Genome-wide evolutionary analysis of precursor sequences of MIR156 and MIR172 family members in *Brassica* species

Agarwal Ritika[1], Dhaka Namrata[2,3], Sharma Rita[2] and Jatain Aman[1]*
1. Amity University Haryana, Amity Education Valley, Gurugram (Manesar), Haryana 122413, INDIA
2. Crop Genetics and Informatics Group, School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Mehrauli Road, New Delhi 110067, INDIA
3. Department of Biotechnology, Central University of Haryana, Jant-Pali, Mahendergarh, Haryana 123031, INDIA
*amanjatainsingh@gmail.com

Abstract

MicroRNAs (miRNAs) are crucial regulators of plant development as well as stress responses. Synteny-based analysis of evolutionary patterns of miRNAs in related species offer useful insights to investigate the roles of miRNAs and their targets in shaping up crucial traits during evolution. With the recent availability of genome sequence data and well-elucidated evolutionary relationships with the model plant, Arabidopsis, members of Brassicaceae family can serve as useful models to study the patterns of miRNA evolution. Here, we report the evolutionary patterns of precursor and mature sequences of two conserved miRNAs of Arabidopsis, miR156 and miR172 in five Brassica species of the U's triangle. A synteny-based comparative genomics strategy was employed to identify Arabidopsis miRNA orthologs in five Brassica species and analyze them further for their chromosomal location, phylogeny and nucleotide variation. This was accomplished using several openly available sequence databases and resources.

The phylogeny and evolution of the precursor and mature sequences of miR156 and miR172 were analyzed across AA, BB and CC genomes of the Brassica species. The analysis of MIR156 family revealed a rather highly dynamic evolution as the total copies of MIR156 genes and their chromosomal location varied considerably across the family members. It also exhibited significant conservation across all the mature miR156 family member sequences with polymorphism at 11th and 14th positions of MIR156i. The phylogenetic analyses of pre-MIR156 sequences showed that MIR156i genes have possibly recently evolved in comparison to the other MIR156 family members and are therefore, interesting candidates for functional characterization and further evolutionary analyses. MIR172 sequences exhibited a greater conservation as compared to MIR156 across the different Brassica species with respect to their chromosomal location but showed a higher variability at the sequence level towards the mature miRNA forming 5' and 3' regions.

Keywords: *Arabidopsis thaliana*, Brassicaceae, MIR156, MIR172, SPL genes.

Introduction

MicroRNAs (miRNAs) are a class of small (21-24 nt) non-coding RNAs that play role in post-transcriptional regulation of gene expression. They are derived from long primary RNA (pri-miRNA) transcripts which are transcribed from *MIR* genes by RNA polymerase II.[9,13,28] Following transcription, these pri-miRNA transcripts undergo a cleavage to form a hairpin-like stem loop structure called precursor miRNA (pre-miRNA). These pre-miRNAs then undergo a sequential processing involving further cleavage to form mature miRNAs.

The mature miRNAs regulate target gene expression either by transcript cleavage or translational inhibition.[1,29] MiRNA-mediated regulation has been demonstrated in several biological processes in plants including development, organogenesis, genome maintenance and biotic and abiotic stress responses.

Based on the extent of their conservation, miRNAs may be grouped as known or novel miRNAs. Known miRNAs are ancient ones, which have been conserved during evolution and they have been crucial in regulating and coordinating plant development as well as response to biotic/abiotic stresses, while evolutionarily novel miRNAs are not conserved among different plant species.[16]

The mustard family, Brassicaceae, mostly comprising of cruciferous vegetables and oilseed crops, includes some of the most prominent dicot model systems such as *Arabidopsis* and *Brassica*.[6] Though, *Arabidopsis* is so far the most studied experimental system in plants, the genus *Brassica*, containing both diploid and polyploid species, provides an ideal system to study evolutionary relationships. *Brassica* species is characterized by several wild and cultivated species, classified into six genome types namely, AA, BB, CC, AABB, BBCC and AACC.[10,19]

Of these, there are five species of genus *Brassica* which are closely related as well as economically important and hence, it would be highly useful to study evolutionary relationship between these species. These six species include *Brassica rapa* (turnip), *Brassica nigra* (black mustard), *Brassica oleracea* (cabbage), *Brassica juncea* (brown mustard) and

Brassica napus (oilseed rape). These *Brassica* species share a genomic evolutionary relationship called 'Triangle of U'.[15,18]

This theory states that the three basic diploid species have generated three amphidiploid taxa by a subsequent intercross between each of the three ancestor species where three dipolid ancestors are *B. rapa* (n=10, AA), *B. nigra* (n=8, BB) and *B. oleracea* (n=9, CC) and the three amphidoploid species are *B. juncea* (n=18, AABB), *B. carinata* (n=17, BBCC) and *B. napus* (n=19, AACC).[17] As seen in their total number of chromosomes, these six species satisfy the chromosomal count as proposed in the theory. Recent studies have explored the evolutionary patterns of miRNAs in individual species. For instance, Sun et al[23,24] have described the effects of whole genomic triplication on the evolution of miRNAs in *B. rapa*. However, not much has been done with respect to studying the evolutionary patterns of miRNAs across the *Brassica* members of the U's triangle.

The main objective of the present study is to trace the molecular evolution of two miRNA families, miR156 and miR172, across the members of the U's triangle in *Brassica* species. miR156 is a highly conserved miRNA present in several plant species and is one of the highly expressed miRNAs in several tissues. It has been shown to modulate the tolerance to frequent heat stress in *Arabidopsis*. In response to the recurring heat stress, miR156 regulates the *SPL* (*SPOROCYTELESS*) genes.[22] miR156 has also been experimentally shown to control miR172 expression *via SPL* genes which further help in the expression of miR172b.[14] In addition, miR156 is also involved in modulation of lateral root development in *Arabidopsis*.[30] miR156 comprises a family of ten miRNAs named from a to j. The mature miRNA sequences corresponding to all these variants are 20-nt long.

miR172 comprises another family of conserved miRNAs that acts downstream of miR156 to support adult epidermal identity.[27] miR172 plays a key role in synchronizing the transformations within the various developmental phases and in stipulating floral organ characteristics by regulating the expression of a small group of AP2-like transcription factors.[32] These functions are conserved across monocot and dicot lineages. Till date, several efforts have been made to study the functions of miR172 and its targets in *Arabidopsis.*[14] These studies revealed role of miR172 in regulating diverse aspects of plant development such as cleistogamy and tuberization.

We ascertained the number of gene copies and their chromosomal allocation in both the families using the pre-existing genome sequence data of *Brassica* AA, BB and CC genomes and studied the phylogeny of their precursors. These analyses significantly contribute to current understanding of patterns of evolution of the precursor and mature sequences of miR156 and miR172 in *Brassica* species.

