TÜBİTAK

Turkish Journal of Biology

http://journals.tubitak.gov.tr/biology/

Research Article

Turk J Biol
(2015) 39: 740-757

doi:10.3906/biy-1501-77

Identification and evolutionary genomics of novel LTR retrotransposons in *Brassica*

Faisal NOUROZ[1,*], Shumaila NOREEN[2], John Seymour HESLOP-HARRISON[1]
[1]Molecular Genetics Laboratory, Department of Biology, University of Leicester, UK
[2]Molecular Genetics Laboratory, Department of Genetics, University of Leicester, UK

Received: 26.01.2015 • **Accepted/Published Online:** 23.03.2015 • **Printed:** 30.09.2015

Abstract: Retrotransposons (REs) are the most abundant and diverse elements identified from eukaryotic genomes. Using computational and molecular methods, 262 intact LTR retrotransposons were identified from *Brassica* genomes by dot plot analysis and data mining. The Copia superfamily was dominant (206 elements) over Gypsy (56), with estimated intact copies of ~1596 Copia and 540 Gypsy and ~7540 Copia and 780 Gypsy from *Brassica rapa* and *Brassica oleracea* whole genomes, respectively. Canonical Copia and Gypsy *gag-pol* polyprotein organizations were observed in most elements with a few displaying 1–3 additional or internally deleted domains. The PBS and PPT motifs were identified with tRNA complementary to $tRNA_{Met}$ or, rarely, other tRNA types. PCR amplification of RT regions revealed their abundance and distribution among A-, B-, and C-genome *Brassicas* indicating a common ancestor. The evolutionary relationship of *Brassica* REs resolved them into superfamily-specific (Copia/Gypsy) lineages. The phylogenetic analysis of 130 *Brassica* Copia clustered them into 2 clades and 10 sub-clades of 18 families; Gypsy elements clustered into 2 clades. The results enabled identification and understanding of the structure and nature of full-length REs and their derivatives in *Brassica*. The markers derived here will be useful for examining chromosome and genome evolution in *Brassica.*

Key words: LTR retrotransposons, *Brassica*, Copia, Gypsy, evolutionary relationship, RTAP markers

1. Introduction

The mobile genetic elements, transposable elements (TEs), 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; Kapitonov and Jurka, 2008). The larger genomes are made up of abundant tandemly repetitive sequences and TEs, which compose a major proportion of DNA, sometimes representing more than half of the genome (Heslop-Harrison and Schwarzacher, 2011). With advances in computer-assisted analyses and genome sequencing projects, it is now known that retrotransposons (REs) are important components of all eukaryotic genomes and play a major role in their evolution (Flavell et al., 1997; Wicker et al., 2007). The eukaryotic TEs are classified into two major types by many authors, retrotransposons and DNA transposons, based on their copy-and-paste and cut-and-paste transposition mechanisms, respectively (Jurka et al., 2007; Kapitonov and Jurka, 2008). Among TEs, the major proportion in plants is represented by long terminal repeat (LTR) REs, which reverse transcribe their RNA to generate DNA copy integration to new host sites (Eickbush and Jamburuthugoda, 2008).

LTR REs have been categorized on the basis of phylogeny of their reverse transcriptase (RT), *gag-pol* domain organization, proliferating devices, and structural features into superfamilies as Ty1/copia, Ty3/gypsy, Bel-Pao, Retrovirales, and ERV-like elements. Copia and Gypsy elements are the most abundant and diverse group of retrotransposons studied in several organisms (Wicker et al., 2007; Kapatonov and Jurka, 2008). They are characterized by 4–6 bp target site duplications (TSDs), 100–5000 bp LTRs, internal regions encoding *gag-pol* protein domains, a primer binding site (PBS), and a polypurine tract (PPT) at 5′ and 3′ LTR, respectively. The LTRs exhibit conserved termini (5′-TG---CA-3′) and carry the promoter elements, TATA box, polyadenylation signals, and enhancers, which regulate the transposition mechanism of LTR REs. The *gag-pol* encodes the 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 et al., 2007; Vukich et al., 2009).

Copia and Gypsy are two major superfamilies of LTR REs dispersed in plants that differ in order of protein domains encoded by the *pol* gene. The canonical Ty1/copia exhibits TSDs, LTRs, displays PBS/PPT motifs,

* Correspondence: faisalnouroz@gmail.com

and has internal *gag-pol* genes that encode the protein domains as 5′-GAG-AP-INT-RT-RH-3′ (Flavell et al., 1992b; 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 REs, which displays 5 bp TSDs, LTRs, and internal-region–encoding *gag-pol* protein domains as 5′-GAG-AP-RT-RH-INT-3′ or have additional domains. On the basis of presence or absence of chromodomain, they are divided into chromodomain- and nonchromodomain-bearing Gypsy; the former are most common in several plants (Novikova et al., 2008; Novikova, 2009).

The genus *Brassica* of family *Brassicaceae* includes several important crops such as oilseed rape (canola), brown mustard, Chinese cabbage, turnip, cauliflower, broccoli, Brussels sprouts, collards, and kale. They are used as valuable and long-standing food and oil sources in both developing and industrialized countries (Monteiro and Lunn, 1999). The diploid genomes of *Brassica rapa*, *Brassica nigra*, and *Brassica oleracea* have been named AA, BB, and CC, respectively, and they have resulted in allotetraploids such as AABB (*B. juncea*), AACC (*B. napus*), and BBCC (*B. carinata*) by hybridization forming 'Triangle of U' (Nagaharu, 1935; Monteiro and Lunn, 1999). Several nondomesticated *Brassica* taxa have been described, mostly related to *B. oleracea* (Ostergaard and King, 2008).

The present study aimed to identify elements in *Brassica* species with retrotransposon characteristics without relying on homology to known elements through dot plot analysis. Bioinformatics and molecular approaches were used to characterize the mobile genetic elements in the genome with the aim of studying the identification of novel retrotransposons, their genetic diversity, distribution, activity, and evolutionary impacts on *Brassica* genomes.

2. Materials and methods

2.1. Plant material for *Brassica*

The DNAs from 40 *Brassica* accessions (cultivars) were used. Seeds from 32 cultivars were brought from Warwick Research Institute, Warwick, UK; 4 (NARC-1, NARC-II, NARC-PK, NATCO) from the National Agriculture and Research Center (NARC), Islamabad, Pakistan; and 4 synthetic allohexaploids *Brassica* (Ge et al., 2009) from the University of Wuhan, China. The seeds were grown in green house at the Department of Biology, University of Leicester, UK, and DNA was extracted by standard CTAB method.

2.2. Dot plot analysis for identification of LTR retrotransposons (REs)

Ninety bacterial artificial chromosome (BAC; Supplementary Table; on the journal's website) genomic sequences deposited in the National Center for Biotechnology Information (NCBI) GenBank database were retrieved/downloaded and surveyed for the identification of LTR REs. A novel approach was used for the identification of LTR REs based on the dot plot comparison of BAC sequences against themselves. The candidates of full-length elements were identified by running each BAC genomic sequence against itself in a dot plot analysis in Dotter program (Sonnhammer and Durbin, 1995). The central diagonal line extending from one corner of the dot plot to the opposite corner represented the homology of the sequence. The LTRs on both termini were represented by 2 small parallel diagonal lines at opposite corners (Supplementary Figure; on the journal's website). The numbers of nucleotides in LTRs were counted, and TSDs were characterized by visual inspection.

2.3. Computational analysis and data mining for LTR REs

The intact or full-length (reference) elements identified by dot plot analysis were blasted against the *Brassica* Nucleotide Collection (nr/nt) database available in NCBI. In the database, searches for LTR REs were performed by using intact elements to find the full-length copies and their deleted elements as defined by Ma et al. (2004). For estimation of copy numbers, the number of strong hits against the reference queries, with >75% query coverage and identity in their entire lengths, were collected. Only intact elements were counted, and the following formula was used to estimate the copy number of intact REs: copy no. = no. of intact copies in database × total *Brassica* spp. genome size/available genome size in NCBI (Tu, 2001). The Conserved Domain Database (CDD; http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) available in NCBI was used to identify *gag-pol* gene encoding proteins in REs. The PBS and PPT motifs were detected in LTR_FINDER program by using the parameter 'Predict PBS by using *Arabidopsis thaliana* tRNA database.' The sequences showing >85% nucleotide identity in their coding regions were considered to belong to the same family, and sequences showing >95% homology were considered copies of a single element. A novel family was defined when no homology was found against known elements and the element showed homology to one or more sequences (Wang and Liu, 2008).

2.4. Characterization, classification, and naming of LTR REs

The Repbase (Jurka et al., 2005) and Gypsy databases (Llorens et al., 2011) were used to characterize the REs (on a homology basis) to known elements. Elements that failed to be characterized by homology searches against TE databases were characterized by visual inspection on the basis of hallmarks such as TSDs, LTRs, PBS, PPT,

and organization of *gag-pol* encoding proteins. REs were classified as Copia if they displayed *pol* gene as 5′-AP-INT-RT-RH-3′ and Gypsy as 5′-AP-RT-RH-INT-3′. The individual elements and their families were defined by the criteria recommended in other works (Wicker et al., 2007; Minervini et al., 2009). A novel family was declared when no homology with any known LTR REs was found (Wang and Liu, 2008). The names of elements were given according to Capy (2005); for example, ***B**rassica **r**apa* **Cop**ia **1** (*BrCOP1*), where the first letter indicates genus, the second species name, three letters indicate the superfamily, and the number indicates the family.

2.5. Polymerase chain reactions (PCRs)

The degenerative primer pairs designated as reverse transcriptase amplification polymorphism (RTAP) markers were designed from the conserved D-DD triad of RT regions with Primer3 (http://frodo.wi.mit.edu/primer3/). PCR was used for the amplification of RT fragments derived from LTR REs. Total volume of the reaction mixture ranged from 15 to 20 µL with 50–75 ng/µL of genomic DNA, 10X Kapa Taq buffer A (Kapa Biosystems, UK), an additional 1.0 mM $MgCl_2$, 200–250 µM dNTP (2–2.5 mM; YorkBio), 10 pmoles of each primer (Sigma-Aldrich), and 0.5–1 U of 5 U/µL Taq polymerase (Kapa Biosystems, UK). The thermal cycling conditions were adjusted as follows: 3 min denaturation at 94 °C; 35 cycles of 1 min denaturation at 94 °C, 1 min annealing at 52–64 °C (depending on primers), 1 min extension at 72 °C; and a final 5-min extension at 72 °C. PCR products were separated by electrophoresis in 1% agarose gel with TAE buffer. Gels were stained with 1–2 µL of ethidium bromide for the detection of DNA bands under UV illumination.

2.6. Multiple sequence alignment and phylogenetic analysis

The RT sequences (~170–220 aa) from identified *Brassica* REs were aligned in the CLUSTAL-W multiple alignments available in BioEdit (Hall, 1999). Small insertions/deletions were removed, and frame shifts were introduced in aligned sequences. Phylogenetic analysis was performed by constructing the neighbor-joining tree with 1000 bootstrap replicates implemented in Mega5 (Tamura et al., 2011).

The overall methodology can be summarized as follows: 90 BACs randomly collected from NCBI > dot plot comparison of each BAC against itself > retrotransposons highlighted in BAC sequences > each retrotransposon subjected to NCBI BLASTN searches to retrieve its complete copies > CDD used to detect domains in identified elements > elements characterized as Copia or Gypsy based on their *pol* polyprotein arrangements > LTR Finder used to detect the PBS and PPT motifs in elements > names given to elements as recommended by Capy (2005) > RT domain sequences aligned to detect polymorphisms and construct phylogenetic trees.

