGTTTTG
CGCCACTAGTTTTTACTTTACCATCCTTTTGCACTGAGCCTCGGGGAGCCGTGCATAATTGGAGTCTAGCGTCGGCAACCC
TGTTTTAGTAAAGGCAAAATAGAAAACTTTGACAAAGGGTGTCGATGGGGGCCAATGATCACCCGACCCTCTGCCTTAGTT
TCTCTATATAAGCCTTAGTTCAGCTCATTGCAGAGTAGTCAGAAAAAATCTGAAGTTCTACTTTGAGTTCCAAGTCATAAA
TCAGAGTCTGTAAGTTTCTTTCTAGTTCTTCATACCTATCTCTGTTTTTATTTCTTGAAGGTTTAGTGAAAGCTTAGTGAA
TACTTTAGCTAAGTGATTTCTTGTTCAGTGAAAGCTTAGTGAATACTTTAGCTAAGTGATTACTTTGGTATCAGAGCTGGT
GCATTTTTAAGCCCGAATGGCATGACTAGCCATTCATATAACCTTACGCCATTTTAGTTGCTGGCTAAAATGGGTATCAGA
GTTGGGTAGCCATTCATATTAGCCATTCATACCATTGGTATCA**GAGCTGGGTAGCGGA**CGGTTCTCCTAGCAGCCTTTACC
CCATTCGCCTTGTTGCCTGGCGAGCATTGAATAGCCTTCTGTGATTACTT***TGGTATCAGAGCTTGTTTTTATCAGTGCTAT***

GGCTTTCTAAATTGCTTCAAGATAGTCATGTGGACAGGTTTTGAGGAGCATAGGTTACATTTTTGTTAAGTTCCTTCTATC
ATTTAGAATCAACTCGTGAAGCTCAAATAACCGTTGGGAAGTACTTTCTTTAATCCAGAGAAAGGTTGTCAAAAGATGTAA
GGTGTCAAAATTGGGCAGGGAGACCCTAAGAAAACACTGATAAGGTGAGGTAACCTAAGACTAGAAAAAGGAAAGGAAGAC
CAATGGTTAGGAATGCCTAAGATAGATGTCTGAAAGATGGGAAAAAGCTATACAACAATGGTATACTAGCTCCCATACCTC
GAAACTAGACTATCTTGACCTTGCTGAAAGCAGTAGCCCGACCCGAAAGGAACTAGTTCACAATCTTGCTGTCATTTACGA
CATAACTTGTTTATCAAGTAAAGCACATCTAAAGAACTTCAAAATCCTCATAGAAAAAATTCAATCTCTCTCGAGAGAGGT
TAAGATACTGAGATCCTCTGTGAAAACTTTCTCTGCCTTGTTCTCCGAAAATAAACCTTTAACAAAGCATGAAGTCAGGGA
TCTAGTTGAAGAAATATCCAAGCAACCAAAGCTGGTTGAAGAAAAAGCTTTAAAACTAACCCTAAACCTCAATCAAAAGCT
TCAAAGAGTAGAACAACTCCTAACAAGGATAGAATAACAAATTTTTGGATGAGCTACTTGTTCACCAAAAACACCGAATCC
TATAAAGAAGCACTCAAAGCTACAGAATCCATTGACTCTTCGTACCTTGGGTTCCTAAAAACCAGTGACTATCCCGGAACC
CTAAGCCATCAATCAGCTGTAATCAAACAACACAACACTCAGCTACAACTACTCGTACAAATTGCTGAGGATATAAAAGGG
ATACGTATAGAACTGCAAACCATATGGGAAATCCAGCAAGCAAAAGCCGGATCATCCCAAGGAATCCTTGAAGATTTGAAT
ACTAAATTATCAAATCTAAATTTAGGTCCTTCAGAAAGACCTAAAGAGGTGCAAGGAAGAATCTTAGTATTCAAAGATCCC
ATCCAGATTCTTAAACAAGTCCAACAATGACAACCCAGACTTCAACCCAAGAAGTCTCCACATCTGCCCCTTTAGTTGAAG
ATCAAATCCGAGACTATAGGAAAAACCACAGAAGGCTATTCAATGCTCGCCAGGCAACCAGGCGAATGGGGCAAAGGCTGC
TAGGAGGACCGTCCGCTACCCAGCAAATCCTTGAGCAGCAGATAGACCCTCAGGCCCAACTGAGGCTATCCATGCGTGAAA
GGGCGACGATAGCACCTGCCGAGGTGCTATATCACTCCAGGCGAGATGATGCACACCATTGAGTTTATGTGCACCGATCCG
AAGAGGCAATGTTGGTCACCAATAATCAGGAAGACAGGGCTTTCATCATAGAGGAAAGCTATGACCGACTGCAAAGGAGCC
GCATGCAATACATTCACCTAGGCATACTGCAAGTCAGATGCAAATGCTTCATTGGCAAGAAGAAGGAACACTAGCACTATT
AGTGTTCAGAGATAATAGATGGACTGACGATAGGTCCATCATAGCCACCATGGAGGTAGACTTGACTCGAGGAAGTCAGCT
GGTCTACATCATTCCTGACATCATGATGACGATTAGGGACTTCTTCCGCAACATTCAAATTTCAATCCTCACTAGAGGATA
TGACACATGGCGGAATGGTGAGGCAAACTTGCTAATCACCCGTGGAATAGTAGGCGACTTTCAAATACCACGAATGCCGCC
TTTGCATATGAAACATCAGGGGTGGTGGACTACCTAACAAGCCATGGTGTCCGGGCACTTCCCGGAAGGAGGTACTCCACA
GTTGAGCTACATGGCAGGGACTGGGTCATCAGACCAACCCAAATTGCAATACCAATACAGCCAACAGAAGTAAGGAGTCGC
AACCTTATTGACTGCAGAATATCTATAAGCTTTGACAACTACAAAGCTGCATCTACATCCAGTCGGATCAACTACAATACC
GCTGATGATGAAACATTCAGCGATGAAGAAGAAATCTGGAGCCACATAGTAGCTGTCAACATCCAATTATCTGATGACCGT
GAAAATGAAGCTAAAGAATTATGTAAAAACCTGAATTCCTATTGTCAGGATTTTAACGTTTCAGAAGGAGGTGGGGAGATG
CCATATCCCCAAAAATTTCAAAAGGAAGTAATTGCAGCAGGACTCGAGGAAGTCCTAGCAATGGAATACCCCCAACTTGCA
AAGTTATCTCAACAGGTATATTCATTCTCTGCTGCATCAAATTACAGACCACCTGCGGATTCAACTATGGGACCAGCAAAT
TACCCCCCAGCAGTGAATGTGGAATCCACAAGCCAGAGGCCCGCATATGAAGGCTACTCAAGACATCCAAGGTTCAAATCA
AAGGATTTCTCAGAAGCTTGGAACCTCCCATCAACCTTCCAACAACAAGGGGCAATGTTTATAATTCCATCCCAACTTGGG
ATGTTTAATGAAGTATTTATGAGATGGGAATCAATCACTAAAAACTTGGTTTCCCTACAAGGATTCACTGATCCCCAAGCA
AAAATGGAATTCATTGAAAACTTGCTTGGAGAAGCAGAGAAATTAGCATGGATCCAATGGCGAATGGCGTACCCAGAGGAA
TACCAACTACTAATGGCCAACGCAGACGGAACTGGAGGAACTCAAAATATCCTCTCACAACTAAGGACGATCTTCATTCTA
GAGAATCCGTTATAGGGTTCAACTACAGCGCAAGAAGAAGCTTACAGAGACCTCGAAAGGTTATCCTACACAAACCTCAAG
TACATTATTCAATTTCTAAACGATTACATGAAGCTTGCATCTAAGACAGGAAGGCTGTTTACAAGCCCAGAACTTTCCAAA
AAACTATGGTCTAAAATGCTTGGAGAATTAGGAAAAAGAATCAAAGAGGTGTTTGAGACGAAATACAGGGGAAATACTATA
GGAGTAATTCCAAGGATACTTTTCTCCTATAAGTATCTGGAAGCAGAATGCAAAGATGCTACCTTCAGAAGGGCGTTAAAA
GACTTATCCTTCTGTAGCGAGATCCCTATCCCTGGATATTATAACAAGCCCGAGAGGAAGTATGGCGTGAGAAGATCAACA
AGCTACAAGGGCAAACCCCACTCGTCTCACGCCAGAATTGAAAAGAGGAAGCACCTGATAAGAAACAAGAAGTGCAAGTGT
TACTTGTGCGGGGAAGAAGGACACTTTGCAAGAGAATGTCCCAATGACCGTAAAAACATCAAGAGAGTCGCTATGTTTGAA
CAATTAGACCTTCCTTATGACTATGAAATTTTGTCAGTACAAGAAGGGGAAAACCAGAGCGATGCAATCTACTCCATCTCT
GAAGGAGAAGACGTAGAAGATCTACAACACGGTATTCACTCCTTCACCCACAAAATTTTTGCACTAATAGAAGATAGTAGA
ACTTGGTGGATAGGACCCGAGTCAGGTTACCGGGCCAGAGTTCAAGTCTCCCAAGCACAAGCTGAATGTAGACACATATGG
AAAATTAACACCGAGTTACCAGCCAATTTGGAGAAGTGCAAATGCTGCAAACGGACATCACAAAGGAGGCGCAGGAGACAC
TGCCCCTTGTGTAAAATTACTTCATGTGAGATGTGCAGTATCTACTACTTTGATAAAAGAACCCCCGTAATGACTGAGGAA
CCACCAAGGTATGAACCAAGAAACTTGCCTCAGCAGCAACGGGATTATATTAATCACTCT

MaLAR1: **4564 bp LARD-like element in *Musa acuminata* (AY484588.1) from 48330-52793 bp**