Material and Methods

Extraction of sequence data: The genome sequence data of five *Brassica* species, *Brassica rapa* (AA, n = 10), *Brassica nigra* (BB, n = 8), *Brassica oleracea* (CC, n = 9), *Brassica juncea* (CC, n = 18) and *Brassica napus* (BBCC, n = 18) were retrieved from the BRAD database v2.0[26] (http://brassicadb.org/brad). The sequences of *Arabidopsis* miRNA precursors used as the query were retrieved from MIRBase v22[11] (http://www.mirbase.org).

Identification of MIR156 and MIR172 family members in *Brassica* genomes: A local BLAST query dataset was created using the sequence data of all the precursor miRNAs of the model plant *Arabidopsis* from MIRBase v22[11] (http://www.mirbase.org). The precursor stem-loop sequences were retrieved in the fasta format. A stand alone version of BLAST was downloaded and BLASTn (E value 10, max. no. of hits 10) was executed with the query dataset against the genome data. The corresponding hits with more than 80% alignment to the query sequence were considered as orthologs and retrieved from the database.

Nomenclature: The orthologs were named as per the nomenclature prescribed by MIRBase database registry v22.1[5] (http://www.mirbase.org). The precursor sequences were labelled as MIR with a three-letter prefix, of which the first letter stands for the genus and the remaining two letters correspond to the *Brassica* species. For instance, bra-MIR156 represents precursor sequence from *Brassica rapa*. Additionally, individual members of the MIR156 and 172 families were differentiated with a letter suffixed as MIR156a, MIR156b, MIR156c etc. In case more than one sequences are mapped to a particular *Arabidopsis* MIR, they were named by adding a number as suffix.

Phylogenetic Analysis: Phylogenetic analyses was performed using MEGA v7[12,25] (https://www.megasoftware.net/download_form). The mature and precursor sequences of miR156 and miR172 families were aligned using the MUSCLE alignment algorithm in MEGA7. The aligned sequences were then used for constructing phylogenetic trees using Maximum Likelihood (ML) method. ML trees for the precursors (MIR156 and MIR172) were derived using a Tamura-Nei substitution model with uniform distribution rates among sites. The ML trees were derived at 1000 bootstraps, carrying out nucleotide substitution with complete deletion of missing sites.

Results and Discussion

Identification of orthologs of *A. thaliana* miRNA precursors in *Brassica* species and assignment of genomic blocks: With an aim to study the extent of conservation of *Arabidopsis* miRNAs in selected *Brassica* species, we extracted 54 *Arabidopsis thaliana* miRNA precursor sequences from miRBase v22.1[11] (http://www.mirbase.org). These sequences ranged from 85 to 377 nt in length. Further, the genome sequence datasets of

five major *Brassica* species *B. juncea*, *B. napus*, *B. rapa*, *B. nigra* and*B. oleracea* were downloaded.

The orthologous precursor miRNAs for *A. thaliana* were identified from the *Brassica* genome datasets using BLASTN v2.2 with default parameters (e-value 10, maximum number of hits 10). In case the same sequences mapped to two different queries, for example 398b and 398c in *B. juncea* and *B. oleracea*, the blast hits with the higher percentage identity were considered as orthologs. The percentage alignment for all the resulting hits was manually calculated and the ones with >80% alignment were considered as orthologs of *Arabidopsis* miRNA precursors in *Brassica* species.

We identified a total of 71 orthologs in *B. rapa*, 74 in *B. nigra*, 55 in *B. oleracea*, 165 in *B. juncea and*147 in *B. napus*. The comparative analysis showed that out of a total of 54 MIR sequences in *Arabidopsis*, 13 MIRs (including MIR164b, 165a, 166a, 166c, 166d, 167d, 168b, 171c, 396b, 397b, 408, 845a and 845b) may have been lost in one or more *Brassica* species, 3 (MIR156a, 156c and 156e) have undergone expansion and 21 show high conservation across the *Brassica* species (Table 1).

The remaining 17 miRNAs exhibited diverse evolutionary patterns. For instance, the number of copies of MIR156f and MIR167a were higher in *B. juncea*, whereas the number of copies were conserved in its progenitor species i.e. *B. rapa* and *B. nigra*. On the other hand, MIR160c with no copies in the progenitors (*B. rapa*, *B. nigra andB. oleracea*) had one copy each in both the amphidiploid species i.e. *B. juncea* and *B. napus*.

A similar pattern was observed in MIR166e orthologs in *B. juncea* in reference with *B. rapa* and *B. nigra*. MIR164a, 171a, 172a, 394a, 156g and 167c showed the presence of orthologs in one of the progenitors i.e. *B. rapa* and expansion (duplication) in the amphidiploid hybrid i.e. *B. napus*. MIR397a showed a similar pattern with presence in *B. oleracea* and expansion in *B. napus* along with MIR398a which has shown the presence in *B. rapa* and expansion in *B. juncea*. MIR166e showed an altogether different pattern with presence of orthologs in both the progenitors of *B. juncea* with complete loss in their amphidiploid hybrid.

In order to further gain insights into the patterns of evolution of these precursors in all five *Brassica* species, we analyzed their chromosomal locations and genomic block localization. The genomic blocks were assigned to all *Brassica* precursors based on the positional and chromosomal information of genomic blocks in *A. thaliana* as proposed earlier.[20]

Out of 54 precursor sequences taken from *A. thaliana*, 53 MIRNA precursors were assigned to genomic blocks while one remains unassigned due to lack of continuity in the genomic blocks (Table 2).

The R block had the highest number of miRNA precursors (8) followed by J block (6) which is very similar to the pattern observed in *A. thaliana* where majority of the studied pre-miRNA sequences also belong to the largest genomic block. Such diverse evolutionary patterns of MIRs have also been reported in other Brassicaceae members previously.[7,23,24]

Size and chromosomal distribution of MIR156 and MIR172 families in *Brassica* AA, BB and CC genomes: The miR156 and 172 are two of the most crucial known miRNAs that regulate various aspects of plant development including transition from the juvenile to adult phase of shoot development. MiR156 facilitates this transition by inhibiting *SPL* expression, while miR172 acts downstream and further promotes adult epidermal identity in *Brassica*.[27]

Therefore, we shortlisted these two miRNAs to further explore their evolutionary patterns. The study of five *Brassica* genomes identified a total of 177 members of MIR156 family; 24 from *Brassica rapa* (AA); 26 from *Brassica nigra* (BB); 19 from *Brassica oleracea* (CC); 61 from *Brassica juncea* (AABB) and 47 from *Brassica napus* (AACC) (Table 1).

