Plant Science 247 (2016) 35–48

Contents lists available at ScienceDirect

Plant Science

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

Microsynteny and phylogenetic analysis of tandemly organised miRNA families across five members of Brassicaceae reveals complex retention and loss history

Priyanka Rathore, R. Geeta, Sandip Das*

Department of Botany, University of Delhi, Delhi 110007, India

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form 4 March 2016
Accepted 5 March 2016
Available online 10 March 2016

Keywords:
miRNA
Tandem organisation
Evolution
Brassicaceae

ABSTRACT

Plant genomes are characterized by the presence of large miRNA gene families which are few in number. The expansion of miRNA families is thought to be driven by gene and genome duplication. Some members of these miRNA gene families are tandemly arranged and their analysis is of interest because such organisation may indicate origin through tandem duplication and also to investigate whether some such tandem clusters have similar expression patterns, and whether these are regulated through a common set of *cis*-regulatory elements (eg. promoters and enhancers). As a first step, we undertake a comprehensive study using micro-synteny analyses of tandemly organised miRNA families across the Brassicaceae spanning an evolutionary time scale of ca. 45 million years, among *Arabidopsis, Capsella, Brassica* and *Thellungiella* species, to address the following questions: Are most miRNA gene families present as tandem clusters? To what extent are these tandem patterns retained? To what extent can family sizes be ascribed to genome duplication? Our analysis of thirteen tandemly organised miRNA families revealed that synteny is largely conserved among *Arabidopsis thaliana, A. lyrata* and *Capsella rubella*, which form a clade spanning approximately between 6.2–9.8 my (Acarkan et al., 2000) [1]. On the other hand, comparison of sequences from these species with *Brassica rapa*, *B. oleracea* and *Thellungiella halophila*, which form a separate clade spanning 31 my (Franzke et al., 2011) [2] reveals many differences. The latter clade reveals several paralogous duplications that probably resulted from whole genome duplication, as well as disrupted synteny. Phylogenetic analyses of precursor sequences generally support the history inferred from synteny analysis. Synteny and phylogenetic analysis of six members of the tandemly organised miR169 family suggest that the Brassicaceae ancestral state consisted of a "dimer as a unit" which may have undergone direct local duplication to retain the transcriptional orientation followed by lineage specific changes. MiR169, to the best of our knowledge, is one of the largest tandemly organised miRNA gene family across plant kingdom and further analysis should reveal the generality of this pattern of evolution. The conserved organisation of miR395A-B-C and miR395 D-E-F as two clusters on same chromosome/scaffold across *A. thaliana*, *B. rapa* and *salsuginea* demonstrates retention of the large chromosomal segment across the two lineages. MiRNA family miR845 was detected only in *Arabidopsis* species and *Thellungiella* indicating a complex loss and retention history. MiR447A-B family was only found in *A. thaliana* indicating that it is a species-specific gene family of recent origin.

1. Introduction

MicroRNAs (miRNAs) are 20–24 nucleotide long noncoding RNAs that regulate target gene expression at various levels via target cleavage, interference with translation or chromatin re-organisation [3–5]. MiRNAs are usually encoded by discrete genomic loci that lie between other protein and non-protein coding genes; in animal systems, intronic encoded miRNAs, known as miRTrons have also been discovered [6,7]. MiRNA genes are regulated by type II promoters and are transcribed by RNA polymerase II [8] or rarely also RNA pol III [9]. Biogenesis of miRNA from primary to mature miRNA is a multistep process, and involves generation of a characteristic hairpin secondary structure [10]. The mature miRNA duplex (5p-miRNA:3-p-miRNA), which is processed from

Abbreviations: my, million years; mya, million years ago; RISC, RNA induced silencing complex.

* Corresponding author.
E-mail addresses: sdas@botany.du.ac.in, sandipdas04@gmail.com (S. Das).

http://dx.doi.org/10.1016/j.plantsci.2016.03.002
0168-9452/

the arms of the hairpin structure, is exported into the nucleus with the help of exportin/Hasty in plants [10,11]. One of the strands from the miRNA duplex, termed as the guide strand is then loaded onto the ribonucleoprotein complex-RISC (RNA-Induced Silencing Complex), which then binds to the target mRNA [5,12]. Positions 2–7 of the miRNA sequence, is called "seed sequence" and confers target specificity.

MiRNAs with similar mature sequences are grouped into families called miRNA families, each derived from a common ancestor [13]. Members of a miRNA gene family are highly conserved in their mature sequence, which decays in the precursor and the flanking regions [14]. Because members of the same family have identical mature and seed sequences (the latter of which is involved in target specificity), they target the same mRNA [15]. Some of the miRNA families are highly conserved through millions of years in both plants, and animals [16]. Plants have fewer but larger miRNA families (in terms of number of members) in contrast to animals which have numerous but smaller gene families [17].