3. Results

3.1. Distribution and copy number estimation of LTR REs in *Brassica*

Ninety *Brassica* BACs (Supplementary Table) were screened for the availability of LTR REs. Seventy full-length (intact) retroelements (Table 1) from *B. rapa* and *B. oleracea* BAC clones were identified by dot plot analyses as belonging to Copia (55) and Gypsy (15) superfamilies. The dot plot analyses revealed that some BAC sequences showed multiple LTR REs, while others displayed only one or two copies or even lacked them. The *B. oleracea* BAC (accession number: AC240090.1; 117.7 kb long) harbored five Copia and one Gypsy element covering 33.3 kb (28.5% of the BAC; Supplementary Figure). Another *B. oleracea* BAC (AC183496.1) contained three Copia (5063 bp, 4616 bp, and 4001 bp) and a Gypsy element (11275 bp), representing 15.5% of total BAC size (Table 1). The intact elements (70) and their solo LTRs identified by dot plot analyses were used as query against the NCBI *Brassica* Nucleotide Collection (nr/nt) database; all full-length, truncated, and partial elements were counted. Around 14,904 copies of Copia and Gypsy elements and their partial fragments were retrieved from the database. Of the 14,904 copies, 262 intact elements belonged to Copia (206) and Gypsy (56) superfamilies. The ratio of intact elements to solo LTRs in *Brassica* BAC sequences was ~2:1. Based on the BLAST survey of intact (262) elements, ~1596 Copia and 540 Gypsy and ~7540 Copia and 780 Gypsy were estimated for *B. rapa* and *B. oleracea* whole genomes, respectively.

3.2. General characteristics of *Brassica* Copia elements

The investigated Copia elements were generally smaller than Gypsy elements with a size of 3.7–8.9 kb (Table 1). The smallest *Brassica* Copia was an internally deleted element, *BoCOP23* (3.7 kb), while *BoCOP22* (8.9 kb) was the largest Copia studied. Most elements were terminated by perfect AT-rich 5 bp TSDs (3 bp in *BoCOP25*), flanked by LTRs ranging in sizes from 121 (*BrCOP19*) to 587 bp (*BoCOP32*), and displayed the canonical Copia domain organization (5′-GAG-INT-RT-RH-3′). A few elements showed internally deleted domains (*BoCOP23, BoCOP26, BoCOP46, BoCOP47, BoCOP48, BoCOP49*), others captured one or more additional protein domains (*BrCOP2, BrCOP5, BrCOP9, BrCOP13, BrCOP19, BoCOP31, BoCOP35, BoCOP44, BoCOP55*), and very few lacked one or more domains in their molecular structures. The position of Copia in various BAC sequences was variable with few elements excised and integrated to nearby places and others integrated to a site away from excision sites.

Table 1. List of Copia and Gypsy retrotransposons with their sizes, TSDs, LTRs, and positions in BAC clone sequences.

Element name	Accession	Species	Size	TSDs	LTRs	Position in BACs
BrCOP1	AC189222.1	*B. rapa*	5366	GTGAA	539/541	54,707–60,072
BrCOP2	AC189222.1	*B. rapa*	4828	ATAAT	312/312	96,814–101,614
BrCOP3	AC189446.2	*B. rapa*	5778	CCTTT	493/493	74,000–79,760
BrCOP4	AC166739.1	*B. rapa*	6020	GTCAT	599/599	2956–8975
BrCOP5	AC155341.2	*B. rapa*	4807	CCGTC	180/180	67,278–72,084
BrCOP6	AC189472.2	*B. rapa*	5029	AGTTG	159/159	51,849–56,877
BrCOP7	AC189496.2	*B. rapa*	4481	ATTAG	152/152	72,529–77,009
BrCOP8	AC189496.2	*B. rapa*	4971	CCCTG	385/385	86,234–91,204
BrCOP9	AC241035.1	*B. rapa*	5313	GGATG	407/488	77,808–83,120
BrCOP10	AC241108.1	*B. rapa*	6489	AACCT	306/299	74,968–81,456
BrCOP11	AC241191.1	*B. rapa*	5630	ATTAA	304/304	60,038–65,667
BrCOP12	AC241195.1	*B. rapa*	4672	TATCT	147/147	5590–10,261
BrCOP13	AC241195.1	*B. rapa*	4117	GTAAG	127/127	54,558–58,674
BrCOP14	AC241196.1	*B. rapa*	4595	AACTT	228/230	2514–29,738
BrCOP15	AC241196.1	*B. rapa*	4585	CTCTA	172/172	80,837–85,421
BrCOP16	AC241197.1	*B. rapa*	4940	CTCTT	345/345	134,939–139,878
BrCOP17	AC241198.1	*B. rapa*	5010	GAACC	170/170	17,376–22,385
BrCOP18	AC241200.1	*B. rapa*	6096	AAAGT	399/399	46,476–52,571
BrCOP19	AC241200.1	*B. rapa*	4196	CACAA	121/121	61,155–65,350
BrCOP20	AC241201.1	*B. rapa*	4838	GAGGT	182/182	35,112–39,949
BrCOP21	AC241201.1	*B. rapa*	5089	ATAAT	266/266	95,924–101,012
BoCOP22	AC149635.1	*B. oleracea*	8922	TAGCT	579/582	23,364–32,285
BoCOP23	AC149635.1	*B. oleracea*	3757	GACTA	296/296	71,762–75,458
BoCOP24	AC183496.1	*B. oleracea*	5063	GAAGT	429/425	34,468–39,530
BoCOP25	AC183496.1	*B. oleracea*	4616	TCC	221/221	146,660–151,275
BoCOP26	AC183496.1	*B. oleracea*	4001	GTGTA	425/425	251,315–255,315
BoCOP27	AC183492.1	*B. oleracea*	4790	CCCCC	368/368	38,224–43,014
BoCOP28	AC183492.1	*B. oleracea*	6395	CATAC	333/333	50,944–57,338
BoCOP29	AC183498.1	*B. oleracea*	6576	ATATT	288/318	162,553–169,128
BoCOP30	AC240087.1	*B. oleracea*	4682	AGTTT	268/253	71,136–75,817
BoCOP31	AC240089.1	*B. oleracea*	6230	ACAAT	249/249	11,346–17,575
BoCOP32	EU568372.1	*B. oleracea*	6160	TGAAC	577/587	31,626–37,785
BoCOP33	EU568372.1	*B. oleracea*	4660	ACTTT	201/252	56,936–61,595
BoCOP34	EU579454.1	*B. oleracea*	6060	ATTAT	233/244	48,881–54,940
BoCOP35	EU579455.1	*B. oleracea*	4769	ACTAA	392/392	61,558–66,325
BoCOP36	AC240081.1	*B. oleracea*	5108	GCACT	366/366	41,065–46,172
BoCOP37	AC240081.1	*B. oleracea*	4879	TTGTA	170/170	59,406–64,283
BoCOP38	AC240082.1	*B. oleracea*	7097	TAAAT	313/313	2322–9418
BoCOP39	AC240082.1	*B. oleracea*	5371	TACAG	304/293	61,467–66,837
BoCOP40	AC240083.1	*B. oleracea*	4778	AAGAG	370/370	43,143–47,920

Table 1. (Continued).

Element name	Accession	Species	Size	TSDs	LTRs	Position in BACs
BoCOP41	AC240084.1	*B. oleracea*	4690	CCTTA	300/303	66,766–71,455
BoCOP42	AC240085.1	*B. oleracea*	4656	GAACA	264/264	71,673–76,328
BoCOP43	AC240087.1	*B. oleracea*	4682	AGTTT	268/253	71,136–75,817
BoCOP44	AC240088.1	*B. oleracea*	4802	CATTG	321/320	48,706–53,507
BoCOP45	AC240088.1	*B. oleracea*	4706	GACAT	400/400	57,933–62,638
BoCOP46	AC240090.1	*B. oleracea*	4450	CTTTT	366/366	8583–13,032
BoCOP47	AC240090.1	*B. oleracea*	4616	CTATA	366/366	42,364–46,979
BoCOP48	AC240090.1	*B. oleracea*	6096	TAAAT	257/248	90,035–96,130
BoCOP49	AC240091.1	*B. oleracea*	6096	ATTTA	248/257	28,774–34,869
BoCOP50	AC240090.1	*B. oleracea*	4748	AAGCA	263/263	63,073–67,820
BoCOP51	AC240091.1	*B. oleracea*	4748	TGCTT	263/263	57,085–61,832
BoCOP52	AC240092.1	*B. oleracea*	4763	GAGAC	288/288	15,999–20,762
BoCOP53	AC240092.1	*B. oleracea*	5887	AATAG	200/198	71,126–77,012
BoCOP54	AC240093.1	*B. oleracea*	4703	TATCG	273/273	41,973–46,475
BoCOP55	AC240094.1	*B. oleracea*	6131	AATTA	251/250	36,442–41,571
BoGYP1	AC240090.1	*B. oleracea*	9161	CAAAA	2004/2035	27,208–36,368
BoGYP2	AC183496.1	*B. oleracea*	11275	GCTGA	1140/1272	283,163–294,437
BoGYP3	AC183498.1	*B. oleracea*	11845	GTGTT	471/476	257,711–269,554
BrGYP4	AC241108.1	*B. rapa*	11744	GATTC	480/480	31,686–43,429
BrGYP5	AC189430.2	*B. rapa*	11872	CTAGG	480/480	107,900–119,771
BoGYP6	EU579455.1	*B. oleracea*	11576	ATGGC	508/509	13,914–25,488
BrGYP7	AC232508.1	*B. rapa*	11664	ATCTT	506/506	118,772–130,435
BrGYP8	AC241108.1	*B. rapa*	5094	TGGGG	331/331	74,345–79,439
BrGYP9	AC241195.1	*B. rapa*	5900	GATTG	346/339	43,731–49,630
BrGYP10	AC189263.2	*B. rapa*	5221	CAAGA	346/346	38,008–43,228
BrGYP11	AC189218.2	*B. rapa*	5173	CTCTA	340/343	68,590–73,762
BrGYP12	AC155338.1	*B. rapa*	5163	CTTAA	360/360	110,515–115,677
BoGYP13	AC240081.1	*B. oleracea*	4168	TGCGC	199/200	89,533–93,700
BrGYP14	AC189233.2	*B. rapa*	7195	ATCAT	1553/1553	66,972–74,166
BrGYP15	CU984545.1	*B. rapa*	5140	GGGAA	369/369	74,632–79,771

3.2.1. Structural features of Copia elements identified from *B. rapa*

A lot of variation in sizes, TSDs, LTRs, *gag-pol* gene domain organizations, and heterogeneous structures were studied in various retroelements. *BrCOP1* identified from *B. rapa* accession AC189222.1 was a 5.3 kb element (Table 1) flanked by 5′-541/539-3′ bp LTRs and PBS/PPT motifs and displayed the canonical Copia 5′-GAG-INT-RT-RH-3′ structure (Figure 1). A 4.8 kb element, *BrCOP2*, was flanked by 312 bp LTRs, and it displayed 5′-GAG-AIR1-ZK-INT-RT-RH-3′ *pol* domains. *BrCOP3* (5.7 kb) displayed PBS/PPT sites with an extra motif (ZK) incorporated in *pol* gene (Figure 1). *BrCOP4* (6.0 kb) was flanked by 599 bp LTRs, *pol* gene encoding polyproteins, and PBS/PPT motifs. *BrCOP5* was a 4.8 kb element flanked by 180 bp LTRs and exhibited the PBS/PPT motifs and a typical Copia domain organization with an additional protein. *BrCOP6* was similar to *BrCOP5* but larger in size with small LTRs. *BrCOP7* and *BrCOP8* were 4.4 and 4.9 kb and flanked by 152 and 385 bp LTRs, respectively (Table 1). *BrCOP9* (5.3 kb), flanked by 5′-488/407-3′ bp LTRs, displayed the PBS/PPT motifs and typical Copia *gag-pol* genes with an additional (HVE) domain (Figure 1).

