Plant Gene 6 (2016) 30–42

Contents lists available at ScienceDirect

Plant Gene

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

Computational identification and functional annotation of microRNAs and their targets from expressed sequence tags (ESTs) and genome survey sequences (GSSs) of coffee (*Coffea arabica* L.)

Karam Jayanandi Devi [a], Sreejita Chakraborty [a,b], Bibhas Deb [b], Ravi Rajwanshi [a,*]

[a] Department of Biotechnology, Assam University, Silchar, Assam 788011, India
[b] Bioinformatics Centre, Gurucharan College, Silchar, Assam 788004, India

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form 6 February 2016
Accepted 1 March 2016
Available online 4 March 2016

Keywords:
miRNA
Coffee
MFEI
RISC
PTGS
TAS3

ABSTRACT

MicroRNAs (miRNAs) are conserved, small non-coding endogenous RNAs, which are typically 21–24 nucleotides in length and play a pivotal role in post transcriptional gene regulation either by translational repression or cleavage. The regulation of target genes by miRNA has a central role in plant growth and development as well as in stress condition; however, only a few reports on coffee miRNA functions have been published. In the present study, we report the computational identification of miRNAs and their targets from expressed sequence tags (ESTs) and genome survey sequences (GSSs) of *Coffea arabica* as well as functionally annotated the target genes based on Gene ontology terms (GO). By following stringent filtering criteria, a total of 20 new potential miRNAs belonging to 13 different miRNA families (miR393, miR390, miR397, miR482, miR2118, miR414, miR1879, miR1134, miR1110, miR533, miR5809, miR426 and miR5212) were identified through homology search. The psRNATarget server predicted 142 potential target genes for 17 miRNAs and their probable functions were illustrated. Most of the predicted target genes encoded transcription factors and genes involved in plant growth and development, signal transduction, metabolism, defense and stress responses. The result from this study will shed light on understanding the functions of miRNAs in *C. arabica*.

1. Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNAs of 20–24 nucleotides in length, endogenously expressed and act as negative regulators of target genes. MiRNAs in plant were first reported from *Arabidopsis thaliana* in early 2002 (Reinhart et al., 2002). Plant miRNA genes that exist in introns as well as exons are transcribed by RNA polymerase II to primary miRNA transcripts (pri-miRNAs) that contain cap structure as well as poly(A) tail, which are subsequently processed to form stem loop structures. Mature miRNAs are produced from these precursor miRNAs by the endoribonuclease III-like enzyme, dicer-like-1 (DCL1) and exported to the cytosol via HASTY5. In the cytosol, the mature miRNAs are incorporated into an argonaute (AGO) containing RNA-induced silencing complex (RISC) that interacts with the complementary sites of the target gene transcript. The miRNAs mediate gene regulation through a phenomenon known as Post Transcriptional Gene Regulation (PTGS) that relies on sequence based interaction with target mRNAs and results in degradation of target transcript or attenuation of translation (Bartel, 2004; Voinnet, 2009).

Plant encodes other small non-coding RNAs such as small interfering RNAs (siRNAs), transacting siRNAs (ta-siRNAs) and they all differ from miRNAs on the basis of the following features: 1) all miRNAs are encoded by MIR genes and subsequently form pre-miRNAs which are exported to the cytoplasm. 2) All pre-miRNAs can form the characteristic stem-loop structure with high negative minimal folding free energy (MFE). 3) All miRNAs possess a miRNA* sequence (a complementary sequence on the opposite site of the stem-loop). Plant miRNAs show a high degree of perfect or near perfect complementarity with their targets. The identification of target genes is an important step for determining the biological function of miRNAs (Rhoades et al., 2002). Multiple miRNAs may control the expression of a single gene or a single miRNA may be involved in regulating the expression of multiple genes (Dehury et al., 2013). Therefore, the identification of target is fundamental for functional analysis of miRNAs. Plant miRNAs have been implicated in various developmental processes such as leaf morphogenesis, floral organ development, flowering time, organ polarity, shoot and root development, vascular development, vegetative phase change and seed development (Zhang et al., 2006a, Yang et al., 2007, Rubio-Somoza and Weigel, 2011). They are also involved in gene regulation during different abiotic stress conditions of the plant such as drought, cold, heat, salinity, chilling, oxidative, nutrient deficiency, hypoxia, UV-B radiation as well as during biotic stress from bacterial pathogenesis, fungal, viral infection (Zhang