GGTT*TGTCACTGACTTAGCTAGTTTTGCCTAAGTCGTGCGGCACCCTTGCATGTCCGTCCGCAAAGGTCAGTCTCCCCGAA*
GCCTCCCATGGTCCCTTAGGACATACGAAAGAGAAAACGAGTTAGAGAGAACGCCTCACTTGGGATCCATAAGCAAACATT
CCAAGAAACACTTCATAGACAATGCAAATTACAAACAGACTTTACAAGCTCTGAACAGTTGCACAACAAAGGGTCAAAATG
GTCTACTACAGACCGAAAATTTCTCACAAGTGTCCACATGACACAACCTTTATTTACAAGCCTAAAGCGGCCACCAAACCC
AACTAAAATGAGGCTATTAAGCCTTCGGTTGTCCCTCTACATGCTAGGCAAAGTATGAACATACCAAAAGATACGGACATA
*CATAAGCATTACATCAAACATCCTGTTTAGAAGTTTATCCGTGACA*TTCTCCCCCACTTATTCCTTCGATGTCCTCGTCGA
AGCCTTTGTGAACACTGTAACTCCTTGCCTTTGTTGAGTCTTCAATCTTTTGCTCCAGCTGCAATGCGCCTCTTGCTCCCA
GTTGCTCTCTACTGCTGTTTTTGAGTAGTCAAACCTTTGATCCGCTATGATGCTTCAACTCACCAATGACTCTGACTTTGG
TGTGGGGTTGGCTGAGTTGTGTTGATCCTTATTGATTCCTGTGGATCCTCCAGATGAAGGAAAATACCATCTTTACTATAC
GCTAGTCTCTCAAATGTCTCATGCTGCTTGAATTGGGTGGATAGTTGTTGGAGCTTTAACAAGTATCGCATCGTAAACTTT
CAAAGCTTTTGGTCCTTCCTCCACAAAATTTGCTTATCGACTTTTCTTTCACTTAGTTGTCACTTCCAAGTAGATTCGCAT
CACTTCCGCTTTTGATTGACATTCTGTTGGGAAATGAGGCGAATACTCTACTCTCAGTAGCACTGATCACCAATGGTGAGG
ATTTGATAACTATTGTCTTCTAATATCTTCAAAGGGTCTTTGAACCTGTGCAGAGCTCCTCTGCTGGATAGATAAGAGAAT
TAGGGTACTCGGTTTCGTCCATTCTCTTAAGAGTTGAGAAGGCAAAGGTTACTTAACTTCGCCCGCTTCCTCGAGGTTGTA
CTCCATGCATCGAGCTGGTTATTGGCCTTCGCCTGCTCTTTGCTCATACTTCTAAAGCACTTGAAGTGTTTGCACTCCTTG
TGTTGAGTTAGTTACTGTGATTCACCTTCTCAATGCCATCGAACTTCTGGAATGCAGGAAGTTTTCACCCCAACTTGGAGT
AATTCTCTAATAGATTTGGTCGCCTCTGGGATTGTACCGTATTCTCCATCAACCCTGCCGCCTACTCCACTAAGTAGCAAA
GGTATAGCACCGCGTACTACTTGCTTCATTCCTTGGTTGTGCATCTTGCCTAACCCGAAGTCCTTCACTTTCGACTATCTT
GATGAGAAGCTCGTTGACACCGGTCTTACGAAGTTCCTTAGCCTCTGCCCTTCAGCCTTGTCTCGGTACTTGGAGTTTGCA

TCTGCATACTCCACCTTCTCGGCCCCTTTTACGATCAAGCACTCTCCCTCCATGAGAGCAAGGGATCAATGACTTTCACGG
AAGTCCTGCCTCTGCGGTACCATGGCACTGCCATGCCCATGGCCCTACTATTCGTCACCTCGCATCTGCATCCCTTTTCTT
CACGATCAGTAGATATGTCTCTGTGGTACTCCTCCGAGTCCACCTCCATTCTAACTGATGCTTGATTTTGGGTAGCTAAGT
CCCTATGGACTTATCGTCGCTTCCTCGCCCCTTTAGATCCTCTGCTTCAACACCTTTGTGTTCTCCAAGCAACTCCCTTTG
ATCGATGGAGAGACAGATTACAACTCCCATGCATGGCCTCTGCCATTACGTTGTAGAGTTTGCATCGATTATGTTCTCCTT
AGCTTCCTTGGTAGCAACATTCGCTTACTCGACCTTGTCCTCTGACTTACCGGGCTCCCTTAAGCAAATATGAGCTCTGGA
GCAGTCCAACTCTCCAGCTGCTTCAATCATACCTCTGTATGATCAACTCCCTCCCATGAGACTCATTGGTACTTGCATTCA
AACTTTTTCCTTGGTGGAACACAACCCTCATATGCTAATGACCAAGGCTTTCATCCGATGCAAAATTCGATGCACGCACGG
AAGACCCGCCTCTACGGTACCATAGCCTTCACTTCTTGAATCCATAGCCCTTCTTGCCGTTGTGTTGTTCACTGAAGTGGA
GCTTCCAATAGCTCCCGATCATACCTCCGTATGATCTAGTCCCTTACAGGACTATGTCATGTGTATCGCATTGCTATAAAC
TCTTCCACCACGATCCGCTGCACAATGCTGCCTCCTGGTGACATCCCCATTGCATTCTGATCCTTATGGGATGAACTCGAA
TTGTGATCCCTCAATGTGTGACCTTTGCCAATACATCGTAGTGTCTCTTCCACCTTCGATTTTGTTCAATCATTTGGCAAT
CGACCTTCATCCACCTACTCTTGGGTCACACCTAGATGAAGCACCGCTCTAGGACAGTCCATCGCCTAGTAGCTCCCAAAG
TCTCCCGACTTCACTGCAATTAGTGCAACATTGTCTGGATCCTGGGACTCTGCCCCTACTAGCACAATCTTCGCTGTGCAC
CACTTCCTTCATAGCAACTTGAATGGCAAAACTGTGGCATATTCTTCAAGAGTACCTGCCTCCACGTCCTCTTGCCCCGTG
CTAAGGCCTTTTGAACCTAACTTCGCCTCTACAAGTTGAGTCGCCTTAGTTCCTCCATCAAATGCTTCTCCGAGATAAGGT
GTATGTGCCTAGAAGCTCCCTTCGTCTTTGGCACCATGCAAGATGAGTCCGCTTCGTCAGAATGAAGGACCCATGGAACAA
CATGATCCTACTCTTGCCTCTGCAAGAGTTCATGTCCTTGACCTCTGTCCAAGGAAAGCACTGTGCCTCCGCTCCATGTTC
CATCTTCTATGCTGGCTCCCTTCATGCGGCTTGGGTACTTTGCCAAGTTACACCCAAGTTGCTCCGCTCCTCATTTTTGCA
TTGAGTTGATGGTGGCCCTCGCGCCCACTATTCCACGGGTCAGCCCTCCCTTGAGTCCGATCTCTATATTGACTCCAAGTG
TGCCTTCATTTGTGTTGCTTTGTGTCGCTCCCCCACTTGATCTCACAATGCATTCTCTCAAGCAAGATCATGCGATGACTC
CTCGCCACTTACTCGGTCCATTGAGCTTCATGGAGTTATTGTTTGTTGAGGCACTCCTCCTCAACATGTGAGGTCCATCCC
ACTTGATTCTCCCTCTGGAGAGCCGAGACTTATCCCTCCTGGATAACTGTCCCATTGGAGCAACATCTCTTTTAGTTTCGG
AGACCACCATCCCCTTGGACTACTCCGATTTGCTGAACAAACTGTGTATTGTTCTACCTCCTGCAAACGCACTTGCTAGAT
TGCGACTCCACGTCAATACAGCCCCCACTGTACCACTCAAGGCCTAGCAACATGCTGAACTCGTTGTACACTTTAGCCTCC
TACGGACGTATCCTTCACATGCCGAAGAGAAAGTTTTAATGCTCCATGGCGTCGAGTTTCGGTCGCCTTGGGATTGCCGCG
AACATTCCATCGTCTGTATACAAGCCCATGTATGAGTACCGAATTCTTTGAGTTAGCAATTCCCCTCATCTCTGTGAGCTT
TACACAACTCTTATGGTCGCTGAGCAACTCATTCTACCTTGCATTGTCTCATCCTTTACCAAGCGCCTCGCTTGCCTTGAG
CACCATCAAGTATAGTTGTCAACGTTGAGCCGTAGCTCAAACTCAACCATCCCAACCTTTGTGCGCTCCGCATTCTTCCAA
GCTTGCCTATTCTCGTGGTGTCTCGTGCGCGAAGGGTTGGCCATTCCTCTGAATGCCAATCTCAGATGCTCGCTCTTTCGA
GCGACTCCTTTTCCCTACATCTCCATACCCGTTTTCCCCCAAACGATCGCGCGTGTGCTGACTGCCCTCAATGCAGCCCCG
CTAGGTCCCCCACATTTGCATTCTAAGTGTTTCTATAAGTGCTTGTCCCACTCTGATACCATT***TGTCACGGACTTAGCTAG***
TTTTGCCTAAGTCGTGCGGCACCCTTGCGTGTCCGTCCGCAAAGGTCAGCTTCCCCGAAGCCTCCCATGGTCCCTTAGGAC
CTACGAAAGAGAAAACGGGTAAGAGAGAACGCCTCACTCGGGATCCACAAGCAAACATTCCAAGAAACACGTCATAGACAA
TGCAAATTATAAACATATTTTACAAGCTCTGAATAGTTGCACAACAAAGGGTCAAAATGATCCACTATAGACCGAAAATCT
CTCACAAGTGTCCAAATGACACAACCTTTATTTCCAAGCCTAAAGTGGCCACCAAACCCAACTAAAATGAGACTATTAAGC
CTTCAGCTATCCCTCTACATGCTAGGTAAAATATGAACATACCAAAAGACACGGACATACATAAGCATTACATCAAACATC
TTGTTTAGAAGTTTGTCCATGACAGGTT