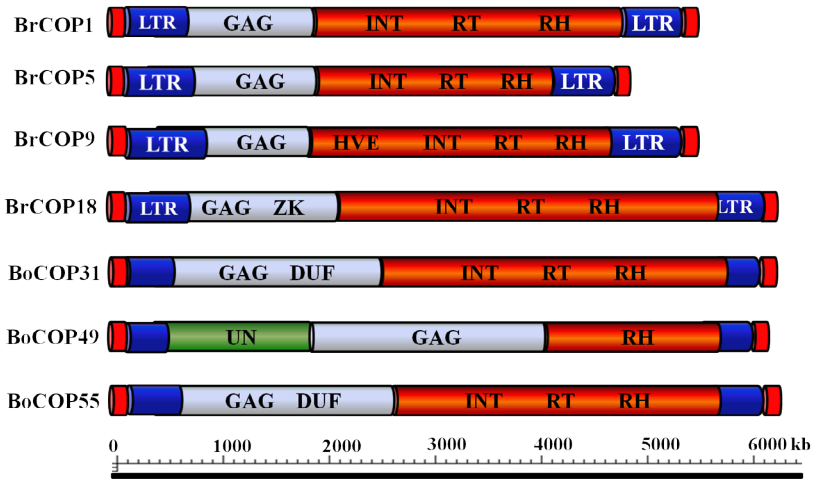


Figure 1. Structures of a few Copia elements in *Brassica*. The discs at the ends represent the TSDs. LTRs are drawn internally to TSDs (blue). The *gag* and *pol* regions are drawn with their protein domains. Scale below measures the lengths of the elements (bp). AP: aspartic protease. RT: reverse transcriptase. INT: integrase. GAG: *gag*-nucleocapsid. ZK: zinc knuckle. DUF: domain of unknown function. UN: unknown.

The *BrCOP10* was a 6.5 kb large Copia including the 306 bp 5′LTR and 299 bp 3′LTR. Two elements, *BrCOP11* (5.6 kb) and *BrCOP18* (6.1 kb), showed >90% similarity in their RT-domains. *BrCOP12* and *BrCOP13* were 4.6 and 4.1 kb and flanked by 147 and 127 bp LTRs, respectively, with an extra phage virion morphogenesis (PVM) protein domain in *BrCOP13*. *BrCOP14* and *BrCOP15*, both 4.6 kb, were identified in *B. rapa* (accession number: AC241196.1) with varied TSDs and LTRs (Table 1). *BrCOP16* and *BrCOP17* (4.9 and 5.0 kb) shared structural features including LTRs of 345 and 177 bp, respectively. They displayed PBS next to 5′LTR complimentary to $tRNA_{Met}$ and 15 bp PPT adjacent to 3′LTR. *BrCOP19* (4.2 kb) displayed the shortest LTRs (121 bp). *BrCOP20* and *BrCOP21*, identified from the same BAC (AC241201.1), showed a distinct mode of *gag-pol* domain organization and varied PBS/PPT motifs and LTRs (Table 1).

3.2.2. Structural features of Copia elements identified from *B. oleracea*

The largest (8.9 kb; *BoCOP22*) and smallest (3.7 kb; *BoCOP23*) Copias were identified from the *B. oleracea* accession AC149635.1 (Table 1). *BoCOP22* was flanked by ~580 bp LTRs and exhibited PBS/PPT motifs, an extra AIR1 domain, and an unrelated insertion towards the C-terminus (Table 2). Detailed analysis of the element revealed the most heterogeneous sequence in comparison to other Copia investigated. *BoCOP23* was internally deleted with 5′-RT-RH-3′ domains only. Three REs identified from *B. oleracea* (accession AC183496.1), *BoCOP24*, *BoCOP25*, and *BoCOP26*, were 5.0, 4.6, and 4.0 kb and flanked by LTRs of 525, 221, and 525 bp, respectively (Tables 1 and 2). *BoCOP27* displayed a size of 4.8 kb including 368 bp LTRs and terminated in perfect 5′-CCCCC-3′ TSDs. *BoCOP28* and *BoCOP29* are 6.4 and 6.6 kb and flanked by 333 bp and 5′-288/318-3′ bp LTRs, respectively. *BoCOP30* and *BoCOP31* are 4.6 and 6.3 kb, terminated by 268 and 249 bp LTRs, respectively, and display the PBS and PPT motifs (Table 2). *B. oleracea* BAC clone 'EU568372.1' harbored *BoCOP32* and *BoCOP33* with a size of 6.1 and 4.6 kb, variable LTRs, and similar *gag-pol* proteins (Table 2).

BoCOP34 (6.0 kb), investigated from *B. oleracea* accession EU579454.1, exhibited 5′-233/244-3′ bp LTRs. *BoCOP35* (4.8 kb), flanked by 392 bp LTRs, lacked any detectable PBS/PPT motifs with complete *gag-pol* polyproteins with additional domains (DUF, ZF) (Table 2). *BoCOP36* and *BoCOP37* were found in *B. oleracea* AC240081.1, sized 5.1 and 4.9 kb, and flanked by 366 and 170 bp LTRs, respectively. *B. oleracea* accession AC240082.1 harbored *BoCOP38* (7.1 kb) and *BoCOP39* (5.3 kb) with no PBS in *BoCOP38* (Table 2). *BoCOP41*, *BoCOP42*, and *BoCOP43* were around 4.6 kb; were flanked by 300, 264, and 268 bp LTRs, respectively; and were coding *gag-pol* proteins; however, *BoCOP43* lacked PBS motif. *B. oleracea* BAC (AC240088.1) harbored *BoCOP44* and *BoCOP45* (4.7–4.8) and was flanked by 320–400 bp LTRs, with additional UKP and ZK domains, respectively (Table 2). The elements *BoCOP46*, *BoCOP47*, *BoCOP48*, and *BoCOP50* were detected in *B. oleracea* BAC clone AC240090.1. *BoCOP48* and *BoCOP49* (6.1 kb) were copies of the same element integrated in opposite orientations in two BACs with the deleted *pol* region encoding RH domain only (Figure 1), indicating a sweep of other domains in the rearrangement of the element during evolutionary phases.

Table 2. List of *Brassica* retrotransposons with PBS and 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 envelope. 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. TLC: TLC domain. CL: copia-like. ND: not determined.

Element name	tRNA type	PBS (5′-3′)	PPT (5′-3′)	Domain structure (5′-3′)
BrCOP1	Met	TATCAGAGCCAGGTT	AGAGAAAGATGGAAG	GAG,INT,RT,RH
BrCOP2	Thr	GCTTTACGTTTGAGAG	ATGATTAAGGAGGAG	GAG,AIR1,ZK,INT,RT,RH
BrCOP3	Met	TATCAGAGCACAGTTGATCG	GAGAGACGAAGTAGA	GAG,ZK,INT,RT,RH
BrCOP4	Met	TATCAGAGCCAGGTT	AAGCTTGAGGGGGAG	GAG,INT,RT,RH
BrCOP5	Tyr	TCCGCTACCAAAAGTTCG	GGAGTATTAGGAAAG	GAG,INT,PRK RT,RH
BrCOP6	Met	GTATCAGAGCATTTCTTT	CATCTTGAGGGGGGG	GAG,INT,RT,RH
BrCOP7	Thr	AGACTGTTCTTGAATGAGTTG	AGAAGAGCAGAGAAG	GAG,INT,RT,RH
BrCOP8	ND	----	AGAGATGGAGGAGCG	GAG,INT,RT,RH
BrCOP9	Gln	AGGTCTTCACCGGTAAGGATT	GGTTGAGAGTATAGA	GAG,HVE,INT,RT,RH
BrCOP10	Trp	TAAATCCCTGAGACCTAAATC	GAATGTTATAAAGAA	GAG,INT,RT,RH
BrCOP11	Pro	TATAGTTGATAGAATCTTG	AGAGAGGTGAAGACA	GAG,ZK,INT,RT,RH
BrCOP12	Met	AACCTCTCTCCCGTGCCCA	CCTCCACCCCTTCTC	GAG,INT,RT,RH
BrCOP13	Thr	TGCCTCCAAGCTAAAACGAT	AAGACTGCGGGGGAG	GAG,INT,RT,PVM,RH
BrCOP14	Leu	GAGCATTCTATTGAATT	TAAGGGGGAGAATGT	GAG,INT,RT,RH
BrCOP15	Gln	AGCGTTCCAAACCGAGTCCTT	ATGGATCGAAAGGTG	GAG,INT,RT,RH
BrCOP16	Met	TATCAGAGCTCAGCAAGT	GAGTTTGCGAGGGGA	GAG,INT,RT,RH
BrCOP17	Met	TATCAGAGCACAAAATTC	CAACTTGAGGGGGAG	GAG,INT,RT,RH
BrCOP18	Met	TATCAGAGCCAGGTT	AGAGAGACGGAGAAG	GAG,ZK,INT,RT,RH
BrCOP19	Val	GGCTTCGTCATGGTGTCG	GGTCTAGGAGCAAAG	GAG,INT,ETS,RT,RH
BrCOP20	Arg	ATCTTGCCAATGAGTGCG	AGCGAGAAAAAGAAA	GAG,INT,RT,RH
BrCOP21	Met	TATCAGAGCCAGGTT	TATCAGAGCCAGGTT	GAG,INT,RT,RH
BoCOP22	Leu	GACAGCTACAGTGAGATGTT	TAAAAAGGGGGAGAT	GAG,AIR1,INT,RT,RH
BoCOP23	ND	----	ND	RT,RH
BoCOP24	Met	TATCAGAGCCTGAGTTACG	AAGACAGAAGACAGA	GAG,INT,RT,RH
BoCOP25	Trp	CATCTCTTTGAATTTG	GATATCAATAAGAAG	GAG,ZK,INT,RT,RH
BoCOP26	Met	TATCAGAGCTGAGGTT	AGGACAAGGAGGAGA	RT,RH
BoCOP27	ND	----	GGGAAGGGGGAGATT	GAG,ZK,INT,RT,RH
BoCOP28	Arg	CGGTCCCCAAGGAGAGT	CCTCTACTATTATTT	GAG,INT,RT,RH,
BoCOP29	Ser	CGTTATCAGCACGATCG	GCATCAAAGGGGGAG	GAG,INT,RT,RH
BoCOP30	ND	----	GAAGTAAAGGAAGAA	GAG,INT,RT,RH
BoCOP31	Lys	ATCACTCTGCGATTCG	GAGAGCGGATAGTGA	GAG,DUF,INT,RT,RH
BoCOP32	Met	TATCAGAGCCAGGTT	AAGCTTGAGGGGGAG	GAG,INT,RT,RH
BoCOP33	Met	TATCAGAGCAAAATCT	AAGGAGATGCGAGAG	GAG,INT,RT,RH
BoCOP34	Thr	CGTTATCAGCACGATT	ACATCCAAGGGGGAG	GAG,INT,RT,RH
BoCOP35	ND	----	ND	GAG,DUF,ZK,INT,RT, RH
BoCOP36	Met	TATCAGAGCTTCGGGTT	AGTCAAGGTGGGGAG	GAG,INT,RT,RH
BoCOP37	Met	TATCAGAGCAGAAAGATTC	CAACTTGAGGGGGAG	GAG,INT,RT,RH
BoCOP38	ND	----	AGGTGGAGAGCACAA	GAG,INT,RT,RH
BoCOP39	Ser	CGTTGTCAGCACGATTACG	GCATCCAAGGGGGAG	GAG,INT,RT

Table 2. (Continued).