Gene families including miRNAs are known to expand by gene and genome duplication, and evolve by neo-, sub- or non-functionalization [18,19]. In plants, some members of miRNAs families are present in tandem clusters and may show similar expression patterns [20–22]. Such clusters of miRNA families are presumed to have originated via tandem duplication, while interspersed blocks of duplicated miRNA families present elsewhere in the genome are presumed to arise via segmental duplication events [19–21]. In one such study comparative analysis of miRNA organisation across *A. thaliana*, *Oryza sativa*, *S. bicolor* and *Populus trichocarpa* does provide evidence that tandem duplication has played a major role in expansion of miRNA gene families in all but *P. trichocarpa* where segmental genomic duplication is likely to have played an equivalent role [22]. In humans proximally located miRNAs present on the same strand have been shown to be co-expressed as a polycistronic unit, suggesting that clustering may be related to co-expression [23]. Lacombe et al. [24] have shown the possibility that two miRNAs can be generated either from two tandemly-organised miRNA precursors or from a single stem-loop structure where they are arranged in tandem in rice, and suggest that such mechanism may be significant regulatory steps during biogenesis of miRNA [24]. The transcription of miRNA genes is known to be regulated via type II promoters that act as *cis*-regulatory regions, and are responsible for conferring spatio-temporal specificity of expression on the miRNA, which lies in the intergenic space in tandemly organised miRNA gene families. However, there are relatively few studies of tandem clusters of miRNA genes that could enable investigation of the forces underlying the origin and evolution of such clusters, and also of the *cis*-elements.

Here, we use genome synteny, secondary structure and phylogenetic analyses of tandemly clustered miRNA families to predict and validate miRNA genes, and to detect phylogenetic patterns of variation across taxa that might point to common underlying evolutionary factors [25,26]. We examine genomes of six members of Brassicaceae and find that the synteny of tandemly organised miRNA family is largely conserved in the *Arabidopsis-Capsella* lineage, but re-organised in polyploids such as *Brassica* and the distantly related *Thellungiella*. Unique and complex evolutionary histories across species were revealed in some tandemly organised families (e.g., miR169, miR395 and miR845).

2. Methods

2.1. Identification of tandemly organised miRNA genes

All miRNA sequences of *Arabidopsis thaliana* were retrieved from the miRBase release 21 (www.mirbase.org) [27] and sorted chromosome-wise based on positional information. The precursor sequences and their genomic coordinates were retrieved from TAIR (www.arabidopsis.org) [28] and used as queries to perform BLASTN [29] against genomes of five species of Brassicaceae, namely *Arabidopsis lyrata* v1.0, *Brassica rapa* v1.3, *B. oleracea* v1.1, *Capsella rubella* v 1.1 and *Thellungiella salsuginea* v 1.0 in the Phytozome database (http://www.phytozome.net) or BRAD (http://brassicadb.org/brad/index.php) [30,31] using default parameters (BLASTN program; e-value = −1; word length–11; Target type: Genome). All BLAST results were examined and sequences with High Scoring Pair (HSP) scores were chosen for further analysis. Sequences five kb upstream and downstream (or more depending on the result of the BLASTN analysis and coordinates of the subject in the HSP) were retrieved as putative miRNA precursors with flanking sequences. We limited further analyses to tandemly repeated miRNA families, i.e., where the distance between members of a family was 10 kb or less. These putative miRNA precursors along with their flanking sequences were compared with the whole genome assembly of *A. thaliana* using GenomeVISTA (http://genome.lbl.gov/cgi-bin/GenomeVista) [32] as previously described [33,34] to validate the homology and extent of conservation among the various species of Brassicaceae.

2.2. Identification of paralogs

Gene IDs of the microRNAs and the flanking protein coding genes from TAIR were retrieved; the protein coding gene IDs were employed to identify paralogs using the Paralogon tool for *A. thaliana* (now non-operational; http://wolfe.ge.tcd.ie/athal/dup) [35] for chromosome 1.

2.3. Secondary structure prediction

Secondary structures of the selected potential miRNA precursor sequences were predicted and generated using mfold (http://mfold.rna.albany.edu/?q=mfold/rna-folding-form) [36] using default parameters (linear RNA, folding temperature = 37 °C, 1 M monovalent ion, maximum internal bulge = 30 nucleotides). The structures with lowest free energy were individually analysed and compared with the reference, the secondary structure of miRNA precursor of *A. thaliana*. The individual structures were analysed for free folding energy of the structures (ΔG = kcal/mol), numbers of arms and loops per structure, complementarity between miRNA and its complementary sequence in the precursor arm-miRNA*, and internal loops within the arms were recorded [37]. Then these predicted secondary structures were manually compared with the previously identified homolog miRNAs of *A. thaliana* to reveal any structural variants.

2.4. Phylogenetic analysis

In order to establish orthology/paralogy relationships among the sequences, phylogenetic relationships within each miRNA family were estimated. Because identical miRNA sequences may come from different precursors, the entities analysed were the precursor sequences. These precursor sequences from each miRNA family were aligned by MUSCLE [38], as implemented in MEGA 6 [39]. The nucleotide substitution models for use in phylogenetic analyses were tested and identified by the Bayesian Information Criterion (BIC) in MEGA 6. Phylogenetic trees were estimated using Neighbour joining (NJ) and Maximum Likelihood (ML) methods as implemented in MEGA 6. Phylogenetic results were compared with results of synteny analysis to infer the evolutionary history of the sequences.