Gene 627 (2017) 94–105

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Identification and characterization of mobile genetic elements LINEs from *Brassica* genome

Faisal Nouroz[a,b,*], Shumaila Noreen[c], Muhammad Fiaz Khan[d], Shehzad Ahmed[e], J.S. Pat Heslop-Harrison[a]

[a] Department of Biology, University of Leicester, UK
[b] Department of Botany, Hazara University Mansehra, Pakistan
[c] Department of Zoology, University of Peshawar, Pakistan
[d] Department of Zoology, Hazara University Mansehra, Pakistan
[e] Department of Microbiology, Hazara University Mansehra, Pakistan

ARTICLE INFO

Keywords:
Retrotransposons
Brassica
LINEs
Reverse transcriptase
Diversity
Phylogeny

ABSTRACT

Among transposable elements (TEs), the LTR retrotransposons are abundant followed by non-LTR retrotransposons in plant genomes, the lateral being represented by LINEs and SINEs. Computational and molecular approaches were used for the characterization of *Brassica* LINEs, their diversity and phylogenetic relationships. Four autonomous and four non-autonomous LINE families were identified and characterized from *Brassica*. Most of the autonomous LINEs displayed two open reading frames, ORF1 and ORF2, where ORF1 is a *gag* protein domain, while ORF2 encodes endonuclease (EN) and a reverse transcriptase (RT). Three of four families encoded an additional RNase H (RH) domain in *pol* gene common to 'R' and 'I' type of LINEs. The PCR analyses based on LINEs RT fragments indicate their high diversity and widespread occurrence in tested 40 *Brassica* cultivars. Database searches revealed the homology in LINE sequences in closely related genera *Arabidopsis* indicating their origin from common ancestors predating their separation. The alignment of 58 LINEs RT sequences from *Brassica*, *Arabidopsis* and other plants depicted 4 conserved domains (domain II–V) showing similarity to previously detected domains. Based on RT alignment of *Brassica* and 3 known LINEs from monocots, *Brassicaceae* LINEs clustered in separate clade, further resolving 4 *Brassica-Arabidopsis* specific families in 2 sub-clades. High similarities were observed in RT sequences in the members of same family, while low homology was detected in members across the families. The investigation led to the characterization of *Brassica* specific LINE families and their diversity across *Brassica* species and their cultivars.

1. Introduction

Brassica economically an important and highly diverse genus of family Brassicaceae is informally known as cruciferous vegetables or mustard plants. *Brassica* is highly valuable due to its important agricultural and horticultural crops used as vegetables, oils, forage and ornamental purposes such as Chinese cabbage, broccoli, cauliflower, Brussels sprouts, kohlrabi, collards, turnips, black mustard, brown mustard and oilseed rape (canola). The genus *Brassica* is native to Europe and temperate Asia, while several species are growing well in Mediterranean regions (Monteiro and Lunn, 1999; Christopher et al., 2005). Three diploid species, *Brassica rapa* (AA), *B. nigra* (BB) and *B. oleracea* (CC) by their hybridization and polyploidization led to the speciation events yielding 3 allotetraploid species as *B. juncea* (AABB), *B. carinata* (BBCC) and *B. napus* (AACC) (Christopher et al., 2005; Sochorová et al., 2017). Of all *Brassica* species, *B. oleracea* (C-genome) is most important species with several important crops including many non-domesticated genotypes (Ostergaard and King, 2008). Like other

Abbreviations: aa, amino acids; BAC, bacterial artificial chromosomes; Bo, *Brassica oleracea*; Bp, base pair; Br, *Brassica rapa*; CDD, conserved domain database; CTAB, cetyltrimethylammonim bromide; Cv, cultivar; DIRs, *Dictyostelium* intermediate repeat sequence; DNA, Deoxyribonucleic acid; EN, endonuclease; EST, expressed sequence tags; g, gram; GAG, gag-nuclocapsid; INT, integrase; Kb, kilo base; LINEs, long interspersed nuclear elements; LTR, long terminal repeat; Mbp, mega base pair; Ml, milliliter; mM, milli molar; Mya, million years ago; NCBI, National Centre for Biotechnology Information; NJ, Neighbor-Joining; °C, degree Celsius; ORFs, open reading frames; PCR, polymerase chain reaction; Pol, polymerase; RNA, ribonucleic acid; RT, reverse transcriptase; SINEs, short interspersed nuclear elements; SSR, single sequence repeat; TEs, transposable elements; TIR, terminal inverted repeat; TPRT, target primed reverse transcription; TSD, target site duplication; UD, undetermined; UTR, untranslated regions; WGS, whole genome shotgun; ZF, zinc finger; ZK, zinc knuckle; µl, microliter

* Corresponding author at: Department of Botany, Hazara University Mansehra, Pakistan
E-mail address: faisalnouroz@hu.edu.pk (F. Nouroz).

http://dx.doi.org/10.1016/j.gene.2017.06.015
Received 20 January 2017; Received in revised form 5 June 2017; Accepted 8 June 2017
Available online 09 June 2017

Table 1
List of 40 *Brassica* accessions representing 6 *Brassica* species. ND: not determined.