Element name	tRNA type	PBS (5′-3′)	PPT (5′-3′)	Domain structure (5′-3′)
BoCOP40	Met	TATCAGAGCCAGGTT	GGGAAGGGGGAGATT	GAG,ZK,INT,RT,RH
BoCOP41	Met	TATCAGAGCCTGAGTT	AAGGAAATGAGAGAC	GAG,INT,RT,RH
BoCOP42	Met	TATCAGAGCGTTAGGTTACG	AGCTCAAGAGAGAGA	GAG,INT,RT,RH
BoCOP43	ND	----	GAAGTAAAGGAAGAA	GAG,INT,RT,RH
BoCOP44	ND	----	GGAAAGGGATAAGGG	GAG,INT,UKP,RT,RH
BoCOP45	Met	TATCAGAGCTACAAGTTCC	AAGTTTAAGAGGGGG	GAG,ZK,INT,RT,RH
BoCOP46	Met	TATCAGAGCTTCGGTTT	AGTCAAGGTGGAGAA	RT
BoCOP47	Met	TATCAGAGCTTCGGGTT	AAGTCAAGATGGAGA	GAG,ZK,RT
BoCOP48	Leu	TGTCATAACCATATAGGGTTT	AAGGGCCGGAAGAGA	RH
BoCOP49	Leu	TGTCATAACCATATAGGGTTT	AAGGGCCGGAAGAGA	RH
BoCOP50	Met	TATCAGAGCCATTCA	AAAGAGATGAGAGAC	GAG,INT,RT,RH
BoCOP51	Met	TATCAGAGCCATTCA	AAAGAGATGAGAGAC	GAG,INT,RT,RH
BoCOP52	Met	TATCAGAGCTCCAGGTTTCG	AATTAAGGGGGAGAA	GAG,INT,RT,RH
BoCOP53	Met	TGTCATAACCATACAGGGATT	AAACATAAAGAGTCA	GAG,INT,RT,RH
BoCOP54	Met	TATCAGAGCAACTAGGT	AAAGAAGATATGAAG	GAG,INT,RT,RH
BoCOP55	Pro	TATCATGTTATAATTG	AAGAGCGGATAGTGA	GAG,DUF,INT,RT,RH
BoGYP1	Met	TATCAGAGCGGGTTCCG	ATTAGTGGGGGAGAA	GAG,TLC,AP,RT,RH,INT
BoGYP2	Cys	AGGTCCCAATGCGTGGT	ND	GAG,AP
BoGYP3	Lys	CGCCCATCGTGGGGCT	GTGAACTGGAGGGGA	GAG,AP,RT,RH,INT
BrGYP4	Lys	CGCCCACCGTGGGGCT	GAACTGGGGGGGGAC	GAG,AP,RT,RH,INT
BrGYP5	Lys	CGCCCACCGTGGGACCG	GAACTGGGGGGGGAC	GAG,AP,RT,RH,INT
BoGYP6	Lys	CGCTCACCGTGGGATCA	ACTGGGGGGGGGGGG	GAG,RT,RH,INT
BrGYP7	Lys	CGCCCACCGTGGGGC	GATGGACTGGGGGGA	GAG,AP,RT,RH,INT
BrGYP8	Phe	TGCGGTGACTCGATCG	AAGCTTGAGGACAAG	GAG,AP,RH,INT,CHR
BrGYP9	Tyr	TTCGAACCTCGGAATC	GGGAGAAGAAGAAGC	GAG,AP,RT,RH,INT,CHR
BrGYP10	Tyr	TTCGAACCTCGGAATC	GGGAGAAGAAGAAGC	GAG,AP,RT,RH,INT,CHR
BrGYP11	Arg	CGATTCTACTCGTGATC	GTACGGGAGGGGACC	GAG,AP,RT,RH,INT,CHR
BrGYP12	Met	TATCAGAGACCTTTAAATTA	GTACGGGAGGGGACC	GAG,ZK,AP,RT,RH,ZF,INT
BoGYP13	Tyr	CGGATGAGCAGCGGCTGTG	AAGTAAAAGAATAAG	GAG,AP
BrGYP14	ND	----	AAAAGAAAATAAAAA	GAG,AP
BrGYP15	Ser	CGAATCCTTCTCACCCG	GCTTTGCTACGCTCC	GAG,AP,RT,RH,INT

BoCOP50 and *BoCOP51* showed homogeneous structures, while *BoCOP52* displayed a structure more similar to them. *BoCOP53*, *BoCOP54*, and *BoCOP55* were about 5.9, 4.7, and 6.3 kb, including LTRs of 200, 273, and 251 bp, respectively (Table 2).

3.2.3. PBS and PPT motifs of *Brassica* Copia elements

The PBS motif towards the downstream of 5′LTR and the PPT adjacent to 3′LTR were investigated in all Copia elements. The size of PBS in a few elements was slightly variable, while same-sized PPT were detected in most elements (Table 2). Around 85% of Copia showed both PBS and PPT motifs, 12% showed no PBS, and only 3% lacked a PPT motif. The PBS and PPT motifs from *BoCOP23* and *BoCOP35* were not detected, while *BrCOP8*, *BoCOP27*, *BoCOP30*, *BoCOP38*, *BoCOP43*, and *BoCOP44* failed to display the PPT motif when scanned against the *Arabidopsis thaliana* tRNA database. Eleven different tRNA types were detected by PBS; the most common type was $tRNA_{Met}$, which was present in 45% of the elements. The second important primer type was $tRNA_{Thr}$, identified in 9% of the elements (Table 2).

3.2.4. PCR detection of Copia RE distribution in *Brassica*

The diversity and distribution of various Copia elements among 40 *Brassica* accessions/cultivars (Table 3) were studied

Table 3. List of *Brassica* species with their accessions names. ND: not determined.

No.	Species	Accession name	No.	Species	Accession name
1	*B. rapa chinensis*	Pak Choy	21	*B. juncea*	Tsai Sim
2	*B. rapa pekinensis*	Chinese Wong Bok	22	*B. juncea*	W3
3	*B. rapa chinensis*	San Yue Man	23	*B. juncea*	Giant Red Mustard
4	*B. rapa rapa*	Hinona	24	*B. juncea*	Varuna
5	*B. rapa rapa*	Vertus	25	*B. napus*	New
6	*B. rapa*	Suttons	26	*B. napus oleifera*	Mar
7	*B. nigra*	ND	27	*B. napus biennis*	Last and Best
8	*B. nigra*	ND	28	*B. napus napo*	Fortune
9	*B. nigra*	ND	29	*B. napus*	Drakker
10	*B. juncea*	NARC-I	30	*B. napus*	Tapidor
11	*B. juncea*	NATCO	31	*B. carinata*	Addis Aceb
12	*B. juncea*	NARC-II	32	*B. carinata*	Patu
13	*B. oleracea*	De Rosny	33	*B. carinata*	Tamu Tex-sel Greens
14	*B. oleracea*	Kai Lan	34	*B. carinata*	Mbeya Green
15	*B. oleracea*	Early Snowball	35	*B. carinata*	Aworks-67
16	*B. oleracea italica*	Precoce Di Calabria Tipo Esportazione	36	*B. carinata*	NARC-PK
17	*B. oleracea capitata*	Cuor Di Bue Grosso	37	*B. napus* × *B. nigra*	ND
18	*B. oleracea*	ND	38	*B. carinata* × *B. rapa*	ND
19	*B. juncea*	Kai Choy	39	*B. napus* × *B. nigra*	ND
20	*B. juncea*	Megarrhiza	40	*B. napus* × *B. nigra*	ND

by PCR analysis using five sets (Table 4) of newly developed reverse transcriptase amplification polymorphism (RTAP) markers. Primer set BrCOP2F/R (Table 4) was designed to amplify a 710 bp RT fragment of *BrCOP2* family. The results showed amplification of RT fragments from 37 cultivars from six *Brassica* species. The products were amplified in *B. rapa* (Pak Choy, Chinese Wong Bok, San Yue Man, Hinona, Vertus, Suttons), *B. oleracea* (De Rosny, Kai Lan, Early Snowball, Cuor Di Bue Grosso, Precoce Di Calabria, GK97361), *B. juncea* (NARC-I, NATCO, NARC-II, Kai Choy, Megarrhiza, Tsai Sim, W3, Giant Red Mustard, Varuna), *B. napus* (New, Mar, Fortune, Drakker, Tapidor), *B. carinata* (Addis Aceb, Patu, Tamu Tex-Sel Greens, Mbeya Green, Aworks-67, NARC-PK), and 4 synthetic hexaploid *Brassica* (Figure 2a). There was no amplification in *B. nigra*, except accession HRIGRU010919, which showed a separate evolutionary history of *B. nigra*.

The amplification of *BrCOP11* revealed the A-genome–specificity of the element. The primer pair BrCOP11F/R (Table 4) amplified 650 bp RT products from 26 of 40 *Brassica* A-genome diploid and polyploidy accessions (AA, AABB, AACC, AABBCC). Each of the 6 *B. rapa*, 6 *B. napus,* 8 *B. juncea*, and 4 hexaploid *Brassica* amplified bands, while only 1 *B. nigra* and 1 *B. carinata* amplified the expected bands and showed polymorphisms for this insertion. There was no amplification from *B. oleracea* and *B. carinata* except one, suggesting its absence in C-genome (Figure 2b). The amplification of 703 bp RT region of *BoCOP25* revealed its C-genome–specific nature. Twenty-four of 40 *Brassica* lines amplified the product (Figure 2c) including all *B. oleracea*, *B. napus*, *B. carinata*, and hexaploid *Brassica* cultivars. The bands amplified from *B. rapa* (Suttons) and *B. juncea* (Giant Red Mustard) cultivars were not as strong as those amplified from other cultivars (Figure 2c). The lack of amplification in *B. nigra* suggests the proliferation of the element after the separation of B- from A-/C-genomes. The PCR amplification of *BoCOP37* with primer pair BoCOP37F/R (Table 4) revealed their abundance and distribution among *Brassica* cultivars (Figure 2d), and amplified 34 products from 40 *Brassica* genome collections. The amplification of *BoCOP44* by primer BoCOP44F/R (Table 4) showed its abundance; it was distributed in all *Brassica* except 1 *B. rapa* and 1 *B. nigra* cultivar (Figure 2e). This suggests the ancient nature of elements that were present in a common ancestor before the separation of B- and A-/C-genomes.

Table 4. List of primers to amplify the RT regions of *Brassica* Copia and Gypsy 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 BrCOP2R	GACGTGGGAACTAGTGGAC CACTCTTGCTGTCTCGCATC
2	Copia	*BrCOP11*	650	BrCOP11F BrCOP11R	CAGCTTTGCAATCTGTCATG GGGAATTCCAGGAGTTGAAG
3	Copia	*BoCOP25*	703	BoCOP25F BoCOP25R	CATTGCACGATCCCATTCCG TGGGATCTCGTTGAACTACC
4	Copia	*BoCOP37*	722	BoCOP37F BoCOP37R	TGAGCTCCACTGGTACATAG GGAGGTTGCTACTCTTCCTC
5	Copia	*BoCOP44*	715	BoCOP44F BoCOP44R	AGGCAGAGGAGTAGGCATTG GGTGCCACCAACTGAAGATA
6	Gypsy	*BoGYP1*	521	BoGYP1F BoGYP1R	AATCACATGGCCCAAAAATC GGCCGAGTACTTCACTGTGG
7	Gypsy	*BrGYP5*	562	BrGYP5F BrGYP5R	AGGTTACTCGGTGCAGGTTC TTCCTCGCTGTGTGACAATG
8	Gypsy	*BrGYP9*	598	BrGYP9F BrGYP9R	AACCGCTTTAACCTTGTTAG GGTTCAAAGTCTGTTGGATG
9	Gypsy	*BrGYP12*	770	BrGYP12F BrGYP12R	CCCCCTTCGAGATATACAGC AGAAAGAGGCAAGTCCGTGA
10	Gypsy	*BrGYP15*	421	BrGYP15F BrGYP15R	CGAGCAATCAACAAGATAAC GTACTTCTGAAGCGCCGAAC

3.3. Evolutionary relationship of *Brassica* retrotransposons

The phylogenetic relationships of 64 RT sequences were performed by aligning the most conserved region around the D-DD triad (~180 aa). Of the 64 RT sequences, 62 belonged to *Brassica* Copia/Gypsy elements, while 2 elements (Ty1B/copia and Ty3/gypsy) were collected from *Saccharomyces cerevisiae*. The elements clustered into two main lineages with 11 and 53 elements, clearly splitting the Gypsy and Copia superfamilies, respectively, with high bootstrap supports (Figure 3). In Gypsy lineage the elements further clustered in two clades, out-grouping the *Ty3* element of *S. cerevisiae*. The elements *Br/BoGYP3-Br/BoGYP7* clustered in one clade, while *BoGYP1* and *BrGYP9-BrGYP12* constituted sister families in another clade. The Copia lineage further clustered into 2 clades and 10 sub-clades (indicated by different shapes) with 1–9 elements in sub-clades. The element *Ty1B* out-grouped the Copia clade indicating distinct *Brassica* Copia sequences in comparison to other plant Copia. *BoCOP45* family clustered in its own sub-clade, thus revealing the most varied and heterogeneous sequences investigated in *Brassica*. *BrCOP2* and *BoCOP22* shared a family with weak bootstrap values. Several other families shared the sub-clades due to homologies in their sequences (Figure 3).