***MbLAR5* 4449 bp LARD-like element *Musa balbisiana* (FN396604.1) from 28462-32910 bp**

ATAC***TGTCACGGACTTAGCTGGTTTTGCCTAAGTCGTGCGGCACCCTTGCGTGTCCGTCCGCAAAAGTCAGCCTCCCCGAA***
ACCTCCCATGGTCCCTTAGGACCTACAAAAGAGAAAACGGGTTAGAGAAAACGCCTCACTCGGGATCCACAAGCAATCATC
TCCGGAAACACTTCTTAGACAATGCAAATTACAAACAAACTTTACAAACTCTGAACAGTTGCACAACAAAGGGGTAAAATG
GTCTACTACAGACCGAAACTAAAATGGGACTATTAAGCCTTCGGCCGTCCCTCTACATGCTGTTCAAAGCATGAACATACC
AAAAGACACGGACATACATAAGCATTACATCAAACATCCTATTTTGAAGTTTGTCCGTGACATTCTCCCCCACTTATTCCT
TCGACGTCCTCGTCGAAGCCTTTGCCGACACTGCAACTCCTCGCCTTTGCTGAGTCTTCAATCTTCTATGCCAGCTGCAAT
GCGCCCCTTGGCTCCTAGCTGCTCTCCGCTGCTGTTTTTGAGTAGTCGAACCTTTGATCCGCCATACTGCTTCAACTCGCC
AGTGACTCTGACTTTTGTGTGGGGTTGGCTGAGTTGTGTTGATCCCTGTTGGTTCCTACGGATCCTCCAGATGAAGAAAAA
GACCATCCTTACTATGCGCCAGTCTCTCAAATGCCTCATGATGCTTGAACTGGGTGGATGCTTGTTGGAGCTTTAACGAGC
ATCGCCTCGCAAACTTTTGAAGTTTTTGGGTCCTTCCTCTACAAAATTTGTTCATCGACTTTTCTTTCACTTAGTTGTCAT
TTTCAAGTAGATTCGCATCACTTCCGCTTTCGATTGGCATTCTGTTAGGAAATGAAGCAGATAATCTACTCTCAGTAGCAC
TGATCACCATTGGTGAGGATTTGACAACTATTGTCTTCTAATATCTTCGAAGGGCCTTTGAACCTGTGTAGAGCTCCTCTA
CTGGATAGATAAGAGAATTGGGGTACTCGGTTTCGCCCATTCTCTTAAGAGTTGAGAAGGCAATGGTTACTTGACTTCGCC
CGCCTCCTCAAGGGTTGTACTCCATGCATCGAGCTGGTTACTGGTTTTCGCCTGCTCTTTGCTCACACTTCTGAAGCACTC
GAAGTGTTTGCACTCCTTGCGTTGAGTTAGCTACTGTGATTCACCTTCTCAATGCCATCGAACTTCTGGAATGCAGGAAGT
TTTCACCCCAACTTGGAGTAAGTCTCTAATAGTTTTGGTCGCCTTTGGGATTGTACCGTCTTCTTCATCAACCCTGCCGCC
TACTCCACTGAGTAGCAAAGGTACAGCACCGCGTACTGCCTGCTTCGTTCCTTGGTCGTGCACTCTTGCACGACCCGAAGT
CCTTCACTTTCGGCTATCTTGATGAGAAGCTCGTTGACACCGGTCTTATGAAGTCCCTTGGCCTCTGCCCTTCAGCCTTGT
CTTGGTACTTGGAGTTTGCCTCTGCATGCTCCACCTTCTCGGCCCCATTCACGACCAAGCACTCTCCCCCCATGAGAGCAA
GGGATCGATGACTCCACAGAAGTTCCACCTCTGTGGTACCATGGCGCTGCCATGCCCACGGCTCTACTATCCGTCGCCTCA
CATCTGCATCCCTTTCTTCACGATCGGTAGATATGTCTCTGTGGCACTCCTCCGAGTCCACCTCCATTCTAACCGATGCTT
GATTTTGGGTAGCTAAGCCCCTCTGGACTCGTCGTCGCTTCCTCGCCCCTTTCGACCCCTTGCTTCAACACCTCTGTGTTC
TCCAAGCAGATCCCTTTGGTCGATGGAAAGACAGACTGCAACTCCCATGCATGGCCTCTGCCATAACATTGTAGGGTTTAC
ACCGATTCTGTTCTCCTTAGCTTCATTGGTAGCAACGTTCGCTTACTCGACCTTGTCCTTTGACTTGCCGAGCTCCCTTAA
GCGAATATTGGCTCTGGAGCAGTCCAACTCTCCAGCTGCTTCGATCATACCTCTGCATGATCAAGTCCCTCTCATGGGACT
CACTGGTACTTGCATTCGAACTTTTTCCGTTGGTAGAACACAGCCCCCATATGCTGATGACCAAGGCTTTCATCCGATGCA
AAATTCGATGCACGCACGGAAGACCCGCCTCTGCGGTACCATGGCCTTCACTCCTTGAATCCATAGCCCTTCTTACCGTCG
TGTTGTTCACCGAAGTGAAGCTGCCAGTAGCTCCTGATCATACCTCCGTATGATAAGTCCCTCACGGGACTTTACCCGTGT

GTATCGCATTGCCACGAACTGTTCCACCACGATCCGCTGCACCATGTCGCCTCCTGGTGACATCTCTATTGCCTTCTGATC
CTTGTGGAATGAACTCGAATTGTGGTCCCTCCATTTGTGGTCTCTGCCACCACATTGCAGGGCCTCTTCCACTTTTAATTG
TGTTCGCTCCTTTGGCAATCGACCTTCGTCCACCCGCTCTTGGGTCACACCTAGACGAAGCACCGCTCTAGGACAGTCCAT
TGCCTAGTAGCTCCCGAAGTCCACCGACTTCGCTGCAATTAGTGCACCATTGTCTGGATCCTGGGCCTCTGCCCCTACCAG
CCCAATCTCCGCTGTGCACCGCTCCCTTTCATGGCAACTTGAATGGCAACACTGTGGCATATTCTTCAAGAGTACCCGCCT
CTGTGTCCTCTTGCCCCGTGCCAAGGCCTTCTGAACCCAACTTTGCCTCCGCAAGCTGAGTCGCCTCAGTTCCTCCATCAA
ATGCTTCTTCGAGATAAGGTGCATGTGCCTAGAAGCTCCCTTCGTCTTTGGCACCATGCAATCCGAGTCCGCTCTATCAGA
AATAAAGGACCCATGGAACAACATGATCCTGCTCTTGCCTCTACAAGAGTTCATGTCCTTGACCTTTGTCTAAGGAAAGCT
CTGCGCCTCCGCTCCATGTTCCCACTTCTATGCCGACTCACTTCATGCGACTTGGGTACTTCGCTAAGTTACACCCAAGTT
GCTTCGCTCCTCGTTTTTGTCATTGAGTTGATGGTGGCCCTCGCGCCCACCATTCCACGGGTCAGCCCTCCCTTGAGTCTG
ATCTCTATATTGACTCCAAGTGTGCCTTCATTTGTGTTGCTTTGGGTCGCTCCCCCACTTGATCTCGCAATGCATCCACCG
ATGCATTCTCTTAAGCGAGATCATACGACGACTCCTCGCCGCTCGCTCGGTCCATTGAGCTTCGTGAAGTTGTTGTTTTTG
AGGTACTACTCCTCAACATGTGTAGTCTATTGCACGTGATTCTCCCTTTGGAGAGTCGGGATTTATCCCTCCTGGATATCC
ATCCCGTTGGAGCAACATCTCTCTTCGTTTCGGAGACCACCATCCCCTTGGACTACTCCGATTTGCTGAACAAACTGCGCA
TTGTTCTGCCTCCTGCAAACACACTTGCTAGATTGCGATTCCACGTCAATACAGCCCCCACTGCACCACTCAAGGCCTAGC
AACATGCTGAACTCGCTGCACACTTCAGCCTCCTACGGACGTATCCTTCACATGCCGAAGAGAAAGTTTCAATGCTCCATG
GAGCCGAGTTTCGGTCGCCTTGGGATGGTCACGAACATTCCGTCATCCGCATACAAGCCCTTGCATGAGTACCGAATTTTT
CGAGTTAGCAATTCCCCTCACCTCTGTGAGCTTTGCACAACTCTTTCGGTCGCTGAGCAACTCATTCCACCTTGCATGATC
TCATCCTTTACCAAGCGCCTCGCTTGCCGTGAGCACCATCAAGTATAGTTGCCAACGTTGAGCCATAGCTCAAACTCAATC
ATCCCAACCTTTGTGCGCTTCGCATTCTTCCCAGCTTGCTCGTTCTCGTGGTGCCTCTCACGCGAAGGGTTGGCCATTCCT
CTGAATGCCAATCTCAGATGCTCGCTCCTCTTAGCGACTCTATTTCCCTACATCGCCATGCTCGTTTTCCTCAAAACGATC
GCGTGTGCTGGCTGCCCTCAATGCAGCCCCGCTAGGTCCCCCACGTTTGTATGCTAAGTGTTTTTATGAGTGTTTGTCCCG
CTCTGATACCATC***TGTCACGGACTTAGCTGGTTTTGCCTAAGTCGTGCGGCACCCTTGCGTGTCCGTCCGCAAAAGTCAGC***
CTCCCCGAAACCTCCCATGGTCCCTTAGGACCTACAAAAGAGAAAACGGGTTAGAGAAAACGCCTCACTCGGGATCCACAA
GCAATCATCTCCGGAAACACTTCTTAGACAATGCAAATTACAAACAAACTTTACAAGCTCTGAACAGTTGCACAACAAAGG
GGTAAAATGGTCCACTACAGACCGAAACTAAAATGGGACTATTAAGCCTTCGGCCGTCCCTCTACATGCTGTTCAAAGCAT
***GAACATACCAAAAGACACGGACATACATAAGCATTACATCAAACATCCTATTTTGAAGTTTGTCCGTGACA*ATAC

