Gene 667 (2018) 112–121

Contents lists available at ScienceDirect

Gene

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

Research paper

Genome-wide identification and characterization of lncRNAs and miRNAs in cluster bean (*Cyamopsis tetragonoloba*)

Sarika Sahu[a,b,1], Atmakuri Ramakrishna Rao[a,*,1], Jaya Pandey[a], Kishor Gaikwad[d], Sabari Ghoshal[b], Trilochan Mohapatra[c]

[a] *ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India*
[b] *Amity University, Noida, Uttar Pradesh, India*
[c] *Indian Council of Agricultural Research (ICAR), New Delhi, India*
[d] *ICAR-National Research Center on Plant Biotechnology, New Delhi, India*

ARTICLE INFO

Keywords:
Long non coding RNAs
Endogenous target mimic
Cluster bean
Micro RNA
Principal component analysis

ABSTRACT

Long non coding RNAs (lncRNAs) are a class of non-protein coding RNAs that play a crucial role in most of the biological activities like nodule metabolism, flowering time and male sterility. Quite often, the function of lncRNAs is species-specific in nature. Thus an attempt has been made in cluster bean (*Cyamopsis tetragonoloba*) for the first time to computationally identify lncRNAs based on a proposed index and study their targeted genes. Further, these targeted genes of lncRNAs were identified and characterized for their role in various biological processes like stress mechanisms, DNA damage repair, cell wall synthesis. Besides, lncRNAs and miRNAs bearing Simple Sequence Repeats (SSRs) were identified that contribute towards biogenesis of small non-coding RNAs. Moreover, five novel endogenous Target Mimic lncRNAs (eTMs) were identified that may disrupt the miRNA-mRNA regulations. For easy understanding and usability, a database CbLncRNAdb has been developed and made available at http://cabgrid.res.in/cblncrnadb.

1. Introduction

With the emergence of high throughput sequencing technology, RNAome analysis has become relatively easier to study. It includes the study of non-coding RNAs (ncRNAs) (Wang et al., 2015). The ncRNAs are a group of RNAs with no coding potential (Mercer et al., 2009) and recent studies have shown that they are functionally and spatio-temporally expressed in tissues (Greilhuber et al., 2005; Lakhotia, 2016). Among the ncRNAs, lncRNAs and miRNAs have been considered to play an essential role in multiple biological processes (Derrien et al., 2012). The functional mechanism of lncRNAs is diverse and highly significant as only expression of lncRNAs is sufficient enough to regulate the nearby or distant genes through post transcriptional chromatin complexes. (Mercer et al., 2009; Ponting et al., 2009; Wilusz et al., 2009). Ponting et al. (2009) divided lncRNAs into sense, anti-sense, bi-directional, intronic and inter-genic on the basis of their cellular localization. Moreover, the LncRNAs are generally expressed at low levels and lack sequence similarities among species (Marques and Ponting, 2014).

A vast number of lncRNAs have been reported in animals, whereas very few are explored in plants till now (Liu et al., 2012). State-of-art technology like next-generation sequencing eases the analysis of thousands of lncRNAs in model plant organism like *Arabidopsis thaliana* (Zhang et al., 2013; Zhu et al., 2013; Xie et al., 2014). Heo and Sung (2011) reported the regulation of flowering time due to the association of lncRNAs like cool-assisted intronic non-coding RNA (COOLAIR) and cold-assisted intronic non-coding RNA (COLDAIR) with epigenetic repression of Flowering Locus C (FLC) in *Arabidopsis*. Another lncRNA: long-day-specific male-fertility-associated RNA (LDMAR) is also found to be involved in photoperiod regulated male sterility in rice (Ding