The evolutionary relationships of 130 *Brassica* Copia RT sequences clustered them into two major clades with 33 and 97 sequences in each clade and then dissolved them in 10 sub-clades and 18 families (Figure 4). *Arabidopsis thaliana* Copia '*Araco*' was used to root the tree. The first clade clustered into 3 sub-clades (indicated by different shapes/colors), where *BoCOP45* family out-grouped making it the closest group with *Arabidopsis* '*Araco*' element. The sequences from this family were the most varied and heterogeneous sequences in relation to all other Copia RT studied. The second sub-clade was represented by members of *BrCOP23/BoCOP23*, *BoCOP41*, *BoCOP42, BrCOP50*, *BoCOP51*, and *BoCOP52*. The third sub-clade comprised 15 elements from *BrCOP14*, *BrCOP21, Br/BoCOP25*, *BoCOP30, BoCOP33*, and *BoCOP43* with mostly homogeneous sequences. The second clade further resolved into 7 sub-clades with 6–20 elements in respective sub-clades (Figure 4). The *BrCOP2* and *BoCOP22*, constituting one sub-clade, out-grouped in the second clade. *BrCOP1*, *BrCOP11*, *BrCOP18, BoCOP26, BoCO36*, *Br/BoCOP46*, and *Br/BoCOP47* clustered in the

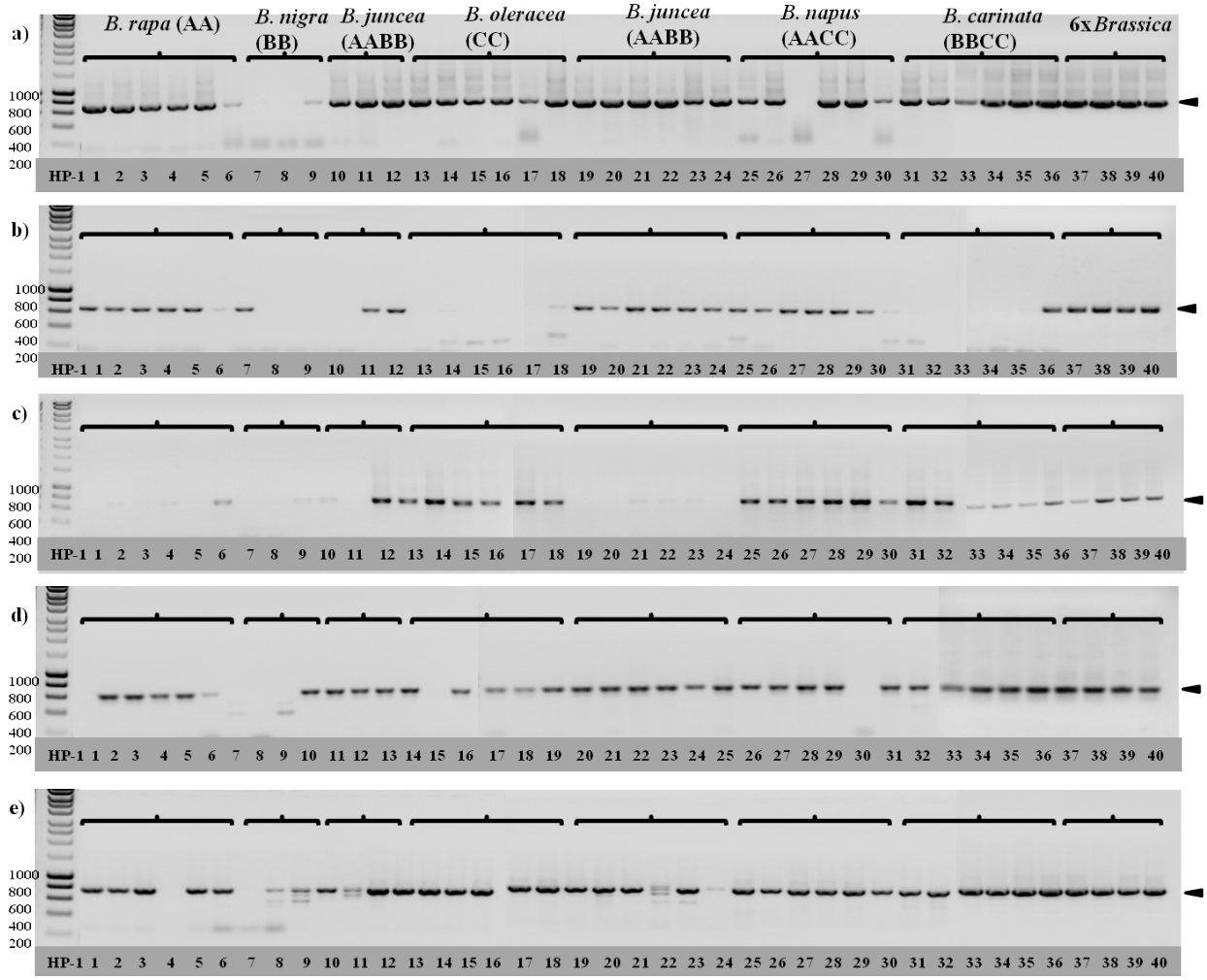


Figure 2. PCR amplification for the detection of Copia RT polymorphisms across 40 cultivars in *Brassica*. Dark arrow heads indicate the expected product sizes. a) *BrCOP2*, b) *BrCOP11*, c) *BoCOP25*, d) *BoCOP37*, e) *BoCOP44* [PCR figures show reversed images of size-separated ethidium bromide-stained DNA on agarose gels after electrophoresis; ladders (HP-I) show fragment sizes in base pairs; lower numbers indicate accessions of the species indicated in Table 3]. Br: *Brassica rapa*. Bo: *Brassica oleracea*. COP: Copia.

same sub-clade, representing their respective families. Due to high homologies in the RT regions of a few families, they clustered in family-specific groups, while others were distributed across their respective sub-clades. *BrCOP7-BrCOP9*, *BrCOP13, BrCOP16*, and *BoCOP44* clustered together in the same group. *BrCOP4*, *BrCOP5*, *BrCOP6*, *BrCOP17*, and *BoCOP32* shared the same sub-clade (Figure 4).

3.4 Overview of Gypsy retroelements

Fifteen full-length *Brassica* Gypsy elements were detected by dot plot analyses with sizes 2-fold larger than Copia (11.9 kb; *BoGYP3*), while the *pol*-region–deleted *BoGYP13* was only a 4.1 kb element. The Gypsy elements were flanked by 199–2035 bp LTRs and terminated with GC-rich perfect 5 bp TSDs. Two major Gypsy groups were distinguished on the basis of their sizes; one representing the small-sized (5.0–5.9 kb) and the other large-sized (11.2–11.9 kb) elements (Table 1). Most elements generated perfect and equally sized LTRs, but in a few (*BoGYP1* and *BoGYP2*), variable-sized LTRs were detected due to the uneven activity of small repeat sequences in one LTR. With the exception of *BoGYP2*, *BoGYP13*, and *BrGYP14* (Figure 5) the rest were complete autonomous elements, showing the *gag-pol* protein domains (Table 2).

3.4.1. Characterization and structural features of *Brassica* Gypsy REs

The element *BoGYP1* was about 9.1 kb in size and flanked by the largest LTRs (5′-2035/2004-3′ bp) investigated in the present study (Figure 5). *BoGYP1* displayed typical Gypsy-like *gag-pol* polyprotein structures with an additional TLC domain. A defective element (*pol* region deleted) *BoGYP2* was about 11.3 kb and flanked by 5′-

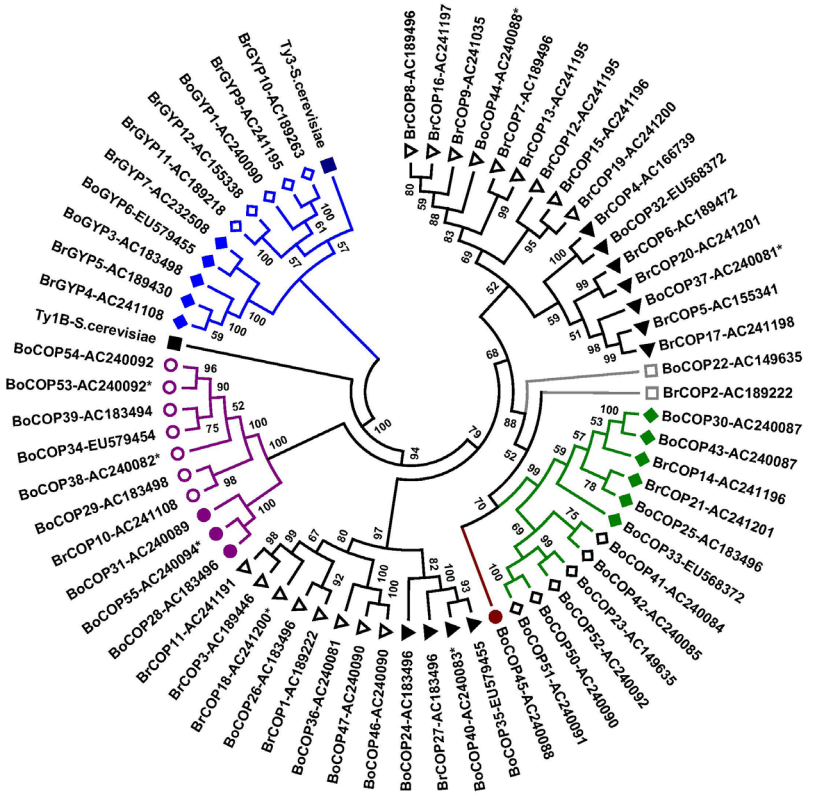


Figure 3. Phylogenetic analysis of 64 *Brassica* retrotransposons RT sequences. The neighbor-joining tree is based on 1000 bootstrap replicates (% shown at nodes), and a Poisson model was used to calculate genetic distance in Mega5. *Ty1B/copia* and *Ty3/gypsy* elements from *S. cerevisiae* were added to observe the evolutionary relation with *Brassica* Copia and Gypsy elements. Two major lineages split the elements into 12 clades (2 Gypsy, 10 Copia) represented by different filled and empty shapes (circles, squares, forward/reverse triangles, rhombuses). The names of the elements are followed by the *Brassica* accession numbers, to which they were identified. Br: *Brassica rapa*. Bo: *Brassica oleracea*. Bn: *Brassica napus.* COP: Copia. GYP: Gypsy.

1272/1140-3′ bp LTRs. The elements *BoGYP3*, *BrGYP4*, and *BrGYP5* were 11.8, 11.7, and 11.8 kb with 471–480 bp flanking LTRs. Their internal regions displayed typical *gag-pol* organization of non-chromodomain Gypsy (Table 2). *BoGYP6* and *BrGYP7* were about 11.5- and 11.6 kb-long elements flanked by 509 and 506 bp LTRs. Typical *gag-pol* organization of non-chromodomain Gypsy was observed in these elements (Figure 5).

