Computational Characterization of Conserved and Novel miRNAs in Cotton

Presented by Vikram Agarwal

In partial fulfillment of the requirements for graduation with the
Dean's Scholars Honor's Degree in Biology

_____	_____
Dr. Jeffrey Chen	Date
Supervising Professor	

TABLE OF CONTENTS

ABSTRACT

Cotton is one of the world's most valuable textiles and a significant oilseed crop. The fiber cells that emerge from the cotton ovule are specified from undifferentiated epidermal tissue and undergo a process of initiation, elongation, secondary cell wall synthesis, and maturation. Because of their unique development features, cotton fibers serve as ideal biological models for investigating the mechanisms of cellular elongation and cellulose biosynthesis. Despite their importance, little is known about how genes are regulated in the initial stages of fiber growth in order to coordinate their development. microRNAs (miRNAs) are small, ~21-nucleotide RNAs that are known to have broad post-transcriptional impact on gene expression and development. In this study, we sequenced nearly 4.1 million small RNAs derived from the early stages of fiber growth in an effort to characterize the diversity of cotton miRNAs and to identify their physiological roles. Comparative sequence analysis identified 26 conserved miRNA families as well as 5 novel, cotton-specific miRNAs. A Perl script was written that systematically predicted 242 miRNA targets in the cotton transcriptome. Nearly all of the targets that were experimentally tested for cleavage were experimentally validated, suggesting a low false-positive rate in the predictions. The results reveal that miRNAs often target families of transcription factors that mediate critical signaling pathways such as those involved in the phytohormonal response to auxin.

INTRODUCTION

Cotton is a cash crop of tremendous agricultural value, both as a source for seed oil and in the textile industry. Currently, cotton is estimated to have an economic impact of nearly $500 billion per year, providing a source of revenue for as many as eighty countries worldwide. The demand for cotton is expected to rise from the approximate 115 million bales that are consumed annually (Chen *et al.* 2007).

Besides their economic importance, cotton fibers provide an ideal biological system to study common developmental pathways in plants. This is largely due to their unique and exaggerated development features. On the day of anthesis, or flowering, fiber cells begin to emerge from the epidermal cell layer of ovule tissue. Over the next forty to sixty days, the fibers develop through a series of four stages: fiber initiation, cellular elongation, secondary cell wall synthesis, and maturation. During initation, the single-celled fibers rapidly differentiate from unspecified progenitor cells, making them a model for studying cell fate specification. Because of their exceptional length, fibers can also be used to investigate the molecular mechanisms that contribute to cellular elongation. Moreover, mature cotton fibers consist almost entirely of cellulose, a structural polymer that can be converted into cellulosic ethanol, a naturally-renewable biofuel and potential source of alternative energy. The extensive deposition of this cellulose in the cell wall of fibers make them an important system for studying the metabolic events that guide cellulose biosynthesis.

Despite the benefits of fibers as model systems, little is known about how genes are regulated in the initial stages of fiber growth in order to specify their cell fate and to mediate the extreme developmental changes that they undergo. In recent years, large-scale sequencing studies have begun to unravel the transcriptional landscape of fiber cells, giving insight into the

tissue-specific expression patterns of messenger RNA (mRNA). These studies reveal that fiber growth is likely under the strict regulation of phytohormonal signals such as auxins and giberellins, as well as important transcription factors such as *MYB* and *WRKY* family proteins (Yang *et al.* 2006).

Recently, microRNAs (miRNAs) have been demonstrated to regulate a variety of transcription factors and phytohormonal regulators that mediate critical morphogenetic events. miRNAs are small (~21 nucleotide RNA) regulatory molecules that are known to have broad post-transcriptional impact on gene expression and development. In plants, precursor miRNA loci are transcribed by RNA polymerase II into primary miRNA transcripts (pri-miRNAs), and subsequently processed into a mature ~21-base pair miRNA fragments by *DICER-LIKE 1* (*DCL1*). The mature miRNAs are exported from the nucleus into the cytoplasm, where they are incorporated into the *ARGONAUTE1* (*AGO1*) silencing complex. Tethered to the protein, the mature miRNA strand serves as a template to guide the base pairing of complementary or near-complementary mRNA targets. These targets, which often encode important transcription factors, are translationally repressed or endonucleolytically cleaved by an RNAse domain in *AGO1* (Jones-Rhoades *et al.* 2006, Dugas and Bartel 2004). This cleavage results in the silencing or repression of transcripts, preventing their translation into a functional protein.