CHAPTER 9

MOLECULAR CHARACTERIZATION OF DNA TRANSPOSONS AND NOVEL MOBILE INSERTIONS IN *MUSA*

***MaN-hAT1*: 273 bp non-autonomous hAT in *Musa acuminata* (AC186955.1)**

TCCCTGAG***CAAGGTCTG****C****CATACCG****T****A****C****CGTACCG****G****CG***TTTCGACCCGGGCTCGGTACGGTACCGGTGTACCGGGCAGTAC
ATCAGGGTGTACCGAATGGTACACCCTGATGTACCGAACAATTTTATACTTTTTCATACTGTAGCAGTGCTACAGTATAAT
ACTGTAGCACTGTAGCGGTATCGGGCGGTCCGCGTACCGGTAACCTGTCGGACCGGTACATACCGCCCGGTAT***CG****G****CGGTA***
CG*C****T****CGGTATG****A****CAGACCTTG***TCCCTGAG

***MaN-hAT2*: 874 bp non-autonomous hAT in *Musa acuminata* (AC186955.1)**

GTGcTAaC***CAGTGATTTaAAaAG***CGCTAAGCGCCAAGCGGCAAAATGCGTCAAGGTCAAAAAACGCCCGAGGCGCTAGGCG
CTCGCCCAAGCGAAGCGAGACGCTAAAATATAAAAATATATAATATAATTAATAAATATAATTATTTAAATAAAAAATATG
CTATTAAATAAATAAAATTTAATGGTATTAAAATCAAAATAATATATTATTAATCTAATAAATAAAAATATTATTATTAGT
ATATAGTTAGAAGTATACTGTTAACAGTATAATGAGAAGAGTGTGAGAAGACCGAGGCTGCTGTGGCAACGACAGCGGTAG
CAGGTGGCAGTCATAGCTGGAGCGAGAGCAGCGAGCGACGGCAATAACGAGAGCAACGACGACAGCGGGAGCGGGAGCTGC
GAGTGGTAGCGGGAGCGGCGATAACAGCGACGAGTGACGACAACAACGACGAGCGACGATTATGGCAGCGGCGAGCAGCGA
TTGTGGCAGTGGCGAGCAGCGACAATAGTGGCGAGCAGCGACAGTGGCAACGACAACGGAGAAGAAATCTCGGCGGCAAAA
TCGTGAGTGTTAGGGTTGGGGAAGTCGGGGAAATCGCGAGAGAGAGCCTGATATCGGTGATTTAGATGGTTCGATTGAACC
AACTAAAGCATTGGAGACCAACCAGACCTAAAAATCTGGTTCGGTCGCCTGGTTTAACCCGGGCGCTCGCCCGAAGCGCCC
GACGCCTGGGCTCGGGCGAGCGCCTAGGCGGTGCCCTCTTTGAAGCGAGGCGCCTGGACATGAAGCGAGGCGCTCGGGCCT
CGTCTCGCCTCGCCCGAGCGCCTAGGCGGGCACCCGAGTGC***CTaTTgAAATCACTG****GTGtTAgC*

***MbN-hAT3*: 1292 bp non-autonomous hAT in *Musa balbisiana* (AC186754.1)**

GTTGCAAC***CAAGGT***CT***GC***A***TACCG***AATCATACCGCTTGGTATGGGCGGTACGTACCGGTCCAACAGGCTCCCGGGCGGTAC
GTCCAAAAACCCCCCGTATCAGACAATACGGGGTAATATCGGGCGGTAACGGTCGAAATTTCGATCGTTACCACCCGGCAC
CACTCGATAACGGTTGAAATCGACCATTATCGCACTGTAGCAGTGCTACAGTGCTCCAACGGTCAATTTAATCGTTGGAGC
CCTTTCTCCTCCTATTTAAACCATTCTTCTCTCCCCTATTTCATTATACTCTCTTAAACTCTCTCTCAAACTTTTTTTTTC
TCTCCTAAACTCTATAAAACAACGATTTGTAAATTCAATCGAGGCTAATTTGGGAAGATTAAGAGGAAGAACTTTCTAATC
AAGAGGTATGTATTCGTTTTCTTTAATTAGAGAGTAATTAAGATTTTACACATGTCACTCAAGATTCAGATCATGGAACTC
GACAAAGTACTAGTCAAGTTTATGCGTGGAAGGGGAAGGCAGTGGATGAATATGAACAAATGCGACAGAGCATACATGATA
TAGACACAGAAAGAGGCTCATCGTATTCACAACCATCGTATTATGGAGAATTATACGAGCAACAACAATATGGTGATAGAT
GGTCATCCTTCTCTGAGCAACAACATGATACAGAATAGCACCAATATATGCCTCAAGAGCTGCCTCGGATAAATATGATTC
ATGACGATCAATCTACGATCAGCACCACATTGATGTATCAATGGCATACGGTGTATCAATACACTATGTCATGTGATCAAT
TTCATGATTGAGTCTAACAAACGTATCATATTGATATGTATCGAATCGAGGACCCAGATCCACCACCCGTGGAGGCTCGTC
GTTCGTTTTGGTGGTAGAATCAACAATCAGGTATATCTAGATATGTGATTATTTAATTCATTTGAATTTTAGAGTTTATAT
TGTAACAAAATAGTCTAACAATTCAACTGATTTTTCTGTAGATATTTCAATATATAACGGGCTGAAATACGAACCTCGAAC

AAGACTACCTCAAAGCCCGAAAAAGATATATGATACTATTCTTACCTAATTTACATTGATTTTGCTAAACTTATACTATTA
CTATATTTATTTGTAACTTGAAAATCCTATCCTAACTTTTTTTTTATTTTCAGGTTCGGAGGGTTAAATCGATGAAATCAG
GTGTACCGCTCGGTACACCGTAACGAGCCCGGGTCGAAACGCCGATACGGTAT***GGTA***CG***GC***GA***ACCTTG***GTTGCAAC

***MbN-hAT4*: 524 non-autonomous bp in *Musa balbisiana* (AC186754.1)**

TTCAAATG*CAAGGT*T***TG***CTGTACCGAATGATACCGCCCGGTACGGGTGGTACGTACCGGTCCGACGGCATACCGGTACTCG
GACCGCCCGCTACCGGGCGAAACAAAAAATAATAATAATATTATATATATATATATATATATATATATATATATATATA
TATATATATATATATATATATATATTAATTAAAAAAAAAGCAATGTTCGGTGACGTTGCCAAGGCGACGTGACGTCGATAT
ATATATATATATATATATATAAATAAACGAGGCGACGTCGCCCCGTGTGGGGAAGGAAAAGGCGACGTCACCGAGGCGATGCG
ACGTCGCCTTTTAAATAAAAAAAAAAATATATAATCTTATATATATATAGACTCGGCGACGTCGCTTCGACGACGTCGCCG
AGTTATATATATATATATATATATACCGAGCGGTATACCATTCGGTATACCGTT**CCGTACCG***T*ACCGAGCGATCGTCGAAA
CTCTGGTACGGTACGAAATTT***CA***G***ACCTTG*****TTCAAATG**