Abbreviations: lncRNAs, long non coding RNAs; SSR, Simple Sequence Repeat; eTM, endogenous Target Mimic; ncRNAs, non-coding RNAs; COLDAIR, cool-assisted intronic non-coding RNA; COOLAIR, cold-assisted intronic non-coding RNA; FLC, Flowering Locus C; LDMAR, long-day-specific male-fertility-associated RNA; miRNAs, micro RNAs; AtIPS1, induced by phosphate starvation1; ORF, Open Reading Frame; CPC, Coding Potential Calculator; PLEK, ***p***redictor of ***l***ong non-coding RNAs and m***e***ssenger RNAs based on an improved ***k***-mer; FPKM, fragments per kilobase of transcript per million mapped reads; HPlncRNAs, highly probable lncRNAs; PCs, principal components; PCS, principal component scores; MFE, minimum fold energy; TE, transposable element; IIS, Internet Information Services; ODBC, open database connectivity; LPlncRNAs, low probable lncRNAs; XyG XT, xyloglucan 6-xylosyltransferase; GMGT, galactomannan galactosyltransferase; pre-miRNA, precursor miRNA; TNRs, trinucleotide repeats; ndG, normalized binding free energy

* Corresponding author.
E-mail address: ar.rao@icar.gov.in (A.R. Rao).
[1] Equally contributed to this work (Joint First Authors).

https://doi.org/10.1016/j.gene.2018.05.027
Received 18 January 2018; Received in revised form 24 April 2018; Accepted 8 May 2018

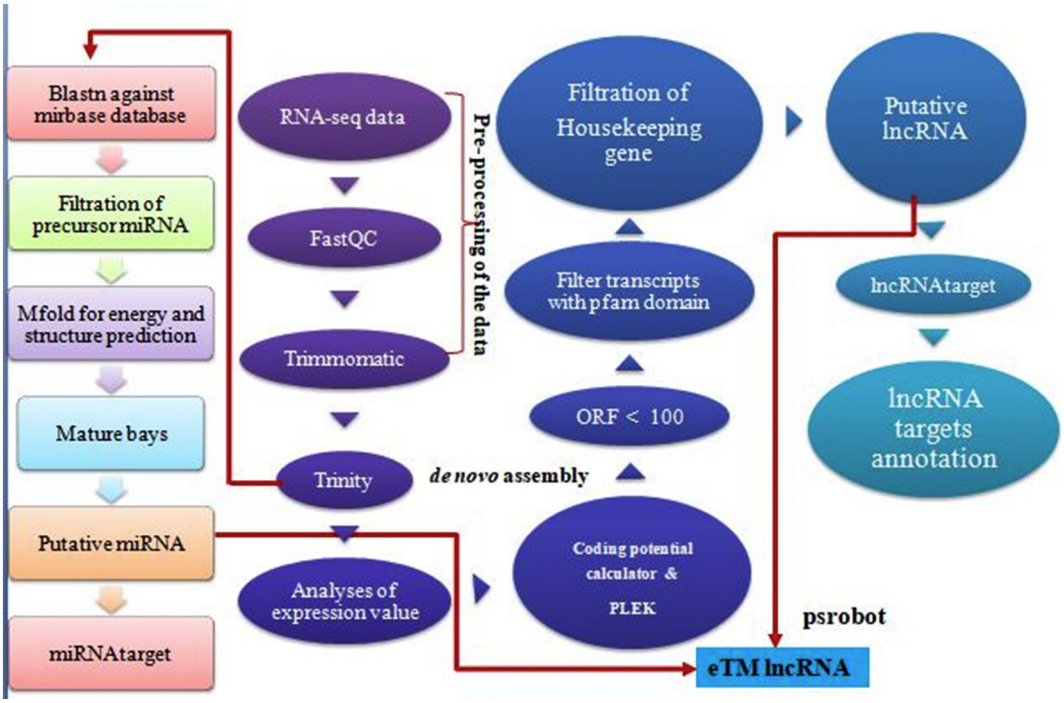

Fig. 1. An integrative computational pipeline for the systematic identification of noncoding RNAs (lncRNA and miRNA) and their targets.

et al., 2012) and ripening in tomato (Zhu et al., 2015).