The number of orthologs corresponding to each one of the MIR156 family member in *Arabidopsis* varied across the *Brassica* AA, BB and CC genomes ranging from zero in *B. oleracea* to 12 in *B. juncea*. Additionally, the chromosomal locations of MIRNA orthologs were quite diverse in various *Brassica* AA, BB and CC genomes (Table 2). *Brassica rapa* genome showed several copies of each MIRNA ranging from one each for MIR156b and 156g to four each for MIR156a, 156c and 156e.

All of these copies, with respect to each MIR156 family member were located on different chromosomes except MIR156e and 156f having two copies each on chromosome A06.

Further, the chromosomal localization in *Brassica nigra* also showed the same pattern of prevalence of copies throughout the genome for both the MIRNAs.

In contrast, *Brassica oleracea* showed that the number of copies with respect to each MIR156 family member were conserved with respect to *Arabidopsis*. *Brassica juncea* and *Brassica napus* showed the presence of copies of each of the MIR156 family members with respect to their progenitors wherein the orthologs were mostly observed to be in the same number and on the same chromosomal location as their progenitors as seen in MIR156d, 156i and 156j.

However, there was a considerable variation in the chromosomal locations for other MIRs. For example, for MIR156b, a single ortholog was present in chromosome A01 in both *Brassica rapa* (AA) and *Brassica juncea* (AABB), but absent in *Brassica napus* (AACC) (Table 2).

Table 1
List of orthologs of *Arabidopsis* miRNA precursors (MIRs) in *Brassica* species.

S. N.	Query							Number of orthologs				
	ID	Start	End	Strand	Length	Chr.	Block	*B.rapa* (AA)	*B.nigra* (BB)	*B.oleracea* (CC)	*B.juncea* (AABB)	*B.napus* (AACC)
1	ath-MIR156a	10676451	10676573	(-)	123	2	K-L	4	5	4	10	8
2	ath-MIR156b	15074899	15075081	(+)	183	4	U	1	2	2	3	1
3	ath-MIR156c	15415418	15415521	(-)	104	4	U	4	5	4	10	8
4	ath-MIR156d	3456632	3456749	(-)	118	5	R	2	2	0	4	2
5	ath-MIR156e	3867207	3867313	(+)	107	5	R	4	4	3	12	10
6	ath-MIR156f	9136106	9136237	(+)	132	5	Q	2	1	1	7	4
7	ath-MIR159a	27713233	27713416	(-)	184	1	E	3	4	3	8	6
8	ath-MIR160a	16340279	16340363	(+)	85	2	J	1	2	1	3	2
9	ath-MIR160b	9888982	9889070	(+)	89	4	U	2	2	2	4	4
10	ath-MIR160c	19009094	19009182	(-)	89	5	V	0	0	0	1	1
11	ath-MIR164a	19520752	19520864	(+)	113	2	J	2	2	0	4	4
12	ath-MIR164b	287584	287736	(+)	153	5	R	0	0	0	0	0
13	ath-MIR165a	78927	79037	(-)	111	1	A	0	0	0	0	0
14	ath-MIR165b	369831	370012	(-)	182	4	O	1	1	0	2	1
15	ath-MIR166a	19176108	19176277	(+)	170	2	J	0	0	0	0	0
16	ath-MIR166b	22922206	22922325	(+)	120	3	M-N	1	1	1	3	2
17	ath-MIR166c	2838635	2838773	(+)	139	5	R	0	0	0	0	0
18	ath-MIR166d	2840622	2840734	(+)	113	5	R	0	0	0	0	0
19	ath-MIR166e	16775520	16775662	(-)	143	5	S	0	1	0	0	1
20	ath-MIR166f	17516301	17516405	(+)	105	5	V	1	2	1	2	2
21	ath-MIR166g	25504798	25504919	(+)	122	5	X	1	1	1	2	2
22	ath-MIR167a	8108072	8108209	(+)	138	3	F	3	2	2	9	6
23	ath-MIR167b	23406168	23406276	(+)	109	3	M-N	0	1	1	1	1
24	ath-MIR168a	10578635	10578772	(+)	138	4	U	5	1	3	7	8
25	ath-MIR168b	18358788	18358911	(-)	124	5	V	0	0	0	0	0
26	ath-MIR171a	19073434	19073556	(+)	123	3	M-N	3	2	0	5	6
27	ath-MIR172a	11942914	11943015	(-)	102	2	I	1	1	0	2	2
28	ath-MIR172b	1188207	1188301	(-)	95	5	R	3	3	3	6	7
29	ath-MIR159b	6220646	6220841	(+)	196	1	A	1	2	2	4	3
30	ath-MIR167d	11137539	11137915	(+)	377	1	B	0	0	0	0	0
31	ath-MIR171b	3961348	3961464	(-)	117	1	A	2	2	1	4	4
32	ath-MIR171c	22930089	22930204	(-)	116	1	D	0	0	0	0	0
33	ath-MIR172c	3599776	3599908	(-)	133	3	F	1	1	1	2	1
34	ath-MIR172d	20587904	20588027	(+)	124	3	M-N	2	2	2	4	4
35	ath-MIR394a	7058194	7058310	(+)	117	1	B	1	2	0	3	2
36	ath-MIR394b	28568808	28568928	(+)	121	1	E	2	2	2	6	4
37	ath-MIR396a	4142323	4142473	(-)	151	2		1	1	1	2	2
38	ath-MIR396b	13611798	13611932	(+)	135	5	S	0	0	0	0	0
39	ath-MIR397a	2625950	2626056	(+)	107	4	O	0	1	2	1	4
40	ath-MIR397b	7878652	7878760	(-)	109	4	T	0	0	0	0	0
41	ath-MIR398a	1040938	1041042	(+)	105	2	K-L	1	0	1	2	2
42	ath-MIR398b	4691022	4691137	(+)	116	5	R	2	2	2	4	4
43	ath-MIR398c	4694694	4694808	(+)	115	5	R	2	2	1	4	4
44	ath-MIR408	19319814	19320031	(+)	218	2	J	0	0	0	0	0
45	ath-MIR156g	8412516	8412618	(-)	103	2	H	1	1	0	3	2
46	ath-MIR156h	22597012	22597117	(-)	106	5	W	2	2	2	3	4
47	ath-MIR159c	18994632	18994856	(+)	225	2	J	0	0	0	0	1
48	ath-MIR164c	9852674	9852775	(+)	102	5	Q	3	2	2	5	6
49	ath-MIR167c	1306622	1306781	(-)	160	3	F	1	1	0	2	2
50	ath-MIR172e	23988472	23988596	(+)	125	5	W	1	2	1	2	2
51	ath-MIR845a	12217406	12217568	(-)	163	4	U	0	0	0	0	0
52	ath-MIR845b	12214069	12214167	(-)	99	4	U	0	0	0	0	0
53	ath-MIR156i	19806917	19807113	(-)	197	1	C	2	2	2	5	4
54	ath-MIR156j	17589020	17589087	(-)	67	2	J	2	2	1	4	4
Total No. of Hits								71	74	55	165	147