Recently, high-throughput small RNA sequencing technologies have opened doorways into the semi-quantitative analysis of the spatial and temporal expression patterns of miRNAs. These studies have discovered and confirmed on the order of hundreds of miRNA families in the genome of *Arabidopsis thaliana*, the model organism in flowering plant biology. Furthermore, statistical studies have given researchers the ability to discern biological parameters involved in miRNA targeting in plants, such as increased base pairing in a seed region near the miRNA 5'

end, as well as the conventional cleavage of the target between nucleotides 10 and 11 relative to the miRNA 5' end (Jones-Rhoades and Bartel 2004, Allen *et al.* 2005).

The discovery of these parameters have greatly facilitated the prediction of hundreds of mRNA targets in the plant transcriptome. Studies indicate that plant miRNAs often target members of protein families that are functional in common pathways (Rajagopalan *et al.* 2006). Since large-scale small RNA sequencing studies have been undertaken in only a small number of plant species, the evolution of miRNA families – their conservation and divergence across species – becomes difficult to evaluate. It remains unclear whether conserved miRNA families have maintained the same regulatory relationships with their targets, or whether they have evolved new biological roles. Finally, the extent to which novel miRNA families contribute to species-specific developmental processes remains unexplored.

In this study, we sequenced nearly 4.1 million small RNAs derived from the early stages of fiber growth in an effort to characterize the diversity of cotton miRNAs and to identify their physiological roles. Comparative sequence analyses identified 26 conserved miRNA families in cotton, as well as 5 miRNA precursor loci that comprise 3 novel families. A Perl script was written that predicted 242 miRNA targets in the cotton transcriptome. Nearly all of the targets that were experimentally tested for cleavage were validated. The results reveal that most conserved miRNAs target homologs of known *Arabidopsis* miRNA targets, including families of transcription factors that mediate critical developmental responses. However, several conserved miRNA families have potentially evolved novel cotton-specific targets, while at least one novel cotton miRNA may influence a phytohormonal response pathway involved in gibberellin regulation.

METHODS

Cotton small RNA Sequencing

Tissue samples were collected from wild-type cotton (*Gossypium hirsutum* L.) three days prior to anthesis (-3 DPA), on the day of anthesis (0 DPA), three days post anthesis (+3 DPA), and from leaf tissue. Cotton total RNA was extracted from these tissues, and small RNAs 17-27 nucleotides in size were purified in a 15% polyacrylamide gel and ligated to the 3' and 5' RNA adaptors. The ligated RNAs from each sample were then reverse-transcribed using primer pairs partially corresponding to the RNA adaptors, and subsequently amplified. The four "barcoded" RNA samples were then pooled in equal amounts. After a series of quality control tests, the pooled cDNA library was sequenced using a high-throughput Illumina/Solexa G1 sequencing machine. This generated a dataset of approximately ~4.1 million sequenced reads.

Bioinformatic Analysis of Conserved miRNAs

After grouping raw Solexa reads by their indexing base and removing adaptor sequences from the raw reads, contaminant sequences were removed. Contaminant RNA was annotated based on homology to: 1) the *A. thaliana* and *B. napus* mitochondrial genomes, 2) the *G. hirsutum* chloroplast genome, 3) plant snRNAs and snoRNAs deposited in NCBI, 4) *A. thaliana* transfer RNA sequences, and 5) annotated *A. thaliana* and *G. hirsutum* ribosomal RNA. Reads with and without contaminant sequences were then mapped onto the TIGR Cotton Gene Index 9 (CGI9) Expressed Sequence Tag (EST) assembly and onto trace reads from the partial *G. raimondii* genome (Table 1). Because no genome assembly is currently available for *G. hirsutum*, it is likely that contaminants and reads containing sequencing error remain. Conserved mature miRNAs were detected based on homology to known miRNAs deposited in miRBase 13.0 (http://microrna.sanger.ac.uk/).

Computational miRNA Target Prediction

After contaminant sequences were removed, miRNA abundances were normalized to transcripts per quarter million (TPQ). The most abundant miRNA variant was used as a query to search for targets, which were predicted as described (Allen *et al.* 2005; Table 2). To annotate the ESTs, a BlastX search was performed against the NCBI non-redundant protein sequence database of flowering plants (taxid:3398), using the top three hits for each query (Altschul *et al.* 1997). Because the ESTs can exist in sense or antisense orientation, candidate targets with an ORF with the same 5'->3' directionality as that of the miRNA were removed. A few partial ESTs derived from miRNA precursors, or sequences with ambiguous orientation, may remain as predicted targets.