* Corresponding author.
E-mail address: rrajwanshi@gmail.com (R. Rajwanshi).

http://dx.doi.org/10.1016/j.plgene.2016.03.001

et al., 2006a, Lu and Huang, 2008, Sunkar, 2010, Sunkar et al., 2012, Rajwanshi et al., 2014). Some plant miRNAs have also been reported to be involved in the regulation of genes such as DCL1 and AGO1 known to play an important role in biogenesis of small RNA as well as in directing siRNA biogenesis (ta-siRNAs) (Yang et al., 2007).

The mature miRNA sequences are highly conserved across the plant kingdom (Zhang et al., 2006b). This conservation of sequences has provided a powerful tool for the identification of novel miRNA genes through comparative genomics based approach. Homology search based on evolutionary conservation has lead to the identification of miRNA families from various plant species whose whole genome sequence is not available. Apart from homology search using Basic Local Alignment Search Tool (BLASTn), another criteria which is considered for miRNA identification is the secondary hairpin loop structure of the pre-miRNA sequence. Expressed Sequence Tags (ESTs) and Genome Survey Sequences (GSSs) provide an alternative powerful data source for plants whose complete genome is not yet available. A large number of miRNAs have been identified in plants by computational and experimental approaches which include high throughput techniques such as direct cloning and deep sequencing. ESTs and GSSs analysis with the aid of high throughput bioinformatics tools have efficiently predicted miRNA with a high degree of accuracy from various plant species including many important agricultural crop species such as wheat (Yao et al., 2007), tomato (Yin et al., 2008), soybean (Zhang et al., 2008), potato (Zhang et al., 2009b), maize (Zhang et al., 2009a), barley (Colaiacovo et al., 2010), *Solanaceae* (Kim et al., 2011), cotton (Wang et al., 2012), sorghum (Katiyar et al., 2012), prairie sunflower (Sahu et al., 2013), garlic (Panda et al., 2014).

The genus *Coffea* belongs to the family Rubiaceae and is a woody perennial eudicotyledons. This genus contains about 100 species but *Coffea arabica* and *Coffea canephora* are two most economically important and widely cultivated species across 80 countries in four continents, Brazil being the leading coffee producing country in the world. *C. arabica* is an allotetraploid ($2n = 4\times = 44$) formed by natural hybridization between ancestors *C. canephora* and *Coffea eugenioides* that took place 10,000 years ago (Lashermes et al., 1999). The expression of these two sub-genomes comprises the transcriptome of *C. arabica* (Vidal et al., 2010). *C. arabica* is one of the most important agricultural commodities by contributing 70% of worldwide production owing to its flavor, aroma, and stimulating effects of caffeine. The aromatic compositions of coffee is extremely complex due to thermal reactions during the roasting process and generates many of the volatile and non-volatile compounds, and are therefore not present in green beans of coffee (Flament, 2001). However, terpenic compounds that play an important role in plant primary and secondary metabolism survive the thermal reactions and are detected in green beans of coffee (Terra et al., 2013). The predominant monoterpenes in *C. arabica* flowers are geranial, linalool, nerol and the minor ones are limonene, myrcene, β-ocimene terpinolene and alpha terpineol (Emura et al., 1997). The development and production of coffee in different countries are affected by environmental factors such as drought, salinity and heat as well as biotic factors namely infection of coffee leaves by the fungus *Hemileia vastatrix* that causes coffee leaf rust, a major disease in coffee (Ganesh et al., 2006; DaMatta and Ramalho, 2006; Lima et al., 2013). Drought conditions affect coffee production by hampering the development of flowers, coffee bean and even cause plant death if the drought period is severe. A thorough study on miRNAs and their target genes function in coffee could provide new insight into the mechanism of coffee development, metabolic processes, and response to various abiotic and biotic stresses. The large numbers of ESTs and GSSs of *C. arabica* provide a good opportunity for predicting novel miRNA and their respective target genes. 17,4275 EST and 3,757 GSS are available in the National Centre for Biotechnology Information (NCBI) for *C. arabica*. There are a total of 5,939 plant miRNAs deposited in the public miRNA database (miRBase, release 19) including many economically important crop species. However, no miRNA has been deposited in miRBase for the genus *Coffea* which includes *C. arabica* despite its economic importance. Considering the economic importance of coffee, we used an EST and GSS based comparative approach for the identification and characterization of conserved miRNAs and their putative target genes. The target genes of the newly predicted miRNAs were functionally annotated based on gene ontology (GO) terms: molecular function, biological processes, and cellular component. Furthermore, the pathways involved by the target genes were investigated using KASS (KEGG Automatic Annotation Server). Although a few miRNAs have been computationally identified from *C. arabica* ESTs (Rebijith et al., 2013; Akter et al., 2013; Guilherme et al., 2014), the present study aimed towards the mining of additional novel miRNAs from ESTs and GSSs as well as performed an extensive analysis of the target genes.