In contrast to lncRNAs, micro RNAs (miRNAs) are very short ncRNAs usually made from long self-complimentary precursor sequences (Rodriguez et al., 2004). miRNAs play key role in various post transcriptional processes such as biotic/abiotic stress responses, tissue differentiation, growth and development (Kidner and Martienssen, 2005). Besides, they are highly conserved among distantly related plant species from non-vascular bryophytes to monocots (Zhang et al., 2006). So far, the miRNAs of many leguminous plants like *Acacia auriculiformis*, *Arachis hypogaea, Lotus japonicas*, *Acacia mangium*, *Glycine max*, *Medicago truncatula*, *Vigna unguiculata*, *Phaseolus vulgaris*, and *Glycine soja* are available in miRBase (Griffiths-Jones et al., 2006), whereas miRNAs of cluster bean are yet to be fully explored.

The complimentary pairing between miRNAs and lncRNAs leads to the development of endogenous target mimics (eTMs) in the cells (German et al., 2008; Cong et al., 2013). The lncRNAs take part in complex biological phenomena, including gene transcription and translation, protein localization, cellular structure integrity and heat shock response (Fan et al., 2015). Some of the mechanisms like modulation of pri-mRNA splicing, RNA editing and abrogation of miRNA-induced repression might involve binding between long non-coding RNA and other RNA molecules. Wu et al. (2013) predicted that eTMs of several miRNAs have potential to abolish the binding between miRNAs and their targets. They found that eTM might have inhibited miRNA function in spatio-temporal manner in plant development of rice and *Arabidopsis*. Franco-Zorrilla et al. (2007) reported that AtIPS1 (induced by phosphate starvation1) eTM acts as target to miRNA ath-mir399 with one bulge of 3 nt and perturbs the cleavage effect of miRNA in *Arabidopsis* and thus regulates the uptake of phosphorous in plant cells.

The cluster bean is very popular since 18th century in Indian textile industry, in which Galactomannan (polysaccharide) is one of the key ingredients that has also been used in other industries like paper, petroleum, mining, pharmaceuticals (Hymowitz and Upadhya, 1963). Its gum has great medicinal value and used widely to control multiple diseases like diabetes, high cholesterol, diarrhea and irritable bowel syndrome (Mudgil et al., 2014). In spite of the economic significance of cluster bean, it is less explored and uncharacterized at genomic level especially the non-coding part of the genome. Moreover, the identification of lncRNAs, miRNAs and eTMs in cluster bean (*Cyamopsis tetragonoloba*) has not yet been fully characterized. Hence, there is a need to explore the role of non-coding RNAs like lncRNAs, miRNAs and eTMs in various biological processes of cluster bean.

2. Material and methods

2.1 Data processing

Initially, the leaf transcriptome of cluster bean variety M-83 was downloaded from ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR321/SRR3218523/ (Tanwar et al., 2017). Subsequently, FastQC program was run to check the quality of transcriptome data (Andrews, 2010). Low quality sequences were then removed using Trimmomatic (Bolger et al., 2014). A minimum quality score of 35 and a minimum length of 25 nt were set to improve the quality and reliability of reads. As the reference genome of cluster bean is not available, we assembled the raw reads using Trinity (Haas et al., 2013b) (Fig. 1) with default parameter, k-mer equal to 25. The contribution of each transcript was identified by mapping the reads against the assembled data via Bowtie2 aligner (Langmead and Salzberg, 2012) that is available in Trinity.

2.1. Candidate lncRNA prediction

The de novo assembled transcriptome data obtained from Trinity was used for the identification of lncRNAs (Mu et al., 2016) by following the pipeline shown in Fig. 1. High stringency measures like (i) length $>$ 200 nt (ii) Open Reading Frame (ORF) length $<$ 100 nt (iii) Coding Potential Calculator (CPC) score < -1 (iv) ***p***redictor of ***l***ong non-coding RNAs and m***e***ssenger RNAs based on an improved ***k***-mer scheme (PLEK) score < -1 were deployed to eliminate the protein coding transcripts. These four steps have been carried out in the following way:

In house *perl* scripts were written to obtain the sequence length $>$ 200 nt from the assembled data. Selected sequences were then submitted to CPC program (Kong et al., 2007) to differentiate putative coding transcripts from non-coding transcripts. The coding potential of