The structural features of chromodomain (CHR)-bearing Gypsy showed relative homogeneity. *BrGYP1*, *BrGYP8*, *BrGYP9*, *BrGYP10*, *BrGYP11*, and *BrGYP12* belonged to the chromoviral branch of Gypsy superfamily based on their structures. *BrGYP8* was detected as a 5.1 kb element flanked by 331 bp LTRs and an internal domain displaying PBS complementary to $tRNA_{Phe}$, an unusual tRNA type identified in plants. The sizes of *BrGYP9* (Figure 5) and *BrGYP10* were 5.9 and 5.2 kb, and they were flanked by 346 bp LTRs. *BrGYP11* and *BrGYP12* showed homologies in their structures, were 5.1 kb, and exhibited LTRs of 340–360 bp. Two incomplete Gypsy *BoGYP13* and *BrGYP14* were identified that were about 4.1 and 7.2 kb elements including 200 and 1553 bp LTRs, respectively. The internal region of *BoGYP13* represented PBS/PPT motifs, but no recognizable PBS was detected in *BrGYP14* (Table 2). Although they displayed typical Gypsy-like ORFs for the *gag-pol* genes, their *pol* polyproteins lost the RT, RH, and INT domains in rearrangements during the ancient evolutionary period (Table 2).

3.4.2. PBS and PPT motifs of Gypsy elements

The PBS and PPT primers necessary for RNA amplification were detected in 93% of elements except *BoGYP14* (Table 2). Six different tRNA types were observed with $tRNA_{Lys}$ occurring most frequently (detected in 35% of elements); this was followed by $tRNA_{Tyr}$ which was observed in 20% of the elements. The most common tRNA type, $tRNA_{Met}$, was detected only in 15% of the elements. The other 3 types of tRNA contributed only 7% each to the tRNA type. A 15 bp PPT motif adjacent to the 3′LTR was detected in 93% of all Gypsy elements (except *BoGYP14*). All PBS and PPT sequences along with their positions within Gypsy sequences were identified and listed (Table 2).

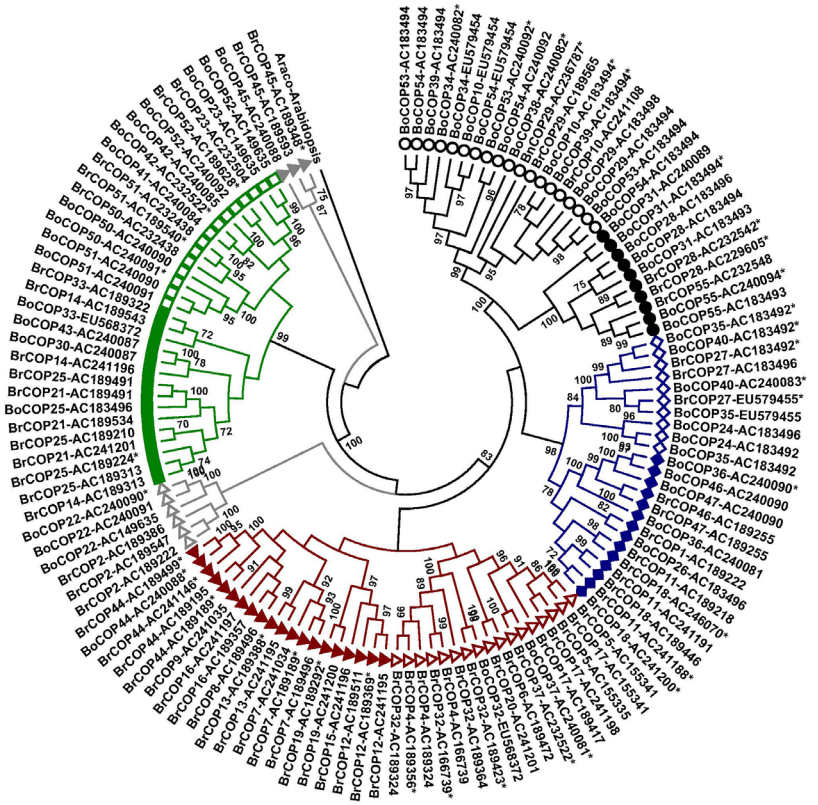


Figure 4. Phylogenetic analysis of 138 *Brassica* Copia-RT sequences. The neighbor-joining tree is based on 1000 bootstrap replicates (% shown at nodes), and a Poisson model was used to calculate genetic distance in Mega5 program. *Arabidopsis thaliana* Copia *Araco* was used to root the tree. Two major lineages resolved the elements into 10 clades (indicated by different filled and empty circles, squares, forward/reverse triangles, rhombuses), which further resolved into 18 families. The names of the elements are followed by *Brassica* accession numbers. Br: *Brassica rapa*. Bo: *Brassica oleracea*. Bn: *Brassica napus.* COP: Copia.

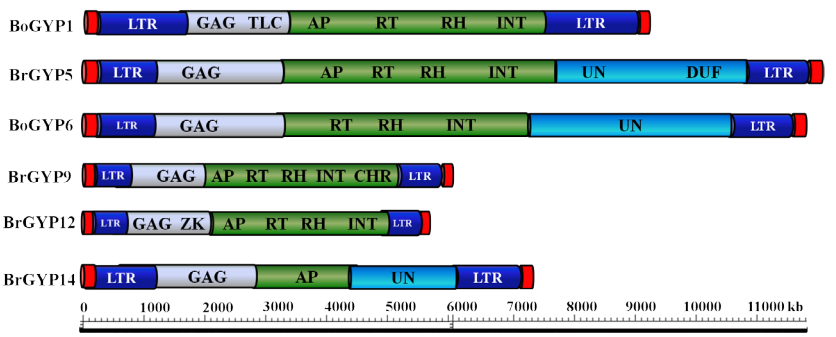


Figure 5. Schematic representation of structures of Gypsy elements in *Brassica*. The red discs at the terminals represent the TSDs, internal to TSDs indicates LTRs. The *gag* and *pol* regions are drawn with their protein domains. The scale below measures lengths of the elements (bp). Additional insertions or unknown sequences are highlighted in 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.

3.4.3. Distribution and abundance of *Brassica* Gypsy elements

The distribution and abundance of Gypsy retroelements in *Brassica* genomes (Table 3) were investigated by RT-based markers using 5 primer pairs (Table 4). The Gypsy elements showed high diversity and distribution across various *Brassica* genomes. The primer pair BoGYP1F/R amplified 521 bp RT regions from all 40 *Brassica* cultivars including *B. rapa*, *B. nigra*, *B. oleracea*, *B. juncea*, *B. napus*, *B. carinata*, and four synthetic hexaploid *Brassica* (Figure 6a). The insertion polymorphism of *BrGYP5* also showed the same pattern, where it was amplified from all 40 *Brassica* cultivars (Figure 6b).

The amplification polymorphisms of chromodomain-containing Gypsy were also investigated, and using primer pair BrGYP9F/R (Table 4) a 598 bp product was amplified from 36 of 40 *Brassica* cultivars tested. All *B. rapa*, *B. juncea*, *B. napus*, *B. carinata*, and hexaploid *Brassica* cultivars amplified the expected product. The B-genome *B. nigra* also amplified the product, with the exception of accession HRIGRU010978; whereas three of six *B. oleracea* (De Rosny, Precoce Di Calabria, Cuor Di Bue Grosso) accessions amplified the *BrGYP9* RT regions (Figure 6c). The polymorphisms of *BrGYP12* revealed its distribution among all six diploids and polyploid *Brassica* species from 'Triangle of U' and their cultivars used in the present study (Figure 6d). Similarly, *BrGYP15* yielded the 421 bp RT domains from all *Brassica* except *B. nigra* (HRIGRU011011) genomes (Figure 6e). The amplification of almost all Gypsy RT products from A-, B-, and C-*Brassicas*; allotetraploids; and hexaploids revealed the abundance and diversity of these elements.

3.5. Phylogenetic analysis of *Brassica* Gypsy RT sequences

The phylogenetic analysis of 40 *Brassica* Gypsy RT sequences clustered them into two major clades, chromodomain Gypsy (rhombus shapes in Figure 7) and

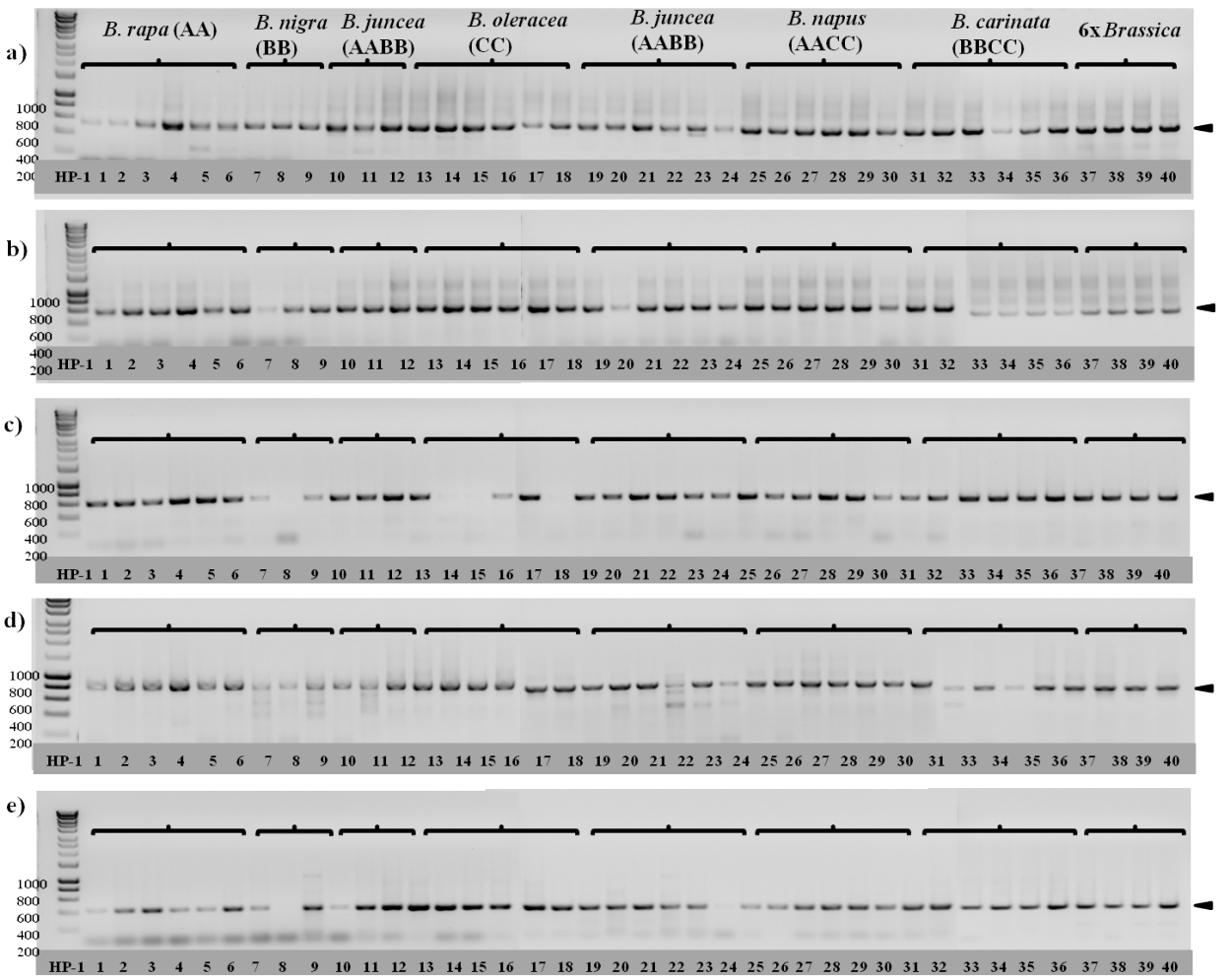


Figure 6. PCR analysis showing fragments with and without Gypsy RT regions between the primers. DNA samples were obtained with primers hybridizing to conserved RT regions of various Gypsy families. Dark arrow heads (right) indicate expected product sizes. Numbers underneath indicate accessions (Table 3). The amplification of a) *BoGYP1*, b) *BoGYP5*, c) *BrGYP9*, d) *BrGYP12*, e) *BrGYP15*.

