Conserved miRNA analysis in *Gossypium hirsutum* through small RNA sequencing

Meng-Bin Ruan\a, Ying-Tao Zhao\b,c, Zhao-Hong Meng\a, Xiu-Jie Wang\b,⁎, Wei-Cai Yang\a,⁎

\a Key Laboratory of Molecular and Developmental Biology, National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100080, China
\b State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, 100101, China
\c Graduate School of Chinese Academy of Sciences, 19 Yuquan Road, Beijing, China

**A R T I C L E  I N F O**

Article history:
Received 9 February 2009
Accepted 12 July 2009
Available online 21 July 2009

Keywords:
Gossypium hirsutum
miRNA
Identification
Sequencing

**A B S T R A C T**

Several miRNA family and their targets in cotton had been identified by computational methods based on the conserved characterization of miRNAs. So far, there are no experiments to validate the existence of miRNAs in cotton. In this study, to analyze the miRNAs in cotton, a small RNA library of sequences from 18 to 26 nt of *Gossypium hirsutum* seedling has been built by high-throughput sequencing. In this library, 34 conserved miRNA families were identified by homology search and the miRNA* sequences of them were also found in the library. Furthermore, potential targets of these conserved miRNA families were predicted in cotton TC library. However, based on the mature miRNAs and their miR* sequences, only 8 conserved miRNA encoding loci (miR156, miR157a, miR157b, miR162, miR164, miR393, miR399, miR827) were identified from cotton EST sequences. Multiple encoding loci of some miRNAs were identified by comparing the cloned miRNA and miR* sequences.

© 2009 Elsevier Inc. All rights reserved.

**Introduction**

Gene expression is regulated by an elaborate network of multiple mechanisms. miRNA, as a class of 20 to 24 nt small regulatory RNAs of gene expression in eukaryotes [1], negatively regulates gene expression at the posttranscriptional levels through RNA-induced silencing complex (RISC) by binding target miRNAs for mRNA cleavage or inhibition of mRNA translation [2]. miRNAs play crucial roles in various biological and metabolic processes in plants and animals [3–11]. Identified miRNAs are found with important regulatory functions in specific biological processes through the whole life cycle of plants [6,10,12–19]. These processes include control of tissue (leaf, root, stem, and flower) differentiation and development, phase switch from vegetative growth to reproductive growth, signal transduction, and response to different biotic and abiotic stresses (e.g. salinity, drought, and pathogens), etc.

Some known miRNA loci are found in clusters on the genome. The miRNAs in a given cluster are often related to each other. These clusters of miRNAs maybe are produced by gene duplication [20–22]. Many reported miRNAs are phylogenetically conserved. This conservation of miRNAs has been used as a powerful strategy for identification or prediction of miRNAs by homology search in other species [23]. Bioinformatically predicted miRNAs should be validated for their expression by experimental methods. Northern blotting and PCR-based amplification of adaptor-ligated cDNA had been used for validation of predicted miRNAs [24,25]. Northern blotting might not be sensitive enough to detect less-abundant miRNAs and it does not reveal the actual miRNA sequences, and PCR-based amplification can be difficult in practice when the actual mature miRNA region is unknown. The recent, dramatic improvements in the second-generation sequencing technology make it possible to acquire even low abundant small RNA sequences of a sample at significantly reduced time and cost compared with traditional approaches [26].

Cotton is an important economic crop. Following many studies involving miRNA computational prediction and identification in *Arabidopsis*, recent works using similar approach have identified many miRNAs in cotton. Qiu et al. [27] and Zhang et al. [28] reported 21 and 22 miRNA families from cSS and EST sequences of cotton by homology search. Using the identified plant miRNA precursors as orthologues to search all sequences of cotton in GenBank, 13 miRNA families were predicted [29]. Recently, 682 miRNAs were identified in 155 diverse plant species using all publicly available nucleotide databases [30]. In this research, 18 conserved miRNA families were identified in cotton. For studying the potential roles of the small RNAs in developing cotton ovules, researchers have cloned and sequenced small RNAs derived from eleven DPA periods (0–10 DPA) of cotton seedlings. After miRNA homology analysis, 34 conserved miRNA encoding loci may present in the cotton genome. For fully identification of conserved miRNAs in cotton, it is necessary to sequence all expressed small RNAs. In this paper, a total of 1,358,811 unique small RNAs have been sequenced in *Gossypium hirsutum* seedlings. After miRNA homology analysis, 34 conserved miRNA families have been identified.
miRNA families have been identified from the small RNA library. Based on the identified miRNA and miRNA* sequences, 8 miRNA (miR156, miR157a, miR157b, miR162, miR164, miR393, miR399, miR827) encoding loci have been identified from the cotton TC library. Diversity analysis of conserved miRNA and miRNA* sequences suggested that the conserved miRNA encoding loci in cotton is far more than that in Arabidopsis and rice.