Experimental Validation of Predicted miRNA Targets

RNA-ligation mediated (RLM) rapid amplification of 5' complementary DNA ends (5' RACE) was employed to map the cleavage sites of target transcripts, using the GeneRacer kit (Invitrogen) that was modified from a published protocol (Llave *et al.* 2002). Total RNA (4 µg) from equal mixtures of ovules (-3, 0, +3 DPA) and fibers (+10 DPA) were ligated to 5' RACE RNA adapter without calf intestine alkaline phosphatase treatment. cDNAs were transcribed with reverse transcriptase and the GeneRacer Oligo dT primer. Non-gene-specific 5′ RACE products were generated using the GeneRacer 5′ Primer (5′-CGACTGGAGCACGAGGACACTGA-3′) and the GeneRacer 3′ Primer (5′-GCTGTCAACGATACGCTACGTAACG-3′). Gene-specific 5′ RACE amplifications were conducted with the GeneRacer 5′ Nested Primer and gene-specific primers. The 5′ RACE amplification products were gel purified and cloned, and 15-50 inserts were sequenced from a typical reaction.

Identification of Precursors for Conserved and Novel miRNAs

CGI9 was searched for precursor miRNAs using mirCheck (Jones-Rhoades and Bartel 2004, Rajagopalan *et al.* 2006, Axtell *et al.* 2007) using a 600-nucleotide window centered on a cotton sRNA sequence (Table 3). Known miRNA families were annotated based upon conservation of hairpin-loop structures as well as mature miRNA sequences, while novel miRNA precursors were accepted only if a miRNA* sequence with a lower abundance was detected on the opposite strand of the hairpin, consistent with the revised annotation criterion for plant miRNAs (Meyers *et al.* 2008).

RESULTS

After processing the small RNAs and categorizing them by their indexing base, it was found that a total of 4,114,570 reads were sequenced, with a greater representation of reads in Leaf and +3 DPA tissue. The size class distribution of the reads was then plotted as a frequency histrogram (Figure 1). It was found that the 21-base pair (bp) small RNAs, the expected size for miRNAs, had a low abundance relative to small RNAs belonging to other size classes. Most noticeably, the 24-bp size class was highly represented in the dataset, with more than 50% of reads belonging to this class in each of the four tissues. Interestingly, the abundance of this size class seemed to increase temporally through the ovule stages, culminating with a representaiton of as much as 75-80% of reads in the 0 DPA and +3 DPA stages (Figure 1).

After mapping the small RNAs, it was found that a large fraction of them mapped to rRNA sequences, especially in the Leaf and -3 DPA tissues (Table 1). It was further found that a greater percentage of reads mapped to the *Gossypium raimondii* (GRa) genome relative to CGI9; this disparity between the mappings was greatest with reads in the 0 DPA and +3 DPA stages