***MahAT1*: 5204 bp autonomous hAT in *Musa acuminata* (AC226051.1)**

TTTTCAAT*CAAGGTTCGTTG*TACCATGGTATATCACCCAGTATGGGTAGTACATACCAATCTAATAGGGCACTGATATGCG
GATCACCCCATACCAGACTATACATGTTACATGACCCTATATACGGGGCTTGTATTAGTCGAAATCCGAATGTTACTGACT
TGTACCGATCGATGACAGTCGAATTTCGACCAATACCGATCAAGGGCTGAGATTGCCCTATTTCAAACGATCAATTTGACC
ATTTAAACCCTTTTCTCCTCCTTTTAATCCATTTTACATTCTTAAACTCTTTTTCACTCACATCTCTTTCTCATTCTCACA
TTTTCAATACTTAAACTCAACAAAGCTACGATTTATGATTTAGATCAAATTTGGAAGGCTAATTTGAGAGGATTAACACTT
AAGTTAAGGGAAGAATACTCTATTCAGGAGATATGTATTCTCTTTTTAGATTTTTGTTGATTTATGAGATGATTAAGACTA
TTCTATGTGGTTAATGTTTAATATATTGTTTGACATATTTTTAATGGTAAAATATTGTTAATTAAGGAGTAATTAAGGCCT
TTAATATTGTGATTAGTGTGTTTAATATTGTTTTTTTTAATTAGACACTTAATGGGTCAAATCATTATATTTCATGCATAT
TAAGGTGATTTACATCTTTGTGATGGTATAATAGATTTAATTAGGTTTTAATTAAGTTTTTTAATGCTATGATTAGTGGTA
TTAATGATGATTTAATCAAAGAGCCGTATTAGGTCAATTCAATGTCTTTAATACCTATTACAATGTTTGACATATTTATAA
TGTGAAAAAAGACTAATTAAGGATTAATTAGCTATATTAATGCTGTGATTAGTGTATTTAATTACGATGAGAACAAAATTT
GACTCATATTGGATCAATTAAATGTCTTTTAATGCCTATTAGAGCATTTGACATGTTTATAGTGGTGAAAGAAAAATAATT
AGAGAGCAATTAACTATGTTAAAGGTGTGATTAGTGTATTTAATGATGATTTAGTCTTAATTTAACCTATATTGGATCAAA
TTTACATATTTAATGCCTCTTAGGGTGTTTGACATATTTATAGTGGTAAAATAAAGTTAATTATGGCTTGATTAAGTTGTT
TAATGCTGTAATTAGTGTATTTATGATAAATTTATCATTATTTGACCTATTATTGGGTCAATTTCATGTATTTAATAATTA
GAGTGTTTGACTCACAATGAGGCATCAACAACTTGACTCACAATGAAAGCGTGGCGATGGTAGTGGGTCTGCGCCACAAGG
ATTTCAAAGGTCGACCGACATGAGACAGCTAACTGCCCCTCCTCAATTGAGTCAGATTGGTAGCATGAGGCAGGGTAGAAT
CCATGATTTTATCAAAAGACTTGGCAGGAGATCAGCATCGTATATTATTAATATTGATCCGTAAACATATCATCCACAAAC
AGCAAAGCAGACATGGATTGACGATGCATATACAAAAGAAAAGAAGTGAGATATTAGGAAGGCAATTATAAAATAATTCAA
CTTCCACAGGATTCCAACAAACACAGCCCACGATTCATATTATCAGAGCATGATCTCCTCTATTCGAAAGGCTAACACAGG
GATTCAACCTCCAATATTCAAGGAGAACCACG**ATGTGTATTTAGATGAGGAAGTGGTAAAATTAAAGAATTTGATCAAATC
CTTCAAGAGGCAATGGGACGAATATGTAGTGATAGTTGGACAAGACCAACAAGAATGAGTATCATCAATTTTCTTATGCAT
TACAATAGACGAGCAGTTTTTCATAAGTCTAATTAATGCAAACTATATTGAAAACATACTTAACACTATAGTAGAGGAAAT
CAGACCATAGTACATTATCCAGATACTCATCGACAATGAAGCCAATTTCAATAAGATCAACTTATAATTGATGGAAAAAAG
GAAAACATCGTTTTGAACTCCATGTGCAGCTCATTACATAGATATAATGCTGAAGGATATTGGCGAGCTACCATTAGTCAG
CA**AGTGTGTAGCTAAGGAATAATCGATTATAAAGTTTACATATAATCATCAGTGGGTCCATTCTTTTATGCAAAAGTATAT
AAATGGTGAGATACTCCGACCTGGAGTCACTTAGTTTGCTATAAATTTCATTGTCTTAAAGTCAATACTCAATACAACAGA
AGAGGCAAGGTCTCATGACAATTGTCACCTCACAGAAATGGTCCGATTATAGGAATTCAAGATCGAGCGATGGAAAAAATG
TAAAAAAGACTAATCTTTCTTCTAAATTTTGGGATACCGTGAATGAAATTATAACAAAAGTAGAACCACTTTATGTTGTCT
TCCGTAAGATCAATAAGGACAAGCATCCCCAAATGCCTTATCTTGACATATGCTGATTACAGCAAGAGAGGAGGCTAAGAA
AGCATTCATAGATTATTTTAATGTTGATCAATACTTATAGATCATCGATCGTAGGACTGAAGTTCATATGAACCAAGACAT
CCATAATACAAATAAATTCATATTCAGTGTAATTTAATTCATAAGTTTTGATATTTAAAAATTTCTTACTAACAAAATTAT
TTGTCTTGCAACATATTATCTAAATTCAAATATTCAATTTCGATATAGTCTTGAAATATGATCAAATTTATTATTGACACT
AAGAAATCTTATATATTGACTCTTGCCAAACACCATCGTTGCAACTGATGCTATTATGGAGGGCCAATTATTCTAGTTCAT
TCTCCAACATTGTAGTTGTATTGTACGTTAAAATATGAATCCTGATAAAGAAAAATTACACACATGTTAATTATAAACTAT
ATATGACATTATTTGTTATATTAATAAAGTTCAATTTATATCTTTTTTGCAATCAAGTGGTAGCTACAATATGGAGGGGAT
ACACCAAATTTAAGGTTGTCATCTATGTAGTATGTACTTCTATAAACGATGACATCAAGTGATTGTAAACATAAATGGTCA
ACATTCACATTGATTCACATGAAGGTCTACAATAGACTATCGTATAGACGATTGGAGAAATTGATATATGTCCACTACAAC
ATGTGACTAAGGTTGCAGTGTGCCAGGCTAAACAAGGAAATAGGATGAAGTGGGTCAAAGCAACAAAGAACCAAAAGGATC
CTTTACTTGATAAGGCAGGAGATCCTCCACGCGCTTCATGTTTTTTCATCGAGACAATAGAAGAAAATGAAGCACATCCCA
ACAGGAGGAAGATCCCCCTCGATCAAATGTAATGGAAGGATCCGGACAACATCGAATCAGACTACTTGAGGAACTAAAGAC
AACTAATAATCAGAGTCGTCCGCCCAGTGTGCAAATAAAACGACAAAGGGGAAGATAGTAGAATCAGTCGCACCGTTAGAA
AGAATCGAGTCGGACGACGAGACACCTCCATCACATTTATCAACACACTCGATTCATAGACATGGTAGCGACGTGAACAAC
AATGCCTCAACTAATGAAGGCGACAACGTCGGGGAGCTAATGATCCTGTCTACATAGTTTGAGGGTGGTGTATGGACCGAG
GACCAATATTTTATGTATGCCACTCAAGATTGATATCATGGAACTCGATGAAGTACTAGTCAAGTTTATACATGGAAGAGG
AAAGGGAAGGCAATGGATGATTTTAAACAGATGTGACAAAGCCTATATGATGTAAACACAGAGTAAAGATCATCGTATTCA
CAGTTGTGGTATTATAGAGAATCTTATGGCCAACAATAGTACGGTGATGGTCATTCTTCTCCGAGTAGTAGTACAGTGATT
GCTCTTATGATACGGATAGTAGATCAATGACGAAAGTAAAAGTTCTGATATTCTCCATGCTTTGCACCAATACATGTCACA
ATTATACTTGCGTTAGGAGCCGCCTTAGACATGTGCGATTCATGACAATCAGATTACAACCATCACTACATTAATGCACCA
ACAATATACGGTGTATCAACATATTATGTCGTGGGATTAATTTCATGATCGAATCCAACAAACATATCAAATTGATATAAA
AATACATAAGATCGACGATCTCGATCTACTGCCCGTGAAGTCTCATCTTTCTTTTTGGTGATAGATCAGATATATCCATTG
ATAGATTACTTGATTTATTTGAATCTTAAAATTTAAATTATTTAAAAAATAATCTAATAATTCAAATCATTTCTTTGTAGA
TAATTCAATATCAACGGAAAAACGATTCGAAATGTGCCACAAAACTACTCCTCAAAGCTTGAAAAAAGATGTCACAATTCT

TTCCTAATTTAGATTATTTTTTGCTAAAATTATAATAAATTTATATTTATTTCCAACTTAAAAACCTAATCTACTTTTTTT
ATTTTTTAGAAATTTTTGGAGTTATTTTCGACCATTTTTTCTGTGGCATCTATACATGACACTGTACTAACACATGGTATG
GTAGTACAAATCCATATGTACTATACCGACAGTCAATAAGTACTCTGGTATAGACCCATAAAACCAAGCGAAAGCAATAAA
GTACACACAAATTGAACCTTCAGACCAGATTGTTAAGAAATTATACAACAAAATATTTTAATTTGGAACAAAAAGAAGCAG
AATCAGTTTATGACTGTTCAAAAAATATGCCACTGAACCTAAAAAGGAATGCAGTTTGCTTGCCTGGTTATGATAAAAAAA
TCACCAACGTTTAATAACCATCTAGATGGCTGCTTCTAACAACACTGCTCCAATACAAATATGGTAAGTTGAATAAACTGT
TCAGGCTTCATTGAGAAAATTTTGATACTCAATAGGTGTCCATGGTCCAGACTAGCTGGGTAACATAACCAAACTCACATT
AACAACTTGGGAAGGTGGATTAGGTTGCAATATCCAATACAAGTTTAATAGATGATCAAGGCTCAAATTATAGTTACTGAG
TCCTAGTAAGATCCACAAAATCTAGATTACTAGAGAGCAATTCTAGAAATATAATTTCCTTTACAATCCATGATATAAGGT
TTCAAATTCCACAAAGAATTCATTAAAGAGTTAAATATTGTAACCAACTTCTAACATAAAAGACTTAAAAACTACCTAAAA
CCCTTAAACACTAGGATAGTAAATAATTTTACAGAACTACTTAGGTCTGGCCCCTGACTTATAATAGGAAGGCTAAGTGGC
C***CAACAACCTTG*****TTCCAAAA**