2. Materials and methods

2.1. Reference set of miRNAs and sequence database

A total of previously identified 5939 plant mature miRNA sequences from Viridiplantae group were retrieved from the miRNA database miRBase (http://www.mirbase.org/) (Release 19: November 2012; Griffiths-Jones, 2004; Griffiths-Jones et al., 2008). The mature miRNAs were duplicated and made non-redundant in order to avoid overlapping of miRNAs. These unique sequences were defined as the reference sequence set and used for identifying miRNAs in coffee ESTs and GSSs. Publicly available coffee 17,4275 EST and 3757 GSS (as of January 1, 2013) were downloaded from National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/). Local database for standalone BLAST were constructed for coffee ESTs and GSSs by using makeblastdb application. Non-redundant protein database was downloaded from NCBI.

2.2. Computational resources

For homology searches BLAST-2.2.27 + program (Altschul et al., 1990; Camacho et al., 2009) was downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/) and locally installed. Secondary structure analysis of miRNA precursors was performed by online version of MFOLD (http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form) (Zuker, 2003). psRNATarget: a plant small RNA Target Analysis Server was used for predicting the targets of the newly identified miRNA (http://plantgrn.noble.org/psRNATarget/) (Dai and Zhao, 2011). Phylogenetic analyses of pre-miRNAs were carried out using web based ClustalW (Thompson et al., 1994) and MEGA version 6.0. Gene Ontology (GO) terms for the target genes were assigned using AmiGO version 1.8 and KASS (KEGG Automatic Annotation Server) (http://www.genome.jp.tools/kaas/) (Ogata et al., 1999; Carbon et al., 2009) was employed for pathway analyses.

2.3. Prediction of potential C. arabica miRNAs

Precursor miRNA sequence identification was carried out using comparative genome based homolog search and secondary structure analysis as described by Zhang et al. (2006b, c; 2007). Mature miRNA sequences were used for BLASTn search as they are evolutionarily highly conserved than their precursor sequences. The non-redundant miRNA reference sequence data were subjected to BLASTn searche for coffee homolog of miRNA against the locally prepared coffee EST and GSS databases. The initial BLAST search was performed with the program of BLAST-2.2.27 +, that was locally installed. The identification of homolog sequences were based on the following parameter: expect values at 10; low complexity chosen as the sequence filter, the default word-match size between the query and database sequences kept at 7. Hits of EST/GSS sequences having 0–2 mismatches and whose matched region with miRNA contains at least 18 nucleotides were chosen to be potent miRNA precursor candidates. The selected ESTs and GSSs were compared for similarity against each other to remove redundant sequences. The filtered sequences were further subjected to