No.	*Brassica* 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 gemmifera*	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 italic*	Precoce Di Calabria Tipo Esportazione	36	*B. carinata*	NARC-PK
17	*B. oleracea capitata*	Cuor Di Bue Grosso	37	*B. napus* x *B. nigra*	ND
18	*B. oleracea*	ND	38	*B. carinata* x *B. rapa*	ND
19	*B. juncea*	Kai Choy	39	*B. napus* x *B. nigra*	ND
20	*B. juncea*	Megarrhiza	40	*B. napus* x *B. nigra*	ND

plants, the genome of *Brassica* has also shown plasticity for transposition and retrotransposition of various transposable elements (TEs) as LTR retrotransposons (Nouroz et al., 2015a) and DNA transposons like hATs, Harbingers and CACTA (Zhang and Wessler, 2004; Nouroz et al., 2015b; Nouroz et al., 2016; Nouroz et al., 2017).

Among TEs, the retrotransposons are widespread in various eukaryotic genomes (Bennetzen, 2000). The genomic and extra-chromosomal copies of retrotransposons proliferate by an RNA intermediate copied into DNA by reverse transcriptase (Kapitonov and Jurka, 2008; Kapitonov et al., 2009). Eukaryotic retroelements are categorized on the basis of phylogeny of their reverse transcriptase (RT), *gag-pol* domain organization, proliferating devices and structural features into long terminal repeat (LTRs) retrotransposons and non-LTR retrotransposons or retroposons [representing long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)], DIRs-like elements and Penelope-like elements (Wicker et al., 2007).

Among non-LTR retrotransposons, the LINEs are copious elements with diversified sequences from various superfamilies such as L1, CR1, CRE, I, Jockey, NeSL, R2, R4, RandI, Rex1, RTE and Tx1. LINEs lack long terminal repeats (LTRs) and range in length from a kilobase (kb) to several kbs. 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 their transcription (Xiong and Eickbush, 1990; Jurka et al., 2007). The autonomous LINEs are characterized by TSDs of variable lengths, typically possessing 1 or 2 (sometimes 3) ORFs and a poly(A) tail at 3′ terminal end. They contain 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. LINEs lacking one or more protein domains required for their transposition are considered as non-autonomous (Malik et al., 1999; Jurka et al., 2007; Smyshlyaev et al., 2013). Based on structural hallmarks and phylogenetic analysis of LINEs RTs, the LINEs are grouped as LINE-1 (LI), retrotransposable elements (RTE), I, R2 and Jockey. These five groups are further subdivided into 28 clades (Kapitonov et al., 2009). The plant genomes mostly harbour L1 and few clades of RTE like elements (Heitkam and Schmidt, 2009; Wenke et al., 2009).

Although widely spread and thoroughly investigated in mammals, few LINE elements have been investigated and characterized in plants till now, although the number of reported LINEs is increasing with the advancement in genome sequencing. The first well characterized plant LINE was Cin4 detected in the A1 gene of *Zea mays* (maize) (Schwarz-Sommer et al., 1987). The well characterized LINEs from plants are *Karma* from *Oryza sativa* (Komatsu et al., 2003), *Llb* described in *Ipomoea batatas* (Yamashita and Tahara, 2006), *BLIN* from *Hordeum vulgare* (Vershinin et al., 2002) and *ATLN* from *Arabidopsis thaliana* (Noma et al., 2001). *BNR* LINE family was described from the genome of *Beta vulgaris* with 3 well characterized elements (*BNR1-BNR3*) displaying a size of 6.4–9.3 kb with two non-overlapping ORFs (Heitkam and Schmidt, 2009). Around 59,390 LINEs RT sequences were investigated from 23 plant genomes. These sequences fall in 2 clades as L1 and RTE of the 28 LINEs clades previously investigated in eukaryotic genomes (Heitkam et al., 2014).

The LINEs are investigated in more details in various mammal genomes, but barely investigated in plants. The availability of less data leads to LINE underestimation and ambiguous annotation of the plant genome projects. The present study was conducted to identify and characterize LINE elements in available sequenced *Brassica* and to study their diversity and evolutionary dynamics across *Brassica* germplasms.

2. Material and methods

2.1. Plant material for Brassica

Of the 40 *Brassica* accessions/cultivars investigated here, seeds from 32 *Brassica* accessions were brought from Warwick Research Institute (WRI), UK, two *B. juncea* accessions (NARC-1, NARC-II) and one *B. carinata* accession (NARC-PK) were brought from National Agriculture and Research Centre (NARC), Islamabad, Pakistan, one *B. juncea* accession (NATCO) was purchased from an Asian store at Leicester and DNA from four synthetic allohexaploid (2n = 6 ×; Ge et al., 2009) were provided by Dr. Xian Hong Ge (Huazhong Agricultural University, Wuhan, China). Standard CTAB method (Doyle and Doyle, 1990) was used for DNA extraction from fresh leaves grown in green house of Department of Biology, University of Leicester, UK. The names of the *Brassica* accessions are listed in Table 1.

2.2. Data mining and computational analysis

The full length or non-autonomous LINE elements were identified by dot plot (Sonnhammer and Durbin, 1995) comparison of *Brassica* homeologous BAC sequences (Supplementary Fig. 1), where gaps in the central diagonal lines running from one to opposite corner indicates an insertion. The detail analysis of these insertions led to the identification of few LINEs, which were further used as query in Blast searches against