**Results and discussion**

The small RNA profile

In plant, three small RNA pathways were discovered recently, including siRNA, miRNA and ta-siRNA [34]. A single miRNA can be produced by the processing of one to several longer precursors which possess stem–loop shaped secondary structures. Deep sequencing approach has been proved as an effective method for analysis of small RNAs including miRNAs in Arabidopsis and rice [32,33]. The small RNA world of cotton developing ovule has been uncovered by using the regular small RNA clone. Now, we are beginning to lift the curtain on the small RNAs especially miRNAs in cotton by high-throughput sequencing method.

A cotton small RNA library was built by high-throughput small RNA sequencing, including 3,129,095 sequences with length from 18 to 26 nt. As the Fig. 1 distribution chart showed, almost 2/3 sequences are redundant sequences. 1,189,529 (~38%) small RNAs have been sequenced only once, 169,282 small RNAs have been sequenced at least two times. Moreover, 322 small RNAs have been sequenced more than 500 times, but only 14 (4.3%) of them were respectively sequenced only once, 169,282 small RNAs have been sequenced at least two times. Moreover, 322 small RNAs have been sequenced more than 500 times, but only 14 (4.3%) of them were respectively sequenced only once (Fig. 1).

For sequence analysis of the small RNA library, 144 BAC and 21,405 TC of cotton were downloaded from GenBank and TIGR. The unique sequences of the small RNA library were mapped to the sequences of these two dataset. As Table 1 showed, 1003 unique sequences were mapped to cotton BACs, and 30,521 unique sequences were located in cotton TCs. These unique sequences that can be located in BAC and TC sequences only comprise of 2.22% of total unique sequences identified in the library. Also, the fragments of repeat regions, tRNAs, rRNAs, protein coding regions, and snoRNAs were analyzed in the library, distribution of these fragments was listed in Table 2.

### Table 1

<table>
<thead>
<tr>
<th>Library</th>
<th>Mapping loci</th>
<th>Unique sRNA</th>
<th>Percent of unique sRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC (144)</td>
<td>2152</td>
<td>1003</td>
<td>0.07</td>
</tr>
<tr>
<td>TC (21,405)</td>
<td>51,051</td>
<td>30,521</td>
<td>2.25</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Small RNA</th>
<th>Unique reads</th>
<th>Total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>1,358,811</td>
<td>3,129,095</td>
</tr>
<tr>
<td>Repeat region</td>
<td>614</td>
<td>10,899</td>
</tr>
<tr>
<td>rRNA</td>
<td>30,152</td>
<td>374,997</td>
</tr>
<tr>
<td>tRNA</td>
<td>573</td>
<td>22,310</td>
</tr>
<tr>
<td>Protein coding</td>
<td>407</td>
<td>687</td>
</tr>
<tr>
<td>snoRNA</td>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td>miRNA</td>
<td>892</td>
<td>135,298</td>
</tr>
</tbody>
</table>