***MahAT2*: 3336 autonomous hAT in *Musa acuminata* (AC226047.1) from 34728-38063 bp**

GAAGGAAG***CAGTGATTTAAAAAGCGCT***GCGGGCGCTCGCCTAGTAGCTCGGGCAAGGCGAGGCGAGGCCCGAGCGCCTCAC
TTCACTTTCAGGCGGCGCGCTTCAAAGAGGTGCTGCTTGGGCGCTCGCCCGAGCCAAGCGCCCAGCGCCTGGGTTAAACCA
GGCGACCGAACCAGGATTTTAGGTCTGGTTCGATCTTGGTGCTTTAGTTGGTTTAATCGAACCAACTAAAACCGATATCAG
CTGTCGCTGCCCAACCCTAACCCTGCTCGTTGCCACTCTCGCTCCCGTTGCTTGCCACTGTCGCTGCTCGCTACCGTTGCT
TGCCGCTGTTGCCGCTCCTGCTCCCACTCCTGCTCCCGCTGCTCACCGCTGTCGCTGCTCGATGCCGCTGTCGCCGCTCCC
GCTGCTGCTGTCGCTTCCTCTTTTCTCAATCAGCACCCCTTTACACTCCTCTTCCTTCTTCTCCTTCTATATACTGTTAAC
AATATACTAATAATAGTATTTATATTTATTAGATTAATAATATATTATTATTATTTTAATACTGTTAATTTTTATTTATTT
GAAATTATTGTTAGGTTTCAAGATAAATGACAAGTGTACAGAGCAACTCAATAGTCTCCAATGGCATCAAAAAAAGATCCT
GCATGGAAGTATAATTATCTGAAGGATCCGAAAGATCATAATGCAGTGACTTGCATATTCTGCGATAAGACTACTAGAGAG
GGTATTTTTCGTGCAAAACAACATCTAATAAGAAATTTCAAGAATGCAAAACAAAAAGTGTCCACCTGAGGTAAGAGAAGA
GTTACTGATATATATGAATGAAAAGAAGATACAAAAGAATGAATCTTACGGGAATTTACCAGAAGACAATGTTGAACATCT
CAGGGATGAAGAAGAAGATTATTTTATGAGTATTAATCCAAGTAGAAAAAAAGTATACGACAAAAATTGAAACGAAGTTAT
GAGTACTAAGAAGGGTAAAAAAGGACCAATGAATCTATATATGTTTCAAGGATCCCAGAAACAACAAGGGCAAGTAGGAGG
CTCAAAATTTAGACAAACAAATATAAGTGATGCTTGTGATAAAAAATAAGAGGAAGAACAATTCAGCACATTGCTCGCTTC
TTCTATCACGTTGGTCTTCCCCTTAATACAACTCGTTTAGACAGTTTTAAGGATATGATTGAAGCTATTGGAAGATATGGT
GTAGGATTAAAACCTCCAAGTTATTATTATGAGATGCGAGTTCCATTGTTGCAAAAAGAGTTGAATTATATAAATGACTTA
CTAAAGGGTCATAAAGAATCATGGGCAACACATGGTTGCTCAATTATGTCAGATGTTTGGATTGACAGGAGGCGTAGGAGT
ATAATTAATTTTATGGTTAATTGTTCTTTAGGGACTATGTTTGTGAAGTCAATAGATGCTTCATCTTTTATAAAATCTGGA
GACAAGATATATGATTTACTTAACAACTTCGTGGAAGAAATTGGAGAACAAAATGTCATTCAAATCATAACCGACAATAGA
AACAACTATGTATTAGCTGGTAATATTCATCTTTTAAATTTTTTTAATTATTTTATCTTCAATTAAATGTGTTAAACTCTT
AAGTCTTATCATTTTTGTTATCCTTTGTCTCAGGTAAATTGCTTGAAACAAAAAGACGACACTTATATTGGACTCCATGTG
CAGCACATTGTATTGATTTAATGTTGGAGGATATTGGAAAGATCTTAGACATAAAAAAAAACCCAAGAAAGGGCAATTTTT
GTTGTTGGATTTCTTTATAATCATTTTGGGGCTTTGAATATGATGAGAGAATTTACAAGGAACAAAGAATTAGTGAGGTAT
GGTGTCACCCAATTTGCTACTTCATTCTTGACATTACAGAGCGTGCATCGTCAAAAACATAATCTGAGAAACATATTTACC
TCTGAGAAATGGGTGACAAGCAAATGGGCAAAAGAAGCAAAAGGCAAGAGGGCTACTGATATCATCTTAATGCCATCCTTT
TGGAATCATGTAGTTTATACATTAAAGGTAATGGGCCCTCTTGTTCGAGTCCTTCGGTTGGTGGATAATGAAAATAAGCCT
GCAATGTGATATATTTATGAGGCTATGGATAGAGCAAAGGAGACGATTAAAAGATCTTTTAATGAAAATGAAGAAAAATAT
GAGAAAATTTTTATAATCATTGACGAAAGATGGAATTGTCAACTTCATCGTCCCTTACATGCAGCAGGATATTATTTAAAC
CCTTAATTCTTTTATAAGATTAAATCTGTTGGGTTTGATGCATAAGTTTTGGATGGGTTATATCAGTGCGTTGTAAGATTA
ATTCCCAGCCTTGAGGTTCAAGGTAAGATTATTCATGAATTATCTTTATATAAAAATGCCGAAGGTCTTTTTGGAATTCCA
ATAGCCGTTCGATCCAGGACAACTACGTCTCCAAGTATTAATAATTTGATATAATTAATTTCATATATATTATGTTACTAT
GTTATTGCTAATAATAACATAAATTTTATAGCTGAATGGTGGAGTCTATTTGAAAATTCCACCTCGAACTTACGGCAATTT
GCTATCAAAGTACTTAGTTTGACATGTAGCGCTTCGAGTTGTGAGCGAAACTAGAGTGTCTTTGAGCATGTAAGGATCACC
AAATATTTTTGTTTTAATTAATTTATTTATTTAGATAGATGTATTAATATATTTTTAAATTATATGACATGACAGATTCAC
TCGAAGAGAAGAAATCGGTTGGAACATCAACGATTGCACGATCTTTTTTACATAAAGTATAATTAAACTTTGAAGGCTCGT
CATAGATTGACATAAGGAGATGTGGCAAGAGCTTCAGGTGCTGGAGAATTACAAACATATACAAGACAGATGACAAAGAGA
AAAATGAGTGCAAAAGCATCAAGCTCGGCTCTTGCTATTATTGAAGACATAGAGAATGAAACATATTTTGATGAAGAGGAA
GAAGTCGAAGGACAAGAGGAAAATGACGAATTTAATGAAGATGATTTGTGTGAAAATGACGATAATATTGATTATGATGAA
TGACTTTAATGTAAAATTTTATTATTTTGAATTTTGAAACTTTTTGTTAATGTGACAATGTGATTTTGTATCTTAGATTTT
CTTAATTTAATAGCATATTTTTATTTAAAATTTTAAATAATTATATTTATTAATTATATTATATTATTATTTTTTAGCA
TCTTGCTTCGTTCAGACGAGCACTTGGGCCAACGCTTAGCGCCTCGGACGTTTTTGGACCTTAGCGCTT***AGCGCTTTTTAA***
ATCACTG**GAAGGAAG**

***MaMITE1*: 781 bp Mutator-like MITE in *Musa acuminata* (AC186955.1)**

ATGCGA***TGTCACGGCCTTAGCTGGAATTGCCTAAGGCGTGAGGCACTCTTGCAGCCAAGACGCAAACTTAGCTTGCGTTAC***
CTAAGTCGCGAGGCACCCTTGCGACAAAGACGCGAACTTAGCTTGCGTTGCCTAAGTCGCGCTTCGCCCTTGAGATATTGC
TCCGCAAAGATCAGCCCACTTGCAACCTCTCGCAGGTCCCGAAGGACCTGTAAAAGAGAAAG***TTGATTAGTTCGAAAGAAC***
GAGCGACGGACAAGTCCTAACATCTCGCGAAAAGAGGGGAAGCTTTACAAGCAATTCAGCGAGCATCTTGTGTGCACAAGA
GAAAAGAAGAGAGGGAAAAAACAAAGACTTTAGAGGGTTGAACGAACAGCTGCAACGGGTGCCGGGCGCGACAACAAGTTC
CCGTCAAGGTAACGTGCGAACTTGCGAAAGGTTGTTCAACACCCGACACCCGGTGAGCAGTTGTCTGCGGAC***TTACAGTTG***
TTCGTTCAACCCTCTAAAGTCCTTGTTTTCCCCCTCTCCCTCTTTTCTCTTATGCACGCAAGGTGCTCGCTAAATTGCTTG
TAAAGCTTCTCCTCTTTTCGCGAGACGTTGGGACTTATCCGTCGCTTGTTCTTTCAAACTAATCAAGACCTGCGAGAGATT

CCAAGTGGGCTGATCTTTGTGGAGCAATATCTCAAGGACGAAGCGCGACTTAGGTAACGCAAGCTAAGTTCGCGTCTTGGC
*CGCAAGGGTGCCTCACGCCTTAGGCAATTCCAGCTAAGGCCATGACA*ATGC

***MaMITE2*: 664 bp Mutator-like MITE in *Musa acuminata* (AC226196.1)**

CTACTATAG*GGGTTAATTGTAGTTAGACCTCCTTGGCATCTCGATTCCTATATTTCAAAAATTTAATTAGCATCTTTATAG*
TTATGAAAGTGAAACATTTAACCTCATTTATTCTAATACCATCGGCTTTACCGACGAAAGATATAACACATGATAGCATAT
ACAGAATTATTGATTTTACTAATTATAAAATGATCATTTCAGGTAATAATTAAAAAGTGAAAGAAAGCAGCCATTGACTTA
*GGCAGTCACATAATTGGTTTCTTTCCCAATTGGATCTCTTAAGGTTTA*CAGCAAAGGTCCTCTACAACGTGGCAATGCTAA
GTCAGTCT*TAAACCCTAAGAGATCGATTTGGGGGAAGAAATCCCAACAGATCGATTTGGGGAAGAAATCAATCATACGACT*
ACCTAAACCAATGGCCCCTTTCTTCCACCTTTTAATTAACACCTGTAATGATCATTTTGCCATTGTCAAAATCGATAATTC
CGTACATGCTGTCATATGTTGTATCTTCCGTCAACAAAGCTGACGGCGTTAGGGTTGATGGGGTTAAATGTTTCGCTTTCA
TAATATAGGGATGCTAATGTAATTTTTTTAACTCTAGAGATCAGAATGCTAAAGATCTCTAACTACATGGATAATTTATAA
*TTAGCCC*CTACTATCG