Lost
Conserved
Expanded

Table 2
Chromosomal locations of miR156 and miR172 family members in *Brassica* AA, BB and CC genomes

MIRNA	Chromosomal location in *Arabidopsis thaliana*	Chromosomal locations in various *Brassica* genomes				
		***Brassica rapa* (AA)**	***Brassica nigra* (BB)**	***Brassica oleracea* (CC)**	***Brassica juncea* (AABB)**	***Brassica napus* (AACC)**
ath-MIR156a	2(-)	A09 A03 A04 A01	scaffold_10.1 scaffold_292.1 B02 B05 B07	C08 C01 (2) C07	A09 B02 (2) B05 (2) Not Determined (2) B01 A04 B03	A09_random C08 C04 C07 C01 A03 A04_random A01
ath-MIR156b	4(+)	A01	scaffold_10.1 B02	C07 C01	Not Determined A01 B05	C01_random
ath-MIR156c	4(-)	A03 A01 A09 A04	scaffold_10.1 B02 B07 scaffold_292.1 B05	C07 C01 (2) C08	Not Determined (2) B02 (2) B05 (2) B03 B01 A04 A09	C07 A03 C01 A01 C04 A09_random A04_random C08
ath-MIR156d	5(-)	A03 A10	B03 B08	NA	B08 A03 B02 A10	A03 A10
ath-MIR156e	5(+)	A02 A03 A06 (2)	B03 B02 B08 (2)	C09 C07 (2)	A02 (2) A03 (2) B08 B05 B02 (2) A10 A06 (3)	Ann_random A03 C09 A10 C03 C07_random (2) A06 (3)
ath-MIR156f	5(+)	A06 (2)	B08	C07	A06 (3) B02 A03 A02 (2)	A06 (3) C07_random
ath-MIR156g	2(-)	A07	B07	NA	B03 A09 (2)	A07 Cnn_random
ath-MIR156h	5(-)	A02 A10	B08 B02	C02 C07	A02 A10 B05	C02 A02 C09 A10
ath-MIR156i	1(-)	A06 A05	scaffold_159.1 B06	C06 (2)	A06 (3) Contig431 A05	A06 C06_random C06 A05
ath-MIR156j	2(-)	A05 A04	B01 B05	C04	A05 (2) B01 A04	C04 (2) A05 A04
ath-MIR172a	2(-)	A07	scaffold_106.1	NA	A07 A08	A07_random C04
ath-MIR172b	5(-)	A02 A10 A03	B02 B08 B03	C02 C09 C03	A02 B05 A10 A03 B02 B08	Ann_random (2) A01 Cnn_random A09_random A10 C03
ath-MIR172c	3(-)	A01	B01	C01	A07 B07	Cnn_random
ath-MIR172d	3(+)	A09 A04	B03 B04	C08 C04	B08 A09 A04	A09 C08 C04

					B04	A04
ath-MIR172e	5(+)	A02	B02 B03	C02	B05 B08	C02 A02

For the MIR156 and MIR172 gene families, the chromosomal locations are listed, without specifying the strand information of the orthologs. For each ortholog, the chromosome number is given along with the number of multiple copies in the parentheses, if any, within the same chromosome. Although, the strand information of *Arabidopsis* query sequences was specified in the parentheses

In contrast to the MIR156 family, members of MIR172 family were more conserved across the *Brassica* species. In the MIR172 family, there were 56 members including 8 from *Brassica rapa* (AA), 9 from *Brassica nigra* (BB), 7 from *Brassica oleracea* (CC), 16 from *Brassica juncea* (AABB) and 16 from *Brassica napus* (AACC). Their chromosomal distribution was comparatively inconsistent as observed for MIR156 (Table 2). With the exception of MIR172a, which was absent in *Brassica oleracea*, all other family members were present in all five *Brassica* species.

Phylogenetic analysis of MIR156 and MIR172 precursor sequences in *Brassica*: The phylogeny of MIR156 and MIR172 sequences was analyzed by Maximum Likelihood (ML) method. In the ML phylogenetic tree, all the orthologs of MIR156 could be divided into two clades: clade I and II (Figure 1a). Clade I contained all the nine MIR156 family members, except MIR156i, with orthologs from all six species. Clade I was further subdivided into seven subclades designated as IA, IB, IC, ID, IE, IF and IG. Subclade IA contained all the MIR156e and MIR156f orthologs from all the six species, showing high conservation between these two MIR sequences throughout the species, with only minor differences. Subclade IB contained MIR156g sequences from all the six species. Subclade IC possessed MIR156b orthologous sequences and subclade ID contained MIR156a and MIR156c orthologous sequences from all the six species. Subclade IE contained the MIR156d orthologous sequences.

On the basis of the genetic distance, bootstrap values and the nodal arrangement in the phylogenetic tree, this clade was identified to show a distinct evolutionary pattern between its orthologous sequences, as MIR156d sequences seemed to have diverged from its *Arabidopsis* homolog and hence, exhibit a clustered branching pattern distant from the *Arabidopsis* sequence. Subclade IF comprised of all the MIR156h sequences while subclade IG consisted of all the sequences of MIR156j across the six species. Clade II contained all the sequences corresponding to MIR156i from the six species.

The genetic distance and the clustering of all the MIR156i sequences in clade II suggest that it is the most recently evolved MIRNA amongst all the known MIR156 family members. Also, since the subclade IB had the maximum number of sequences belonging to MIR156a and MIR156c families, it seems the most conserved throughout the six species.

The ML phylogenetic tree constructed for MIR172 showed a clearer division between the clades as compared to MIR156. On the basis of the bootstrapping values, the tree was divided into three clades I, II and III (Figure 1 b). Clade I contained all the orthologs of MIR172e across all the studied *Brassica* species, showing slight deviation of all the *Brassica* species from *Arabidopsis*. Clade II contained all the MIR172b orthologs across *Brassica* family members. Clade III was further divided into three subclades: IIIA, IIIB and IIIC. Subclade IIIA showed the presence of all the orthologs of MIR172d across all five *Brassica* species. Subclade IIIB and IIIC have orthologs of MIR172c and 172a respectively with some evolutionary divergence from *Arabidopsis*. Overall, clade III showed the maximum number of sequences with MIR172c being the most evolved among all the MIR172 family members on the basis of genetic distance.

Sequence variation of MIR156 and MIR172 precursor and mature sequences: The mature miRNAs directly target genes for repression through post-transcriptional silencing.[2] This direct regulation is accomplished via binding of the mature miRNA molecules to a complementary site in the target genes and hence, the functional diversity of the miRNAs can be predicted by studying the nucleotide sequence diversity of MIRs as well as miRNAs.[4] Therefore, evolutionary patterns of the mature miRNA sequences were also studied with respect to their sequence diversity. The sequence alignment was carried out for 67 to 204 nt long orthologs of all the MIR156 and MIR172 family members.