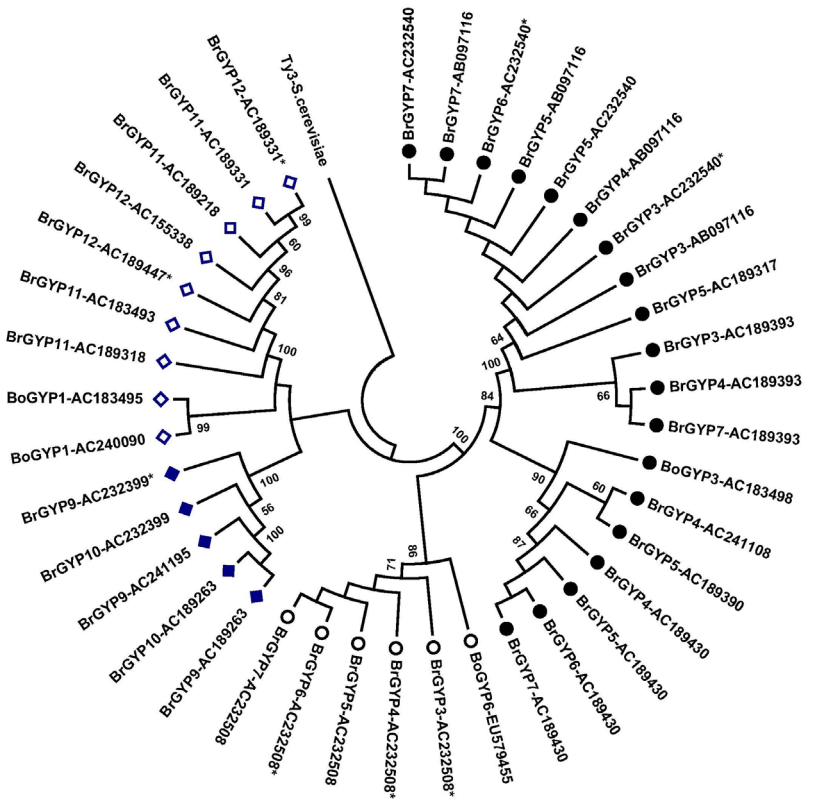


Figure 7. Phylogenetic analysis of 40 *Brassica* Gypsy elements based on the amino acid alignment of the conserved RT domains (~180 aa). Rooted (*A. thaliana* Gypsy *Tat4*) neighbor-joining method with Poisson model was used to construct the tree in Mega5. Tree was generated with 1000 bootstrap values with >50% values shown in the tree. Two main lineages separate the chromodomain-containing from non-chromodomain group; they are indicated by filled/empty rhombuses and circles, respectively. The names of the elements are followed by *Brassica* accession numbers, to which they were identified. Br: *Brassica rapa*. Bo: *Brassica oleracea*. GYP: Gypsy.

nonchromodomain Gypsy (represented by circles). Two sub-clades from chromodomain and 2 (filled/unfilled shapes) from nonchromodomain Gypsy were distinct, with elements possessing homogenous sequences in sub-clades. The members from *BoGYP1*, *BrGYP9*, *BrGYP10*, *BrGYP11*, and *BrGYP12* share one clade representing the chromoviruses-like elements, while *BrGYP3*, *BrGYP4*, *BrGYP5*, *BrGYP6*, and *BrGYP7* clustered in another group making a larger clade of nonchromodomain Gypsy. In the major clade, *BoGYP1, BrGYP9*, and *BrGYP10* clustered in one, while *BrGYP11* and *BrGYP12* made up the other sub-clade supported by high bootstraps values. The second major clade clustered into 2 sub-clades with weak bootstrap values, where *BrGYP3*, *BrGYP4*, *BrGYP5*, *BrGYP6*, and *BrGYP7* elements were dispersed together, suggesting their common ancestry (Figure 7).

4. Discussion

Comparative sequence analyses have shown very fast variations in plant genomes; repetitive DNA sequences, or REs, are major sources of such rapid changes in many genomes (Bennetzen, 2000). As genome sequencing progresses and is updated, there is a need to discover and characterize the TEs, especially LTR REs, which are major drivers of gene and genome evolution. To our knowledge, the present study is the first detailed survey of Copia and Gypsy elements in *Brassica* genomes by the novel approach of comparative analysis of BAC sequences in dot plot for RE identification. This strategy helped to identify most of the elements present in *Brassica* BAC sequences, which are not detectable with other bioinformatics programs and tools.

The LTR REs are highly abundant in plants and were investigated in A- and C-genome *Brassica*. We estimated high copy numbers in *B. oleracea* (7540 Copia, 780 Gypsy) in comparison to *B. rapa* (1596 Copia, 540 Gypsy). In a recent study, LTR Finder was used to screen the 2020 *Brassica* BACs, and around 9956 retroelements were identified with greater proportions in *B. oleracea*. Six BACs showed nested structures and a high proportion of REs including 20%–50% of BAC sequences (Wei et al., 2013). These six BACs were not analyzed in the present study, but a few BACs such as AC240090.1 (Supplementary Figure), AC183496.1, and AC240090.1 showed 15%–40% of RE proportions. The PCR amplification showed activity in

both genomes with higher levels in *B. oleracea* suggesting higher RE proliferation in C-genome. In a study using universal PCR primers, 80 RT fragments were isolated from 16 *Brassica* lines of the 3 diploid and 3 polyploid *Brassica* species. The study confirmed the availability of LTR REs in *Brassica* (Alix and Heslop-Harrison, 2004); however, the present study is more informative and descriptive, as all the structural features and abundance of the elements were investigated in *Brassica*, and the element RT fragments were PCR amplified. The present results further strengthen the hypothesis of Fujimoto et al. (2008), which suggests more REs in *B. oleracea* than *B. rapa*. This is confirmed by a comparison of TEs in *Arabidopsis thaliana* and *B. oleracea*, which shows a high percentage and transduplication of TEs in *B. oleracea* and, hence, a larger size compared to *A. thaliana* (Zhang and Wessler, 2004). The activation and transposition of REs is activated during stress and hypomethylated conditions (Hirochika et al., 2000). Considering these conditions, we suggest that differences in RE copy numbers between A- and C-genome *Brassicas* may be due to variable environmental stress. The reduction in genome size of *B. rapa* may be due to deletion of REs that were swept from the genome by stress conditions or DNA replication and cross over.

The organization of *gag-pol* protein domains in LTR REs is highly conserved and can be used to classify the REs into their respective superfamilies as Copia (5′-GAG-AP-INT-RT-RH-3′), Gypsy (5′-GAG-AP-RT-RH-INT-3′), and retroviruses (5′-GAG-AP-RT-RH-INT-ENV-3′). The arrangement of these domains varies considerably, and some additional domains or ORFs were also identified in a few retroelements (Wicket at al., 2007; Novikov et al., 2012), as observed in the present study. Several elements from both superfamilies harbored additional domains in their structures such as ZK domain in *BrCOP2*, *BrCOP3*, *BrCOP11*, *BoCOP25*, *BoCOP27*, *BoCOP40*, and *BrGYP12*; AIR1 domain in *BrCOP2* and *BoCOP22*; HVE in *BrCOP9*; DUF in *BoCOP35* and *BoCOP55*; and TLC domain in *BoGYP1* (Table 2). Several elements from both Copia and Gypsy superfamilies lacked 1 or more *pol* protein domain revealing defective or deleted derivatives of autonomous elements as domain-lacking elements, as described by Novikov et al. (2012).

The LTR REs were investigated in many eukaryotic genomes with newly developed markers such as SSR, SSAP, IRAP, REMAP, and RBIP. The amplification of RE insertions in host genomes provides strong markers for studying genome evolution and diversity (Flavell et al., 1998; Schulman et al., 2004, 2012). In recent years transposon- based markers have remained highly informative for genetic diversity and genotype/varietal identifications, including RADP markers in *Gossypium hirsutum* (Surgun et al., 2012); RAPD and ISSR markers in sugar beet (Izzatullayeva et al., 2014); and SSR markers in *Rhodiola rosea*, *Salvia*, and *Thymus* (Gyorgy et al., 2013; İnce and Karaca, 2015; Karaca et al., 2015). RTAP markers were developed to conduct PCR analyses to reveal distribution of retroelements among various *Brassica* species. The majority of elements from Copia and Gypsy superfamilies were amplified from all *Brassica* species including *B. nigra*, while a few elements were found proliferating in A- (*BrCOP11*) or C-genome (*BoCOP25*) alleles. The Gypsy sequences were more abundant, distributed across almost all *Brassica* species, and PCR-amplified from most of the tested *Brassica* accessions, which reveals their ancient nature and activity before the separation of A-, B-, and C-genome *Brassicas*. Abundance, diversity, and activity of REs were studied in several other plants (Defraia and Slotkin, 2014) including wheat, barley, rice, and *Arabidopsis* (Wicker and Keller, 2007; Tsukahara et al., 2009; Tomita et al., 2010) and sunflower (Kawakami et al., 2010). Activity, diversity, and abundance of Gypsy REs were also investigated in *Brassica* (Alix and Heslop-Harisson, 2004), wheat (Tomita et al., 2010), soybean (Du et al., 2010), pepper and tomato (Park et al., 2011), and *Arabidopsis* (Tsukahara et al., 2009), suggesting their role in plant genome size duplication and diversification. A small number of inconsistent results were found in the present RTAP RE insertion assays, where one accession or another did not include an element amplified from related accessions. This could result from mutation in the primer sites or excision of this genomic region in some accessions.

The RT sequences of REs can be used to deduce the phylogeny between various elements. The evolutionary relationship of 64 elements from *Brassica* with known elements segregated them into Gypsy and Copia lineages, further clustering them into respective clades and sub-clades revealing separate lines of evolution in both superfamilies (Figure 3). The detailed analysis of 130 Copia RT sequences segregated them into 10 sub-clades resolving them into 18 families. A few sub-clades were family specific, such as *BoCOP45* sequences in one sub-clade and *BrCOP2* and *BoCOP22* in another sub-clade (Figure 4). The phylogeny of 40 *Brassica* Gypsy sequences clearly clustered them into chromodomain (CHD) and nonchromodomain clades (Figure 7). Previous studies confirmed that CHD-containing Gypsy are more advanced and form a separate group from non-CHD Gypsy (Novikov et al., 2012). The distinct nature and placement of Copia and Gypsy in separate clades or groups is evident from several studies, such as one conducted in the yeast genome (Neuveglise et al., 2002).

The retrotransposition of elements required primer-related sites such as PBS downstream to 5′ LTR and PPT towards upstream of 3′ LTRs. The PBS and PPT

sequences were identified from more than 80% of the elements investigated in the present study (Table 2). In *Brassica* elements, the most commonly used tRNA was complementary to $tRNA_{Met}$, with a few other types that were detected in several other plants including PBS of *Ty1*, *Ty2*, *Ty3*, and *Ty5* retroelements (Voytas and Boeke, 1993). *S. cerevisiae Ty4* PBS is complementary to $tRNA_{Asn}$ (Stucka et al., 1992), and the PBS of *Tca1* and *Tca2* are complementary to $tRNA_{Arg}$ (Goodwin and Poulter, 2000), which were also detected in the PBS of a few *Brassica* tRNA types in present study.

The present study is an extensive and detailed compilation of the LTR RE landscape of the *Brassica* genomic BAC sequences and their distribution patterns among various *Brassica* species. The results enabled identification and understanding of the structure and nature of full-length elements and their derivatives. The BAC-based approach not only relies on the conserved protein domains most often analyzed but also ensures that all the families studied have shown activity during their recent evolutionary history within the genus *Brassica*. 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 'wild' *Brassica* species outside the 'Triangle of U' and other genera to explore the value of RBIP-type and RTAP markers for identifying alien chromosome and alien genome introgression.