***MaMITE3*: 1042 bp Mutator-like MITE in *Musa acuminata* (AC226047.1)**

TTTGATTTCA*ACTAATTACATATTACCCCCCATAGTTAGCTACCTTTAGCATCTCGGTCCTTACACTTCAGAAATTTACA*
TTGGCATCCCTATAGTTACGGTGAAACATCTAAGTCAATTTACCCATACACCATTAATTTTACCGATAGAAACATGAAAAT
AAAGAGCAAAAAGGTACTCTTAATGTTTCAGTTGGTCATGGCAAACAACGTCAGTGGTGGCAAACGGCAACGCTGCGGGTG
ACTGTTGTGGATGAGGAGAGCGACGATGAAAGATAAGGGTCGCTATGCATCTATATCGACATCGATGCAGTTGCTGAGCGG
*GGAAAGGGT*CATTGATGCAGATGTAAAGTGATGCTACTCGGTATATGCATCGATGGCCCTTTTGCTCGGCAACTACGTCGC
GCCGACGTAGATGCAAAGTTGCGTCGCTCAGCATCTGCAACGACGGCCCTTGCAGATGCAGAGTGATGTCGCTAGGCAACT
ACGTCAACGTCGACGTAAATGTAAAGCAACGTTACTCGACAATTGCATCGACGTCGATGCAGATACAAAGCGACGTCACTC
GACAACTACGTCGGTCGCTCTGCATTTGCGTCGGCGCCTACGTAGATGCAGAGCGACACCGCTTTGCATCTACGTCAACGC
TGATGCAATTGTCAAGCGACGAAAGAGCCACCAATGCAGATGTCGAGCA*ACCCTTTAGCCACTCGACAATGCCAACGACGC*
CGACACCGATGCAGAGGCACTCTTACCTCTCACCGCTACTCTCCTCATCCATAATAGCCACTCGCAACTCTCAACAATGCC
ACCATCCGCCACCACCGACGTCATCCACCATCGTCAACAAGAATGCTAAAATTACCTTTTTGTCTTTTGTTTTTATATTTC
CGTCGATAAAACTAACAAGGTTATGGTAAATAGACCTAAATGTTTATTTTCGTAATTATAGGGATGCCAATGAAACATTTT
*AAAGCATAAGGACCGAGATGCTAAAGGTAGTTAAATACAAAGGGTATTCTGTAATTAGTT*CTTTGATTTC

MaMITE4: 2067 bp Mutator-like MITE in *Musa acuminata* (AC186754.1)

TAATTACAT*ATTACCCTTCTAATTAACTATGATTAGTGTTTTAATTTTTATATTTTAAAAAATAATATTGGGATCTTTATA*
CTTATGAAAATAAAATATTTAACCTCAATTCTTCGACGAAAAAAATTACAAAGTGTTATCAAATGAAAAAAATATGAACAA
AAATAATAAAAATAATTTATTATCTTTTAAATTATTAATTTCATATAAACTTCTCACTCGTAACATAGAGTTTTAATAGAA
ATTGCTTTTATTCTCTTCGGTTCCCTCCCTTTCGATTGTATTACAAATGACAAACTAGCGACGATGTCCGTAACGCGCAAG
GCCAATGGCATTCTAAAAGAAATATGAGCGTTGCCTTTGACCATAATTGATTTCCTCCTCCCCTCTCTTTTGCACTAGCGT
*GATGTCACTAGTCCATGGCAACCAGCCCTTT*GTCCTCACGATCTTTGCTAAGTTGCAATCAAGGAACACTGTCCACAAGTA
CTCCTTCCATGGCCTCCCAAGGTACGTGCGGTGGACTGCAAGCTTTAAGTAGTAAAGTGTGAGGAACTCGTCGCTCTCGTT
GATGGTGTAGTGCAAAACACGAAGTCGGTGGAGTGGGTGGGGTCGGTGCGATCGTGCACGGTGATGGTGTTGGACTCGCCG
TGCTCGAGGTTAAGCTGGCGGGGGAGGATGAAGATGAGACAGTCACAGAAGAAGGTGACGGGGTTGCCAAAGATGAAGTCG
ACGAAGCCAGAGATGCAACAAGACTTGTAGAATTGGCGGAAAGAGTGGGCGTAGAGGGTGTCCTGGTGGCCAAGGTACTCC
ACCACCTGGTGCTTGTCCGACTTAAAGATGTTGGCGAAGGTCAAGTCTTGGGCCACGTCTAGGTCAAGATTGGACTAAGGA
TTTGTGGCAAATTCTTATTGATTATAGAAAGGAAACTATAAGTGGTAACTATGAGGAAAGTTAAGTTTGTACTCATTGAAG
GTGGTTGTGTTATAGGTGGAGACGTTGACTGTGTCAACGTTGAGGGAACCCGTAATGATTATTTTGCCCATCCCATCCCCG
ATAAATTCCAAGTTCATCTTCTCAAATGACACTCGGATCATCTCCTCGTGAACCCCTTTCTTTATGTAGATTATGCATCGA
TTGGCGTTGTGCTCTAGCGTAGCGTTGACCACCACCTATACCACTCCGAACTGGCACTCACCATCCTTGTAGACCGTTGCA
TCTAACAGTGTCCTTGGTCGTAGGAACTCCTTCCGGATGGCCTCTCTGCTTGGTGATGTCGGGGTGACGTTCGGGCAGTAC
CCATCCCTCTTGATCTGCAATAGGGCCCAGAGTGAAATGTCGACACACTACGGGGTGGCCACCATGGTGATGGCGCCGTCG
ATGACATTAGTGAGGTTGGCAAGGAAGGCCATGGCATCTGCCACCTAGTAGGTATCGTTGATGTACTAGTGGGTAAACCAA
TAATCGTACTGGTAGAGCTCCACGACATGTGAGCGAGCAGGTGCAATGGGCAATACTCGTACAGCACTGGGGGATCAGCAT
TCTCTCACGTATAGAACTAGTTGCGATAGGTGAGCAAATGTGATAGGTGTTCATCTCCAATGGGCAAGCAAGATCAAT*AAA*
GGGGTGATGGGCAAGCAGGTTGGAAAGAGGACGGGAGGAGGAAATTAGTCACGGTTGAAGATAGCGTCCAGATCTCCTCTG
GAACGCCACTGGCCTCGTGCGTCATCGATGTTGTCGTTGGTCTACTATCGACAACATAGTCGGAGGGGAGGGAATCGAAGT
GAGGGAGAGCGATTTCTTTATTAAAGCTTCTACACTACAAGCAAAGAGTTTATATGAAATTAATAATTTAAACGATAATAA
AATTATTTTTTTATCTTTTTCATTCGTTTTTTTTTTGTATGTGACAACACATATAATTTTTTCCATCAACGAAATTAATGA
CGTGAGGAGAATTGAGACTAAATGTTTCACTTTCGTAAGTGTAGGGATCTCAATACTATTTTTTAAAATATAGGGATCGGG
*ATGCTAATCATAATTAACTAAGAGGATAATATA*TAATTACAT

***MaSTE* transposon in *Musa acuminata* (AC186955.1)**

CATAA*TGTAACAcCCTTGA*TTAGTCTCACATCGAAAATGGGCAAGATTAAGATTGACTTATAAGGGTCTGATGAGTGTATT
ATTATTATTAACTTCAGCTTAAGTATTTTGGTCAGTGGTTTAGACCAAATGAAGTTGATAGGCTAGTTAGCCCATCAGGCT
CGGGTCATGATATTTGGTATCAGAACAATATTTCGAGGACATATTGTAATATGGGATCCTAGGACTAGAATATTAGATGAC
CTAATTAAGTGTTTGAGATTTAGATTTTGAGAAGTAGATTAAATACAACATGTATTTAAGATATTATAATCATTTTTAGTC
GAATATAATTTTTTTTATAAAAATAGTTGATTGAAGATGTTAGTTGAATTTTGTTAAATGACTTGAAATAGGATAATAGAA
GATCAAAATTTTAAATTTTCACTAAATTAAAATAAGTGTCATATTAGTTTATCTTATAAACTCCCATAAGATTTTATATTT
TGAATATGATTTTTTTCATGATCCTCTTGTCCCCTTTTTGAAAAATAGATTAGGGATGTTAATTTAATTATAGAATATTTA