The multiple sequence alignment in MIR156 revealed a total of 15 conserved sites throughout the 265 nt long sequence alignment (Supplementary Figure 1a) whereas, the sequence alignment of MIR172 orthologs showed a greater degree of conservation with a total of 35 conserved sites throughout the resulting 148 nt long sequence alignment (Supplementary Figure 1b). Importantly, the highly conserved residues were located near the most probable mature miRNA sequence in the precursor sequence.

The 15 conserved sites in the resulting sequence alignment of precursor sequences of MIR156 (positions 48, 49, 50, 51, 52, 53, 54, 56, 58, 59, 61, 62, 63, 137 and 150) showed conservation across the 16 nt subsequence from position 48 to position 63, with 3 polymorphic sites (positions 55, 57 and 60). The polymorphic sites with positions 55 and 57 showed single nucleotide polymorphism (G to A), whereas position 60 showed all possible types of nucleotide polymorphism (Supplementary Figure 1a).

Thus, the phylogenetic tree corresponding to MIR156 and its orthologs showed a clear subdivision into 2 clades, where clade II notably contained sequences corresponding to

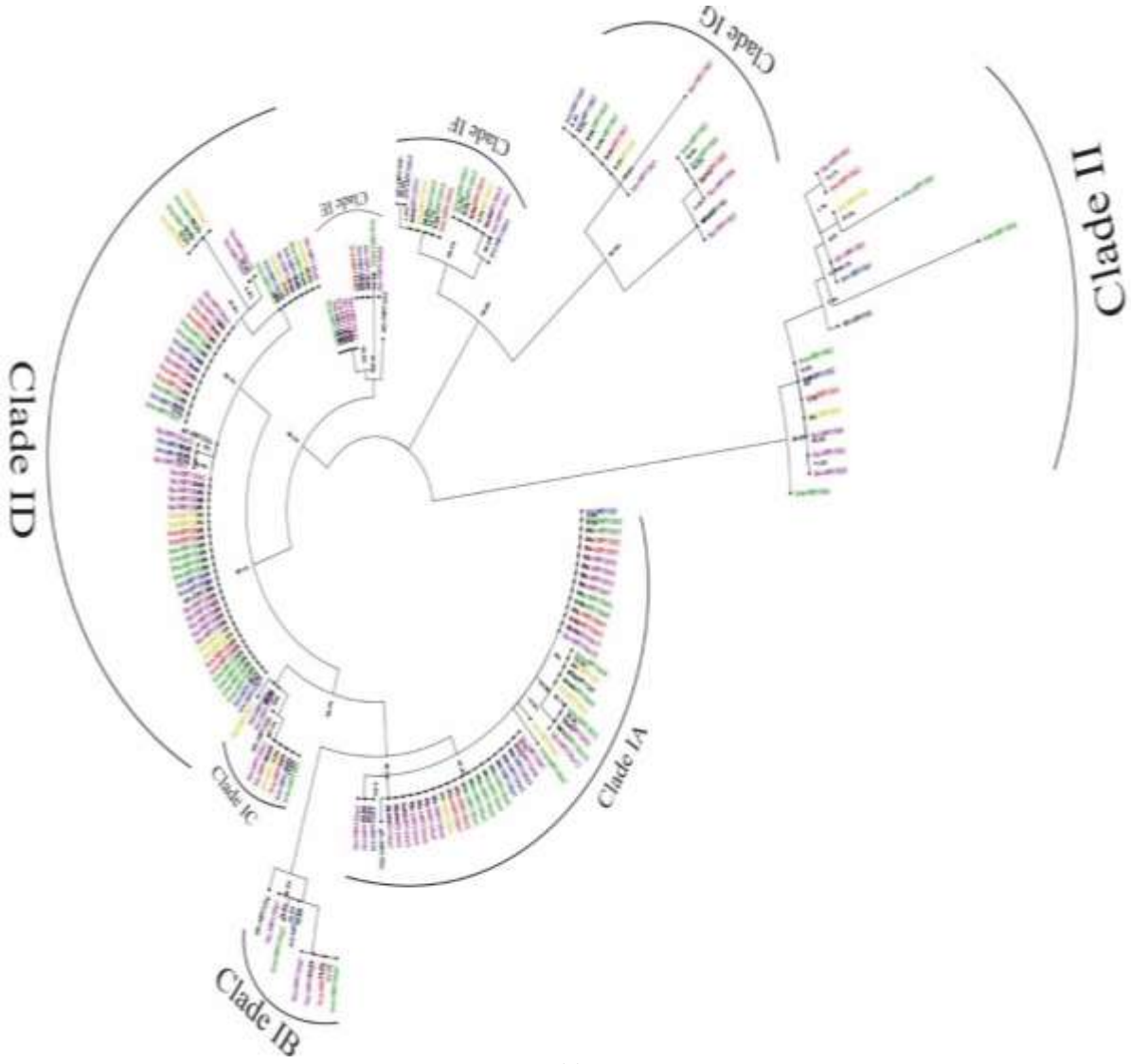
MIR156i. This suggests that MIR156i is the most recently evolved MIRNA family member and likely responsible for the phenotypic differences that stem due to the variation of miR156-SPL modules among different *Brassica* species.

In the sequence alignment of precursor sequences of MIR172, the degree of overall conservation was higher andthe conserved sequences (with 35 conserved sites) lied at the two extreme ends in stretches of two highly conserved regions (15-nt long and 24-nt long) across the sequence (Supplementary Figure 1b). This is because of high conservation in the probable mature miRNAs that result from the processing of these MIRs, from the ends of the precursor sequences (Supplementary Figure 1b).