### Identification of conserved miRNAs

For identification of conserved miRNAs in cotton, unique small RNA sequences had been aligned with conserved plant miRNAs in the miRBase (release 12.0) with 0–3 bases of mismatch. 892 unique sequences were identified as homologs of 33 known plant miRNAs. To be annotated as an miRNA, sequences representing both miRNA and miRNA* should be identified for the miRNA candidates [35]. MiRNA* sequences of most of these conserved miRNA families have been identified from our cloned small RNA library by homology search. Target prediction of the conserved miRNA families was performed by mirU using nearly perfect sequence complementary as criteria [36]. The identified conserved miRNA and their predicted targets were listed in Table 3. Most conserved miRNAs have different sequences, which were shown in Fig. 2. In this result, 13 conserved miRNA families (miR156/miR157, miR159, miR164, miR165, miR166, miR167, miR172, miR399, miR827, miR444a.2, miR444b.1, miR414, miR390), 892 unique miRNA and miRNA* sequences, 8 miRNA (miR156, miR157a, miR157b, miR162, miR164, miR393, miR399, miR827) encoding loci have been identified from the cotton TC library. Diversity analysis of conserved miRNA and miRNA* sequences suggested that the conserved miRNA encoding loci in cotton is far more than that in Arabidopsis and rice.

#### Table 3

<table>
<thead>
<tr>
<th>miRNA family</th>
<th>Reads*</th>
<th>Precursorb</th>
<th>miRNA* sequence</th>
<th>Predicted target family</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR156/miR157</td>
<td>22,560</td>
<td>Y</td>
<td>Y</td>
<td>Squamosa-promoter binding protein</td>
</tr>
<tr>
<td>miR159</td>
<td>178</td>
<td>N</td>
<td>Y</td>
<td>ATM synthase</td>
</tr>
<tr>
<td>miR160</td>
<td>67</td>
<td>N</td>
<td>Y</td>
<td>Auxin response factor</td>
</tr>
<tr>
<td>miR162</td>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>Zinc finger protein</td>
</tr>
<tr>
<td>miR164</td>
<td>297</td>
<td>Y</td>
<td>N</td>
<td>NAC domain protein NAC1</td>
</tr>
<tr>
<td>miR165</td>
<td>49</td>
<td>N</td>
<td>Y</td>
<td>Class III HD-Zip protein</td>
</tr>
<tr>
<td>miR166</td>
<td>1715</td>
<td>N</td>
<td>Y</td>
<td>Class III HD-Zip protein</td>
</tr>
<tr>
<td>miR167</td>
<td>6030</td>
<td>N</td>
<td>Y</td>
<td>LIM domain protein</td>
</tr>
<tr>
<td>miR168</td>
<td>11,627</td>
<td>N</td>
<td>Y</td>
<td>AG01-1</td>
</tr>
<tr>
<td>miR169</td>
<td>362</td>
<td>N</td>
<td>Y</td>
<td>Unknown</td>
</tr>
<tr>
<td>miR170/miR171</td>
<td>77</td>
<td>N</td>
<td>Y</td>
<td>GRAS transcription factor</td>
</tr>
<tr>
<td>miR172</td>
<td>843</td>
<td>N</td>
<td>Y</td>
<td>AP2 related transcription factor</td>
</tr>
<tr>
<td>miR319</td>
<td>26</td>
<td>N</td>
<td>Y</td>
<td>ATP synthase</td>
</tr>
<tr>
<td>miR390</td>
<td>19</td>
<td>N</td>
<td>N</td>
<td>Unknown</td>
</tr>
<tr>
<td>miR393</td>
<td>601</td>
<td>Y</td>
<td>Y</td>
<td>Transport inhibitor response 1</td>
</tr>
<tr>
<td>miR394</td>
<td>11</td>
<td>N</td>
<td>N</td>
<td>Unknown</td>
</tr>
<tr>
<td>miR395</td>
<td>13</td>
<td>N</td>
<td>Y</td>
<td>ATP sulfurylase</td>
</tr>
<tr>
<td>miR396</td>
<td>268</td>
<td>N</td>
<td>Y</td>
<td>Unknown</td>
</tr>
<tr>
<td>miR397</td>
<td>115</td>
<td>N</td>
<td>Y</td>
<td>Laccase, diaphenol oxidase,</td>
</tr>
<tr>
<td>miR444a.2</td>
<td>130</td>
<td>N</td>
<td>Y</td>
<td>Heat stress transcription</td>
</tr>
<tr>
<td>miR444b.1</td>
<td>113</td>
<td>N</td>
<td>Y</td>
<td>Choline monoxygenase</td>
</tr>
<tr>
<td>Osa-miR528</td>
<td>122</td>
<td>N</td>
<td>N</td>
<td>Pathogen-related protein</td>
</tr>
<tr>
<td>miR827</td>
<td>42</td>
<td>Y</td>
<td>Y</td>
<td>Exp1 protein</td>
</tr>
<tr>
<td>miR860</td>
<td>1</td>
<td>N</td>
<td>N</td>
<td>Unknown</td>
</tr>
<tr>
<td>ppt-miR894</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>Unknown</td>
</tr>
<tr>
<td>miR845</td>
<td>5</td>
<td>N</td>
<td>Y</td>
<td>Histone H2B</td>
</tr>
<tr>
<td>Osa-miR1318</td>
<td>3</td>
<td>N</td>
<td>Y</td>
<td>RACK1-like protein</td>
</tr>
<tr>
<td>Osa-miR1436</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>Triacylglycerol lipase</td>
</tr>
</tbody>
</table>