Acknowledgments

This work was supported by Hazara University Mansehra, Pakistan, and the Higher Education Commission of Pakistan. We are thankful to them for the funding and support of this work. We thank Dr Graham Teakle and Dr Guy Barker from Warwick University, UK, and Dr Xianhong Ge from University of Wuhan, China, for supplying seeds or DNA from the *Brassica* accessions.

References

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

Bennetzen JL (2000). Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251–269.

Capy P (2005). Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 110: 457–461.

Defraia C, Slotkin RK (2014). Analysis of retrotransposon activity in plants. Methods Mol Biol 1112: 195–210.

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63: 584–598.

Eickbush TH, Jamburuthugoda VK (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134: 221–234.

Flavell AJ, Knox MR, Pearce SR, Ellis TH (1998). RE-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J 16: 643–650.

Flavell AJ, Pearce SR, Heslop-Harrison P, Kumar A (1997). The evolution of Ty1-copia group Retrotransposons in eukaryote genomes. Genetica 100: 185–195.

Flavell AJ, Smith DB, Kumar A (1992b). Extreme heterogeneity of Ty1-copia group retrotransposons in plants. Mol Gen Genet 231: 233–242.

Fujimoto R, Sasaki T, Inoue H, Nishio T (2008). Hypomethylation and transcriptional reactivation of RE-like sequences in ddm1 transgenic plants of *Brassica rapa.* Plant Mol Biol 66: 463–473.

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.

Goodwin TJ, Poulter RT (2000). Multiple LTR-RE families in the asexual yeast *Candida albicans*. Genome Res 10: 174–191.

Gyorgy Z, Fjelldal E, Szabo A, Aspholm PE, Pedryc A (2013). Genetic diversity of golden root (*Rhodiola rosea* L.) in northern Norway based on recently developed SSR markers. Turk J Biol 37: 655–660.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. Nucl Acids Symp Ser 4: 95–98.

Hansen CN, Heslop-Harrison JS (2004). Sequences and phylogenies of plant pararetroviruses, viruses and transposable elements. Adv Bot Res 41: 165–193.

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

Hirochika H, Okamoto H, Kakutani T (2000). Silencing of retrotransposons in *Arabidopsis* and reactivation by the ddm1 mutation. Plant Cell 12: 357–369.

İnce AG, Karaca M (2015). E-microsatellite markers for some naturally occurring *Salvia* species in the Mediterranean region. Turk J Biol 39: 69–77.

Izzatullayeva V, Akparov Z, Babayeva S, Ojaghi J, Abbasov M (2014). Efficiency of using RADP and ISSR markers in evaluation of genetic diversity in sugar beet. Turk J Biol 38: 429–438.

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007). Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8: 241–259.

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: 462–467.

Kapitonov VV, Jurka J (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9: 411–412; author reply 414.

Karaca M, İnce AG, Aydın A, SY Elmasulu, Turgut K (2015). Microsatellites for genetic and taxonomic research on thyme (*Thymus* L.). Turk J Biol 39: 147–159.

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

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: 821–831.

Kumar A, Bennetzen JL (1999). Plant retrotransposons. Annu Rev Genet 33: 479–532.

Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, Anguilar-Rodriguez J, Vicente-Ripolles M, Fuster G, Bernet GP et al. (2011). The Gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39: 70–74.

Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-RE structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860–869.

Mills RE, Bennett EA, Iskow RC, Lutting CT, Tsui C, Pittard WS, Devine SE (2006). Recently mobilized transposons in the human and chimpanzee genomes. Am J Hum Genet 78: 671–679.

Minervini CF, Viggiano L, Caizzi R, Marsano RM (2009). Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. Gene 440: 42–49.

Monteiro A, Lunn T (1999). Trends and perspectives of perspectives of vegetable *Brassica* breeding world-wide. WCHR—World Conference on Horticulture Research ISHS. Acta Hort, p. 495.

Neuveglise C, Feldmann H, Bon E, Gaillardin C, Casaregola S (2002). Genomic evolution of long terminal repeat retrotransposons in Hemiascomycetous yeasts. Genome Res 12: 930–943.

Novikov A, Smyshlyaev G, Novikova O (2012). Evolutionary history of LTR retrotransposon chromodomains in plants. Int J Plant Genomics, 874743.

Novikova O (2009). Chromodomains and LTR retrotransposons in plants. Commun Integr Biol 2: 158–162.

Novikova O, Mayorov V, Smyshlyaev G, Fursov M, Adkison L, Pisarenko O, Blinov A (2008). Novel clades of chromodomain containing Gypsy LTR retrotransposons from mosses (Bryophyta). Plant J 56: 562–574.

Ostergaard L, King GJ (2008). Standardized gene nomenclature for the *Brassica* genus. Plant Methods 4: 10.

Park M, Jo S, Kwon JK, Park J, Ahn JH, Kim S, Lee YH, Yang TJ, Hur CG, Kang BC 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 retrotransposons in rye (*Secale cereale*). Genome 40: 617–625.

Schulman AH, Flavell AJ, Ellis TH (2004). The application of LTR retrotransposons as molecular markers in plants. Methods Mol Biol 260: 145–173.

Schulman AH, Flavell AJ, Paux E, Ellis TH (2012). The application of LTR retrotransposons molecular markers in plants. Methods Mol Biol 859: 115–153.

Sonnhammer EL, Durbin R (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA/protein sequence analysis. Gene 167: 1–10.

Stucka R, Schwarzlose C, Lochmuller H, Hacker U, Feldmann H (1992). Molecular analysis of the yeast *Ty4* element: homology with Ty1, copia, and plant retrotransposons. Gene 122: 119–128.

Surgun Y, Çöl B, Bürün B (2012). Genetic diversity and identification of some Turkish cotton genotypes (*Gossypium hirsutum* L.) by RAPD-PCR analysis. Turk J Biol 36: 143–150.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.

Tomita M, Asao M, Kuraki A (2010). Effective isolation of retrotransposons and repetitive DNA from the wheat genome. J Integr Plant Biol 52: 679–691.

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009). Bursts of retrotransposition reproduced in *Arabidopsis*. Nature 461: 423–426.

Tu Z (2001). Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. P Natl Acad Sci USA 98: 1699–1704.

Nagaharu U (1935). Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Jpn J Bot 7: 389–452.

Voytas DF, Boeke DJ (1993). Yeast retrotransposons and tRNA. Trends Genet 9: 421–427.

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 H, Liu JS (2008). LTR RE landscape in *Medicago truncatula*: more rapid removal than in rice. BMC Genomics 9: 382.

Wei LJ, Xiao ML, An ZS, Ma B, Mason AS, Qian W, Li JN, Fu DH (2013). New insights into nested long terminal repeat retrotransposons in *Brassica* species. Mol Plant 6: 470–482.

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: 1072–1081.

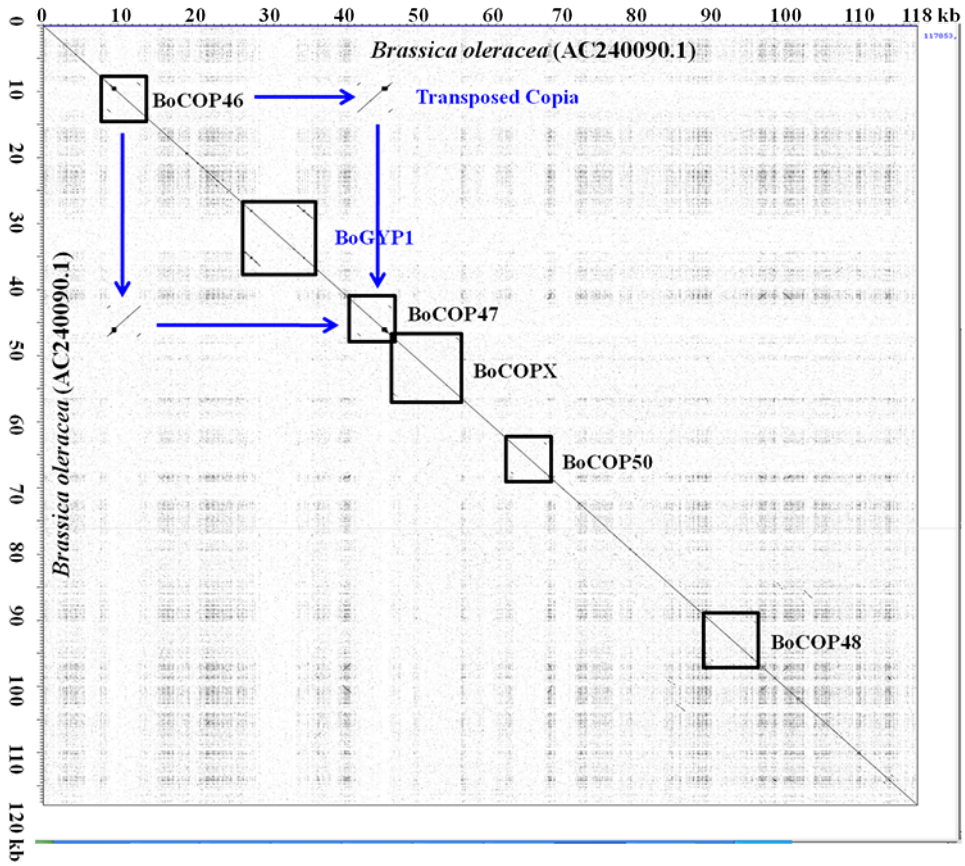
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavel A, Leroy P, Panaud O, Paux E et al. (2007). A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973–982.

Xiong Y, Eickbush TH (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9: 3353–3362.

Zhang X, Wessler SR (2004). Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. P Natl Acad Sci USA 15: 5589–5594.

Supplementary Table. List of 90 *Brassica* BACs screened for LTR retrotransposon identification in the present study. Most of the BACs showed the presence of elements (Table 1), while no element was detected from other BACs.

No.	Accession	No.	Accession	No.	Accession	No.	Accession
1.	AC189222.1	24.	EU568372.1	47.	AC189218.2	70.	CU695254.1
2.	AC189446.2	25.	EU579454.1	48.	AC155338.1	71.	AC155344.1
3.	AC166739.1	26.	EU579455.1	49.	AC189233.2	72.	AC189458.2
4.	AC155341.2	27.	AC240078.1	50.	CU984545.1	73.	AC189472.2
5.	AC189472.2	28.	AC240079.1	51.	EU642505.1	74.	AC189496.2
6.	AC189496.2	29.	AC240078.1	52.	EU642506.1	75.	FP340380.1
7.	AC241035.1	30.	AC240081.1	53.	EU579454.1	76.	FP340381.1
8.	AC241108.1	31.	AC240082.1	54.	AC122543.1	77.	FP340382.1
9.	AC241191.1	32.	AC240083.1	55.	EU581950.1	78.	AC234770.1
10.	AC241194.1	33.	AC240084.1	56.	EU579455.1	79.	AC234770.2
11.	AC241195.1	34.	AC240085.1	57.	EU568372.1	80.	AC237303.1
12.	AC241196.1	35.	AC240088.1	58.	EU579454.1	81.	AC189529.2
13.	AC241197.1	36.	AC240090.1	59.	AC166739.1	82.	AC232592.1
14.	AC241198.1	37.	AC240091.1	60.	AC155341.2	83.	AC237304.1
15.	AC241199.1	38.	AC240092.1	61.	AC166740.1	84.	AC152123.1
16.	AC241200.1	39.	AC240093.1	62.	AC155340.2	85.	AC189656.2
17.	AC241201.1	40.	AC240094.1	63.	AC166741.1	86.	AC189415.2
18.	AC149635.1	41.	AC183496.1	64.	AC155337.1	87.	AC241138.1
19.	AC183496.1	42.	AC183498.1	65.	AC155338.1	88.	AC155342.2
20.	AC183492.1	43.	AC189430.2	66.	CU695282.1	89.	AC241138.1
21.	AC183498.1	44.	EU579455.1	67.	EU642505.1	90.	AC241201.1
22.	AC240087.1	45.	AC232508.1	68.	EU642506.1		
23.	AC240089.1	46.	AC189263.2	69.	AC189222.1		



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