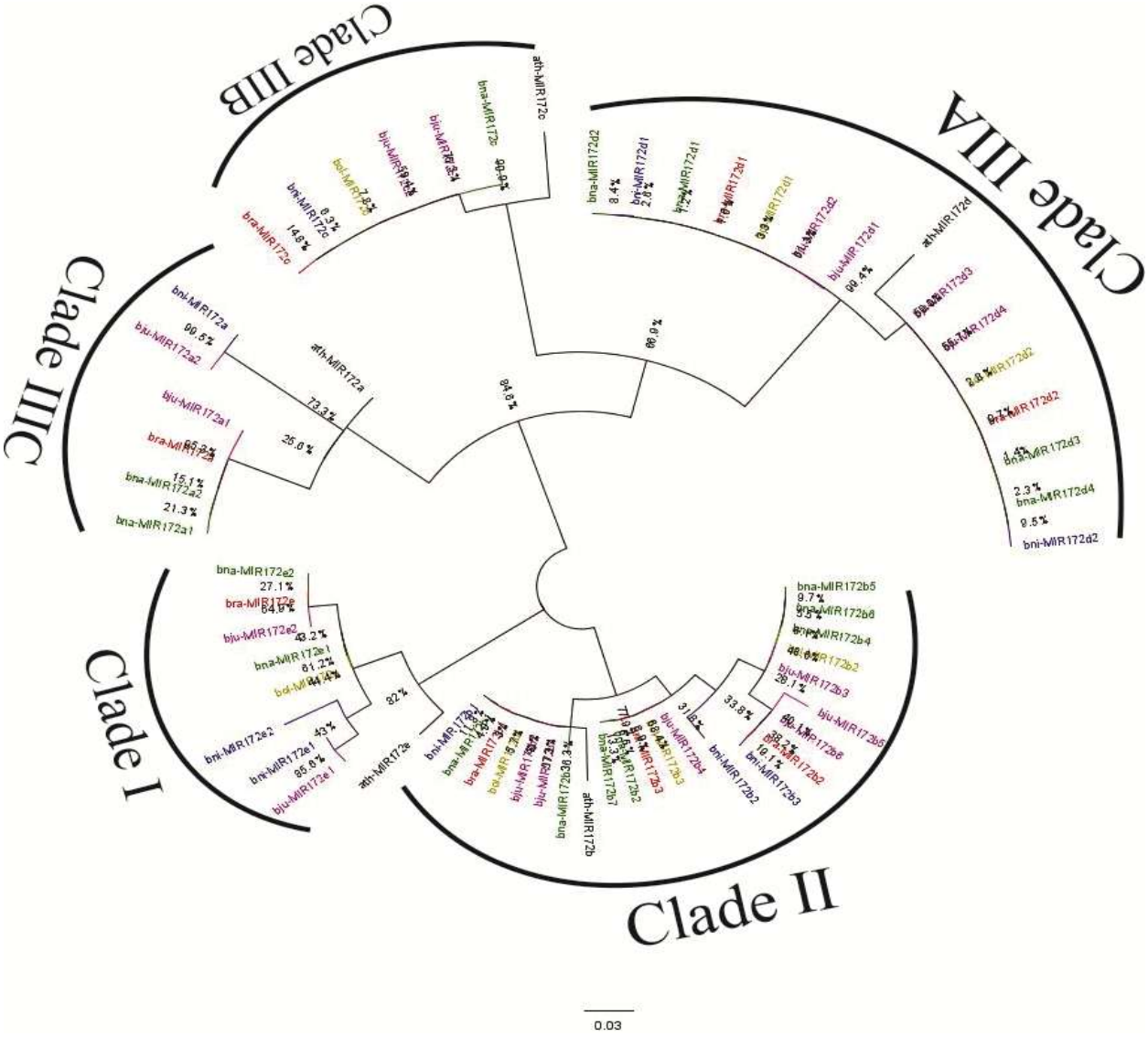
The two highly conserved regions across the 148 nt long sequence alignment were 15 and 24-nt long. The highly conserved subsequence with 15 nucleotides (from position 22 to position 36) had 3 polymorphic sites. All of these polymorphic sites exhibited single nucleotide polymorphism at position 24 (T to C), 26 (A to T) and 28 (C to T). The other highly conserved region has a total of 24 nucleotides (from position 110 to 133) with 2 polymorphic sites. These polymorphic sites also show single nucleotide polymorphism at positions 112 (A to G) and 132 (T to G). Thus, the phylogenetic study of MIR172 resulted in a subdivision into 3 clades, with clades I and II containing all the orthologs of MIR172e and MIR172b, respectively. This suggests that the role of MIR172b and MIR172e is more prominent in overall regulation of SPL genes by MIR172.

The conservation around the most probable mature miRNA sequence is another notable observation with high interspecies conservation in the *Brassicaceae* family in the precursor sequences of MIR156 and MIR172 family members, suggesting structural and functional conservation of miR156 and miR172 functions during evolution.

Further, we also analyzed the nucleotide diversity of *Arabidopsis* mature miRNA family members and observed that miR156 sequences from miR156a to miR156j had a 20-nt long mature miRNA sequence which was highly conserved throughout the miR156 family members in *Arabidopsis* (Figure 2).



(a)



(b)

Figure 1: Unrooted ML phylogenetic trees of MIR156 and MIR172 sequences using MEGA 7. The trees were edited for color coding in FIGTREE v.1.4.4. The trees were color coded for each node, representing precursor sequences of *Arabidopsis thaliana* with black, *Brassica rapa* with red, *Brassica nigra* with blue, *Brassica oleracea* with green, *Brassica juncea* with violet and *Brassica napus* with blue. The trees have been divided into clades based on the genetic distance and the bootstrapping values. The trees are drawn to scale, with branch length measured in the number of substitutions per site. a) A phylogenetic tree for all precursor sequences of MIR156 across *Brassica* AA, BB and CC genomes. The tree has been divided into 2 clades which have been further subdivided. b) A phylogenetic tree for all precursor sequences of MIR172 across *Brassica* AA, BB and CC genomes. The tree has been divided into 3 clades which have been further subdivided