* Reads per million in the small library.

b Precursor located in cotton EST sequences.
miR168, miR169, miR172, miR319, miR393, miR399) have more than 40 different small RNA sequences. The mature and miRNA* sequences of 8 miRNAs can be mapped to cotton TCs. The predicted second structure of these miRNAs are shown in Fig. 3.

Since the *G. hirsutum* full genome sequence is not available yet, it is unfeasible to determine the precise genomic loci of each miRNA family. According to the numbers of different sequences homologous to known miRNAs identified in our library, it is likely that

---

**Fig. 2.** Small RNA number distribution of conserved miRNAs in cotton.

**Fig. 3.** Hairpin secondary structures of conserved miRNA precursors, mature miRNA positions were highlighted in red. The secondary structures were produced by mfold (V3.2) at URL: [http://www.bioinfo.rpi.edu/applications/mfold/cgi-bin/mrna-form1.cgi](http://www.bioinfo.rpi.edu/applications/mfold/cgi-bin/mrna-form1.cgi). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
the *G. hirsutum* genome contains more miRNA encoding loci than *Arabidopsis* and rice. In the *Arabidopsis* plant, whose genome has been fully sequenced, 191 miRNA encoding loci have been identified in the *Arabidopsis* MPSS plus database [33]. And there are 372 miRNA encoding loci that have been identified in rice [32]. Comparing these two libraries suggested that the bigger a genome, the more miRNA encoding loci. Therefore, *G. hirsutum*, a tetraploid (2n = 4x = 52) plant, whose genome is bigger than that of *Arabidopsis* and rice, should have more miRNA encoding loci than that of *Arabidopsis* and rice. As the MPSS plus database showed, the miR156/miR157 family has 12 encoding loci in both *Arabidopsis* and rice. We analyzed the sequences that have been identified as miRNA and miRNA* sequences of miR156/miR157. As shown in Fig. 4, many miRNA and miRNA* sequences of miR156/157 have been identified in our sequencing results. Although some of these sequencing variants might be generated as a result of RNA editing or SNP, it is also possible that the *G. hirsutum* genome contains more miR156/miR157 loci than that of *Arabidopsis* and rice.

Bioinformatics prediction of the conservation of the conserved miRNA families in cotton may not be completely informative at this time because of the lack of complete genome information and the search for these miRNA precursor sequences in TC library has been unsuccessful (only 8 precursors were identified). Recent deep sequencing of *Arabidopsis* and rice small RNAs suggested that the
plant genome encodes more non-conserved miRNA families than
conserved miRNA families [32,33]. Although we have predicted
several potential precursors of non-conserved miRNA from cotton
TC library, these non-conserved miRNAs have been sequenced only
one or two times in the small library, but the miRNA* sequences of
these non-conserved miRNAs were failed to be identified in the library.

miRNA and miRNA* sequence diversity analysis

To analyze the diversity of miRNA and miRNA* sequences that
have been identified from our small RNA library, multiple alignments
of several conserved miRNA families (miR156/miR157, miR167,
miR168, miR169, miR393) were performed. As Fig. 4 showed (only
the 195 sequence that has been read at least 3 times is listed), among
the sequences identified as miR156/miR157, five nucleotide variable
positions were found, and 9 nucleotide variable positions were found
among the sequences of miR393 (Fig. 5). Sequence variation were also
found in the miRNA* sequences, mostly at positions different from
variations observed in the mature miRNA sequences.