TTTTATTTTTGAGATATATATGACTTAGTTATTCATTAGGATTATAGTAACTTTAAGTCATTCTCTCTTAATGTAAAAGTT
TTGATAATTTTAAGTGTAAAAATTTTATAGAAACTTTATAAATATAAAGTATTTACCTTTATAATTTTTATTACCATATAT
TATTCTTGACATATGCTAGGATGGTTTGTAGAAGTTATTGTTGACATTAATAAAATATGTGACTTTAAAGTTAGAATACAT
TTGATCGTGCATTGATATTTAAACTTAATTGTATGAGTAATTAATTTAAATATATTTTTGATTTAATATAAAATTAATTAT
AAGATATTATTTGTTTAATGAAATGACGAGTCTAATGAGAATTAGATATAATGTTATTTCTTTGACACATATAAAATATAA
ATCATTCGAAATTATAAAAGAATATATGCTGAATAGTTGGATAAATTTTCATCATTTCAATTGCTTTTGTTGAGTTGTACC
TCAACTTATATTTTATGTGAACAATCAATTAATTTTTAAAACTTAAAAATTTTATTGATTTATTAGATATAGAATTTGATA
TACTTGGTAATGTAAAACCAAATTTATTCAATAAGTTTCATTAATTTGTCATATATTTTATGTTAAAAATATATAAATATA
TATTTTTGATTGAAAAGTTTTCGAGTGGTTCATTTAATTAAAATATTTTTCAAGCTAATTTAGTTTTGTCCGACTTAGTAT
TTCAAAAGCTAACTATATTGATTAATGTTATTCTTAATCAAATTTGGAGACCAAATTTCTATTAAGTGGGGGAGAAATGTA
ATCACTGATAATTTCTCCTTTAACATAAATTAATTACTATAAATAATATATTACTTTATTATTAACTTCTATATTTGTGTA
TTATAGTTAAGGAAATGATTAAATTTGGTTTCCTTAATCGTACGGATAAGTTATGAATTCTACTAATCAATTAAATTATTA
ATTGATTTATTTCCTAATGAATGATTTGATATTTAGGAAGTAAATAGTTTACATTCTATTTTTTGTGTGTGAACTATAAGA
AATAAATATTATTTTATTTTTTTTTTTATGTGAAAGCAAATGTAAAATAAATTAAAAATAAAAATTAAATTATTACTTACCT
TTCGGGGAGATCTCATCTTCTCCTATATAAAGGGACTATTCATCCCTCATTCTCTACTAAGCCTAACAATAGGAGGATCGA
GGAGATGACTTCCCGGAGAGATGATATCGAGGAAGGGTATCGAAAAAAGATAGGATCCTTTCTTTTTAATTAGCTTTGATT
TATGTTTTAAAATCTAGATATTAAGATTGTTACAATTTTATAATGAATGATGATATTTTAATTTCGAAATTTTATGATTTA
TAAATAATAATAATTGATTTGTGATGTTAGATTGATTATTTGATTTATATTGAACTTTTAAGCCTAAGTAAATTTTAATAT
TAATTTAATTATTAAAATTATTATTTTTGGATTAGTATAATAGTTCTAATTTATTTAATTAGATATATCTATGTTTCAAGA
ATTATGTATGAATTTCTGTAATGTAATTAGTTTTAAATTTTAGATCATGAATTTAAAGTTTTAATTAATATATTTGAGTCA
TTATTTAATAATTTTAAATATTAAAATTTAATTTGATAAAAGGAGTCTAATTGGGGTTTGACTTGAATCAAGTTGATCGAT
CAGATCAAATTGACTCAACCCATGTGGTCATGTACAAACCTAGCACATGTGACTAAGTACAACCTAGCTCATATGACTCGG
TATGACACAATCTATGTGGCCGGATAAGAGACAACCCATATGACAAGGTACGACCTAACCCATGTGGTCAGGCATGACCCA
ACCCATGTGGCTAGGTACGATCTAGTCCATGTGGCTATGTATGACTCGGCCCATGTGGTAAGGTACGACCCAACCATGTGG
CTAGGCATGGGCCAGCTCATGTGGCTAGGTACGACCTAGTACATGTGGCCGTGTATAACCCAACTCATATAGTAAGGTATG
ACTTAACTATATGGTCACGCATTGGCTAGCTCATGCAGCTAGGTACGACCTAATCCATATGGTTATATACAACCCCAACTC
ATACGTTTAGGACCAACTCGCGAGCCAAGTAAGACCTAGTTCAAATGGTAAGGCTCAGCCCAACTTATAAGTGAGGCTTAG
CCTAGTTCATGAAGCTAAGTTTGCCGATTGGCATTAGACATATCGTCGCTCTTTCTATATAGCCTATCGTACCTCAACTCA
ACCTAGTACTTGGAACATGTATATGCAGTGTATGTGACCCGTCCAAGACTCAGGTGGGTTGGTTCGGTTTGGGTTAGCACT
ATACGACATAGTCCAATCTCCTTCCTCTTTATTGGTTTGAGCTCATTAATTGACGAAATCAAACAAGTTTGATTAGTTAAA
ATCAAGGTGATTCCAAACTAACTAAAATTAGACTGGTTTAGGGTGATTCCAGACCGGTTCTATAAGGATTTGATATTGGTT
CCGTTAGATCATAATCGAATTGAGTTTGGTTTAAGTCAATTGAGCTATTCCAATTAGGGTTTTGATTGATTGAAGACTATT
GAGATATTAATATTTCATGCTTTTATGATTTGATGAATTTAAAGATTTTATTTGAAGATATCGTTTGAAAATGATTATTGG
TTTTTTATGTTTATGATATTGATATGATAAGTATCATGTAGTAGATGTGACTAGACTAGTAGTTCACATTTAGAGTGCAAC
CTGTGAAAGGAGCTACGGGCCCATTATAATGCGGAGCCACCAATGAACATAGTCTAGTATATTTACATCAAGTATGGTCTC
TCATAATGTTTAATCATATTAATTTCTTGATGTGTGCATGAGTGATTGAATGAGTATAAATGAATGATCATGGATGTTTAT
GTATCCCTGCCGAGGGCAGGTATGACGATTCCCTTCGGGGATTAGCGGGCCGTCAGATGTGATGGTTAAACGGTTTCTCCA
TCCGTTGTTGGGAGGAACGTGTCCCCACTTGGGCGGGTGAGGGATACCACTCCATCCGCTATTGGGAGTGCGATATGGATT
ATCTCCATCCGCTGTTGGGAGGAATCCATCTCGTGGAATATGCCTCTCCATCCACTGTTGGAGAGTGCACCATTAGGTGAT
GTACGCCTTTATTATTATGTATGTGCTGCATGTGGTTGATGCATGATTTTGTTCATTATGTAAGAAATTATCATCATTTTT
CTGATGCATAAGTTTTTATGATGAATAGATATATTTTTATATTAAATTATTGAAATTTGTTTATATTTCACGAATATTGAA
TATTATTTTTATGATTTTTCTGAGGTTCCTACCAGCATGTGTGGTTGCTGATGTGTTTTTATTTTATATTTATTTTCAGAG
TAATCTACTATTGTGTAATAGATCTGAAGCTTGGGTGAAAGAGACTTGTGACCCTGTGAAGAAGATGTCATTTTTCTAGCT
AATAGAATATTTATTATTATAAAGACAATGATTTGAATTCAAAGCCAAGTAAAAAGAAAAGAAGGGAGTTTATTATGTAAA
TTATGAATAATAAATTATTGATTTATTATTTTTTTTATAAAT***TCAAGGaTGTgACA*****CATAA**

***MAWA* transposon in *Musa balbisiana* (AC186754.1)**

AGTGCTAGTCATAGGTGTCTTACAAGCCAATCACGTTAATAATGACACATGTGACATGACATGCACTCTTTTTGCTTATTA
TTATTATGATATTTTCTCACTTTATATTGCTTGATGTATAAATATATTGTGATGTCCATGGATTTGTGCAATGGGAATCGG
ATCATGATGAGATCGTAATAATGAGAGTGATTCACCTCTAAACACAGACATTAAATAATCATGATCATAGGTTACTCGAGA
TAGACATCGAGATAATTGGACAGACTGGTGTGCTGTATACCCATCCATATGATGGAGGTAACTGATCTCATAGCTACTCGT
GTGGGGACAGTAGGGCTACAGTGCATGTACTCATTGGAGACTGAGTTTACTGATTGATCCACTCACGAAATGCTGGATGGT
TGATGATACCTCATTGTCAGACAATAATTCCGTTGTCCCAGTGGTGTACTTGGTCCTTAGACTTGAGATACTAAGGATGTT
CTGTATGAGTACTCAACTTTTTGATACCGACCTTATAGGTTTGAAATTTCAGATGTAGCACAGTTGGTCATCGGAAGTGGC
AGCCAACCTTACGAGGGCTATTGAGTGTCGATAGAAAATCATCCGCTCTCAATATCATAAGAGGAATATCTCATGTATTCT
TGTTCAGACAAATCCTTGACCAAGATCATTTGAAATAAGAGAGAAAGAGTTCTCCGGGAGAATTCGATTAGAGCAAGATTA
GAGGAGAAACCGTATGGGCTTGACAATACCATACCCGGTGTACGATTTCTAGGATATTAGATGGATAAGAGACCATAGGTA
CACGACAATTGAGGACAGATATGTCCAAAGGATTAGGTTCCCCTATATCGTCTAGGGACTACGACATACTGGCCTAGTACG
TCCGCAGTCGATGAGTCGAGTGAACTATTATAAAGATAATAATTCATTAAGCCGGAAGGAATTCTGACATATATGACTCAC
GGCCAGCTCAATATTGGGCCTAGAAGGTCACACATATATGGTAGGTGTTGCGACGAATAGAGGTTTAGATATGAGATATCT
GCCGAAGCCCCTATTTTTTTGGATATCCATTAAGCCCCTGAATTATTGAATCCTATAGATGAGATCCAATAAGAGCTAATA
AGAGATTATTGGATAGAGATCCACTAATCTAATAAACTTAAGTAATTGGATAGAAATTCAATACCCAATAGGGTAAGATCT
ATTAGGGTTAAGTTAATAGAGGACCTCTATAAATAGGAGGGAACCAAAGGGCCATAGCTAGGCTCTTTGACTGTCACCTCC
TATTCTCCTCTCCCCCTCTCCTCCTCAGCCTGCAACCCTTGTTTGAGGCGTGTGGATAGCAAGAAGGGTCGATCCCTTCTT
GATTGCGTGGTGCGCACAGTAAGGAGATTTGAGGAGCGTATTCGCAACCCTTGGCGTGTGAATCACCGCTAGAGATGAGGG
CGCTTGACTTCTTTCATCCCTCCCACAGATCTGCAGAAATTCATAGATATACGATCTTCCTATATAACACAACTATCTTAC
ACATGGTTTTCAGTTTCGTGAGTTTTTGCGCATCAATCTTCGTACGACGATAAACACCTTTCTGGGAAATCTAAGATTTTT
ATTTTTTGTTCTTCCGCTACGCATATAATGTCGCCCATAGATTTCCCTACACGG**AG**