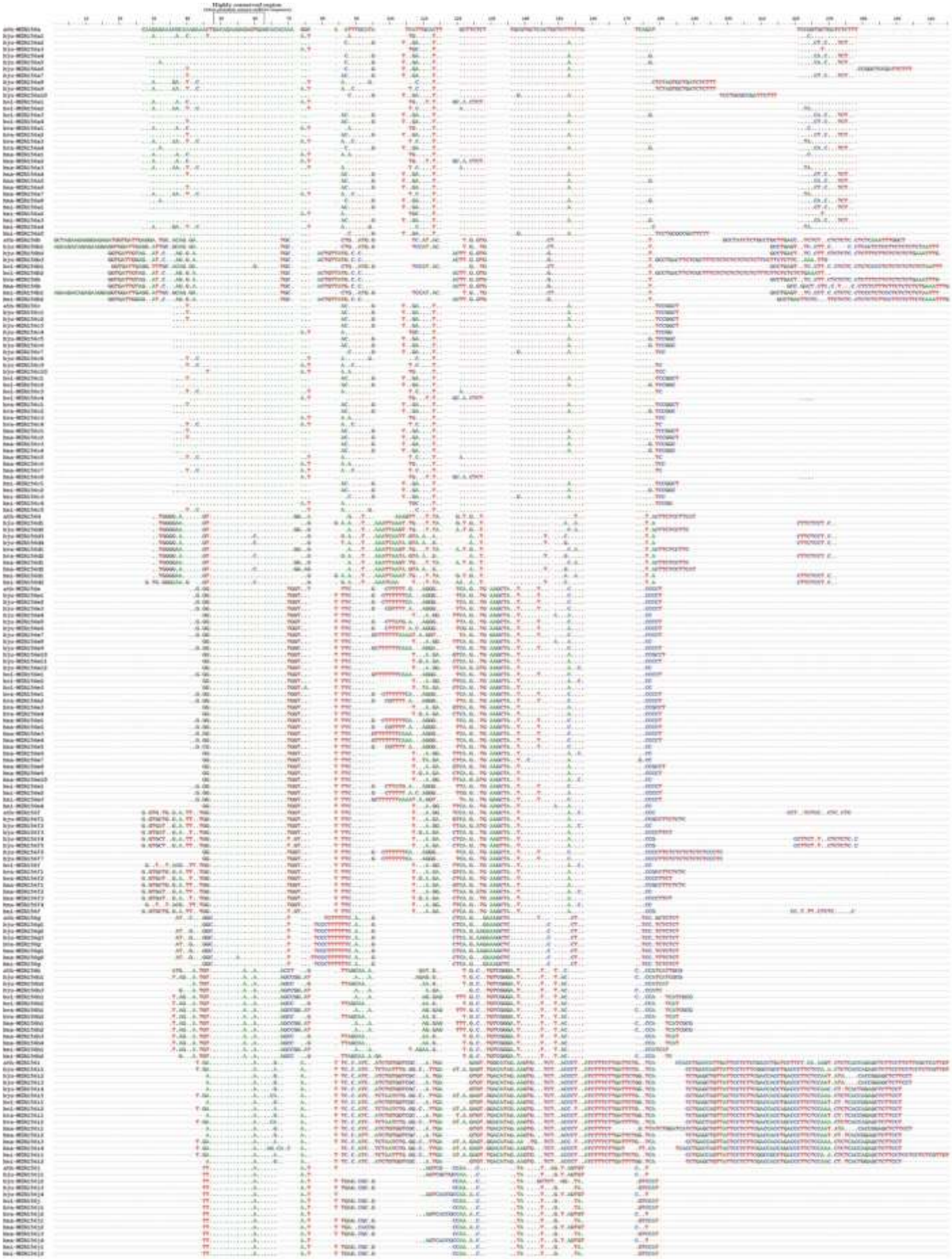
Interestingly, there were only four polymorphic sites (including site 1, 11, 14 and 20) throughout the 20-nt long mature miRNA, while the others were completely conserved.

Further, amongst these polymorphic sites, position 14 was the most prominent one, showing a single nucleotide polymorphism (U to A) in 3 of the 10 miR156 family members including miR156h, 156i and 156j. Sites 1, 11 and 20 also showed a single nucleotide polymorphism in miR156g (U to C), miR156h (G to A) and miR156i (C to G).

On the other hand, the mature miR172 sequences from miR172a to miR172e showed a variable length of 19 to 20-nt with much less overall conservation through the different miR172 family members. This analysis provides a foundation to further investigate the sequence conservation and diversity with respect to the sequence diversity in the target genes and correlate it with the evolution of corresponding miRNAs.

Conclusion and Future Scope

This study reports the evolutionary analysis of miR156 and 172 families across *Brassica* AA, BB and CC genomes.



(a)

(b)

Supplementary Figure 1: Multiple sequence alignment of precursor sequences of MIR156 and 172. The multiple sequence alignment was performed using MUSCLE and visualized in BIOEDIT. The conserved sites are designated by a dot and gaps as hyphens. a) Multiple sequence alignment of all pre-MIR156 family members across the Brassica rapa, Brassica nigra, Brassica oleracea, Brassica juncea and Brassica napus genomes have been presented along with Arabidopsis precursor sequences with a total of 15 conserved sites throughout 265 nt long sequence alignment. b) Multiple sequence alignment of all pre-MIR172 family members across the Brassica rapa, Brassica nigra, Brassica oleracea, Brassica juncea and Brassica napus genomes have been provided along with Arabidopsis precursor sequences with a total of 35 conserved sites throughout 148-nt long sequence alignment

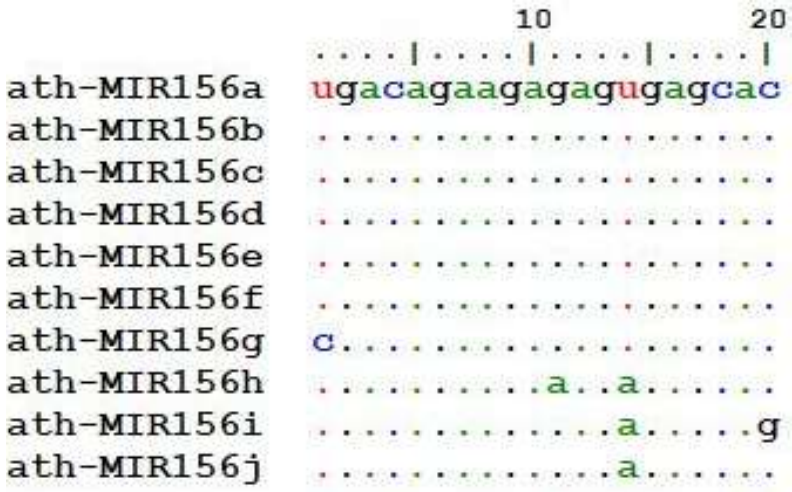


Figure 2: Multiple sequence alignment of all mature miR156 family members of *Arabidopsis.* The alignment shows a total of 4 SNPs in the 20-nt long sequence

A genome-wide analyse was performed and miRNA precursor sequences were identified in five different *Brassica* species using the information of *Arabidopsis* MIR precursor sequences. We analyzed these MIR orthologs on the basis of their size and chromosomal distribution throughout the AA, BB and CC genomes. A considerable variation was observed in the chromosomal locations of miRNA members along with a high level of variation in the number of orthologs, each miRNA family member possessed in the five *Brassica* species.

Besides the identification and analysis of MIR sequence locations, we further analyzed the sequence variation in mature and precursor sequences of MIR156 and MIR172 family members. We concluded that the precursor sequences of MIR156 exhibited significant sequence variation whereas the mature miR156 sequences were mostly conserved. Within the MIR156 precursor sequence, the region of the 'most probable looping structure' exhibited the maximum conservation.

Further, the mature miR172 sequences were less conserved across the 5' to 3' sites as compared to the miR156 sequence alignment. These findings shall be useful to trace the phylogeny and gain more insights as to whether the number of copies of each miRNA differs only due to the whole genome duplication events during the *Brassica* evolution or they point to the evolutionary trajectories followed by individual miRNAs. Further, it would be interesting to analyze in future how these differences affect the regulation of the SPLs and corresponding functions in different plant species.