Some known miRNAs are found in clusters, they are transcribed as
polygenic primary transcripts and miRNAs in a cluster may be func-
tionally related [37]. One evolutionary forces to produce miRNA clusters
is gene duplication. The genus Gossypium occurs naturally throughout
tropical and subtropical regions, and includes about 45 species split
across two ploidy levels, diploid \((2n = 2x = 26)\) and tetraploid \((2n =
4x = 52)\) [38]. G. hirsutum is a tetraploid \((2n = 4x = 52)\), so the gene
duplications in it might be more complex than other plants. Due to the
lack of complete genome information, most of our sequence reads can
not be mapped to the genomes. Therefore, no miRNA gene cluster was
identified in this study.

In summary, we have built a small RNA library of G. hirsutum
seeding by high-throughput sequencing technology. Conserved
miRNA and miRNA* sequences have been identified from the library.
Along with more and more information about cotton genome
becoming available, the miRNA world of cotton will be uncovered
little by little. With the availability of cotton BAC and genome

sequences in the future, we expect more novel cotton miRNAs being
identified in the future.

Materials and methods

Plant materials and total RNA extraction

Upland cotton (G. hirsutum) C312 seeds were grown on 1/2 MS
media and 1/2 MS media containing 200 mM NaCl. Total RNA of
6-day seedlings of two treatments was isolated by cold-acidic
phenol method with a modified extraction buffer (50 mM Tris–HCl
pH 6.0, 10 mM EDTA pH 8.0, 100 mM LiCl, 2% SDS, 2% PVP). Then
RNA was precipitated by ethanol, dissolved in DEPC water and
stored at \(-80 °C\).

Small RNA isolation and sequencing

The extracted total RNAs were resolved on a denatured 15%
polyacrylamide gel. Gel fragments spanning the size range of 18 to
26 nt were excised, and small RNAs were eluted overnight with 0.5 M
NaCl at 4 °C, precipitated by ethanol. The small RNAs were sequenced
by SOLEXA sequencing method (BGI, Beijing).

Small RNA library analysis

144 BAC from GenBank and 21,405 TC from TIGR of cotton were
downloaded. Sequences from 18 to 26 nt in the sequence result were
mapped to the cotton BAC and TC library.

Conserved miRNA identification and target prediction

For analysis of conserved miRNAs in cotton, unique small RNAs
were aligned with the plant mature miRNAs from miRBase (release
12.0) [39]. Cotton small RNA sequences with 0-3 bases mismatch were
identified as known plant miRNAs. miRNA* sequences of conserved
miRNAs were identified by using the miRNA* sequences of Arabidopsis
and rice as homology. Conserved miRNAs were mapped to cotton TC

![Fig. 5. Multiple alignments of miRNA and their miRNA* sequences of miR393. The sequence Logo was produced by WebLogo at URL: http://weblogo.berkeley.edu/logo.cgi.](image-url)
sequences from TIGR. Sequences of 120 bp long around the mapping loci were processed by mfold (V3.2) [40] at the URL: http://www.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi to find whether they could be folded into a hairpin secondary structures. Targets of conserved miRNA were predicted by the web tool psRNA-Target [36] at the URL: http://bioinfo3.noble.org/psRNATarget/ using the Gossypium (cotton) DFCI Gene index (CGI) release 9 as the sequence library for target search, set 3 as the maximum expectation.

Conserved miRNA sequence multiple alignments

Conserved miRNA multiple alignments were performed using the WebLogo at URL: http://weblogo.berkeley.edu/logo.cgi.

Acknowledgments

The author gratefully acknowledges the support of the Chinese National 863 Plans Projects Foundation (grant no. 2006AA10A108), the support of K.C. Wong Education Foundation, Hong Kong, and the support of the Knowledge Innovation Program of the Chinese Academy of Sciences (grant No. KYQY-Y011).

Appendix A. Supplementary data


References