Acknowledgement

ND thanks the Department of Science and Technology (DST) for DST INSPIRE Faculty Research Grant [DST/INSPIRE/04/2017/002597]. RS acknowledges the support from grant number BT/BI/03/004/2003(C) of the Government of India, Ministry of Science and Technology, Department of Biotechnology, Bioinformatics division and research grant from Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India (ECR/2016/001581).

References

1. Bartel D.P., MicroRNAs: genomics, biogenesis, mechanism and function, *Cell*, **116(2)**, 281–97 **(2004)**

2. Catalanotto C., Cogoni C. and Zardo G., MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions, *Int. J. Mol. Sci.*, **17(10)**, 1712 **(2016)**

3. Djami-Tchatchou A.T., Sanan-Mishra N., Ntushelo K. and Dubery I.A., Functional roles of microRNAs in agronomically important plants—potential as targets for crop improvement and protection, *Frontiers in Plant Science*, **8**, 378 **(2017)**

4. Felekkis K., Touvana E., Stefanou C.H. and Deltas C., MicroRNAs: a newly described class of encoded molecules that play a role in health and disease, *Hippokratia*, **14(4)**, 236–240 **(2010)**

5. Griffiths-Jones S., The microRNA registry, *Nucleic Acids Research*, **32(suppl_1)**, D109-11 (**2004)**

6. Hooks C.R. and Johnson M.W., Impact of agricultural diversification on the insect community of cruciferous crops, *Crop Protection*, **22(2)**, 223-38 **(2003)**

7. Jain A. and Das S., Synteny and comparative analysis of miRNA retention, conservation andstructure across Brassicaceae reveals lineage- and sub-genome-specific changes, *Funct Integr Genomics*, **16**, 253–268 **(2016)**

8. Jeong D.H., Thatcher S.R., Brown R.S., Zhai J., Park S., Rymarquis L.A., Meyers B.C. and Green P.J., Comprehensive investigation of microRNAs enhanced by analysis of sequence variants, expression patterns, ARGONAUTE loading and target cleavage, *Plant Physiology*, **162(3)**, 1225-45 **(2013)**

9. Kim Y.J., Zheng B., Yu Y., Won S.Y., Mo B. and Chen X., The role of Mediator in small and long noncoding RNA production in Arabidopsis thaliana, *The EMBO Journal*, **30(5)**, 814-822 **(2011)**

10. Koch M., Al-Shehbaz I.A. and Mummenhoff K., Molecular systematics, evolution andpopulation biology in the mustard family (Brassicaceae), *Annals of the Missouri Botanical Garden*, **90(2)**, 151-71 **(2003)**

11. Kozomara A., Birgaoanu M. and Griffiths-Jones S., miRBase: from microRNA sequences to function, *Nucleic Acids Research*, **47(D1)**, D155-62 **(2019)**

12. Kumar S., Stecher G. and Tamura K., MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, *Molecular Biology and Evolution*, **33(7)**, 1870-4 **(2016)**

13. Lee Y., Kim M., Han J., Yeom K.H., Lee S., Baek S.H. and Kim V.N., MicroRNA genes are transcribed by RNA polymerase II, *The EMBO J*, **23(20)**, 4051-60 **(2004)**

14. Li H., Zhang Q., Li L., Yuan J., Wang Y., Wu M., Han Z., Liu M., Chen C., Song W. and Wang C., Ectopic Overexpression of bol-miR171b Increases Chlorophyll Content and Results in Sterility in Broccoli (Brassica oleracea L var. italica), *Journal of Agricultural and Food Chemistry*, **66(37)**, 9588-97 **(2018)**

15. Maggioni L., Domestication of Brassica oleracea L, *Doctoral thesis*, Acta Universitatis Agriculturae Sueciae, Alnarp (Sweden) **(2015)**

16. Millar A.A., The Function of miRNAs in Plants, *Plants*, **9(2)**, 198 **(2020)**

17. Nagaharu U. and Nagaharu N., Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization, *J Japan Bot*, **7**, 389-452 **(1935)**

18. Prakash S. and Hinata K., Taxonomy, cytogenetics and origin of crop brassicas, a review, *Opera Botanica*, **55**, 3-57 **(1980)**

19. Rollins R.C., The Cruciferae of continental North America: systematics of the mustard family from the Arctic to Panama, Stanford University Press **(1993)**

20. Schranz M.E., Lysak M.A. and Mitchell-Olds T., The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes, *Trends in Plant Science*, **11(11)**, 535-42 **(2006)**

21. Sharma N., Panchal S. and Sanan-Mishra N., Protocol for artificial microRNA mediated over-expression of miR820 in indica rice, *American Journal of Plant Sciences*, **6(12)**, 1951-1961 **(2015)**

22. Stief A., Altmann S., Hoffmann K., Pant B.D., Scheible W.R. and Bäurle I., Arabidopsis miR156 regulates tolerance to recurring environmental stress through SPL transcription factors, *The Plant Cell*, **26(4)**, 1792-807 **(2014)**

23. Sun C., Wu J., Liang J., Schnable J.C., Yang W., Cheng F. and Wang X., Impacts of whole-genome triplication on MIRNA evolution in Brassica rapa, *Genome Biology and Evolution*, **7(11)**, 3085-3096 **(2015)**

24. Sun R., Guo T., Cobb J., Wang Q. and Zhang B., Role of microRNAs during flower and storage root development in sweet potato, *Plant Molecular Biology Reporter*, **33**, 1731-1739 **(2015)**

25. Tamura K. and Nei M., Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*, **10(3)**, 512-26 **(1993)**

26. Wang X., Wu J., Liang J., Cheng F. and Wang X., Brassica database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources, *Database*, **2015,** 1-8 **(2015)**

27. Wu G., Park M.Y., Conway S.R., Wang J.W., Weigel D. and Poethig R.S., The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis, *Cell*, **138(4)**, 750-9 **(2009)**

28. Xie Z., Allen E., Fahlgren N., Calamar A., Givan S.A. and Carrington J.C., Expression of Arabidopsis MIRNA genes, *Plant Physiology*, **138(4)**, 2145-2154 **(2005)**

29. Yu B. and Wang H., Translational inhibition by microRNAs in plants, miRNA Regulation of the Translational Machinery, Springer, 41-57 **(2010)**

30. Yu N., Niu Q.W., Ng K.H. and Chua N.H., The role of miR156/SPL s modules in Arabidopsis lateral root development, *The Plant Journal*, **83(4)**, 673-85 **(2015)**

31. Zhang B., MicroRNA: a new target for improving plant tolerance to abiotic stress, *Journal of Experimental Botany*, **66(7)**, 1749-61 **(2015)**

32. Zhu Q.H. and Helliwell C.A., Regulation of flowering time and floral patterning by miR172, *Journal of Experimental Botany*, **62(2)**, 487-95 **(2011)**.

(Received 31st March 2020, accepted 08th June 2020)