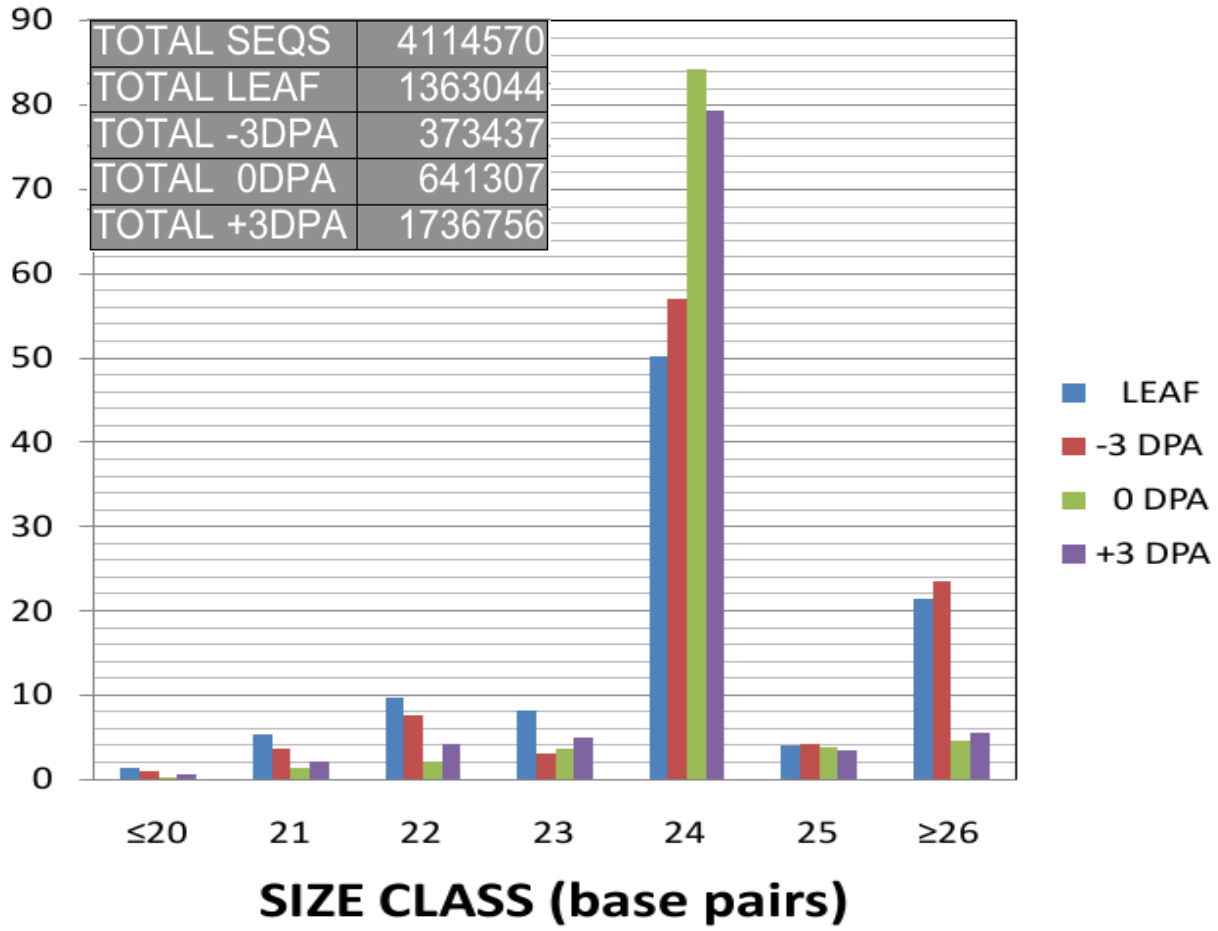

Figure 1. Statistics of small RNA sequence reads. Percentages plotted represent reads belonging to a size class normalized to the total number of reads in the tissue after the removal of contaminant sequences.

specifically, indicating that the reads likely originated from intergenic regions rather than coding ones (Table 1). When searching for homology to known miRNAs deposited in miRBase 13.0, it was found that a greater percentage of miRNAs matched in the -3 DPA stage, with an accumulation of only 50-60% of this in the other three tissues. This pattern was preserved when considering only distinct reads, suggesting that the greater representation of reads homologous to known miRNAs was likely due to a diverse set of microRNAs rather than a select few (Table 1).

In total, 26 conserved miRNA families were detected in the reads (miR156 to miR894, Table 2). A majority of their predicted targets were homologs of known targets in *Arabidopsis*.

	All reads (%)					Distinct reads (%)				
Library	**Leaf**	**-3 DPA**	**0 DPA**	**+3 DPA**	**Total**	**Leaf**	**-3 DPA**	**0 DPA**	**+3 DPA**	**Total**
Contaminants	57.88	39.10	8.21	9.39	27.96	17.87	14.92	3.55	3.03	5.95
rRNA	53.83	29.56	6.37	7.03	24.47	16.25	11.42	2.82	2.20	4.88
tRNA	3.09	7.97	1.48	2.00	2.82	0.90	2.54	0.45	0.58	0.68
snoRNA	0.02	0.31	0.01	0.02	0.05	0.03	0.14	0.01	0.02	0.02
snRNA	0.02	0.04	0.01	0.01	0.01	0.03	0.06	0.02	0.01	0.02
mitochondria	0.02	0.17	0.04	0.04	0.04	0.05	0.27	0.05	0.04	0.06
chloroplast	0.91	1.05	0.29	0.30	0.57	0.62	0.49	0.20	0.18	0.29
CGI9 (EST assembly)	51.60	31.83	9.48	10.58	25.92	13.91	12.12	4.52	3.97	5.54
GRa (Genomic reads)	61.25	42.67	24.99	24.57	38.42	26.83	23.47	19.26	17.56	18.66
Matching miRNAs	0.80	1.82	0.82	1.19	1.06	0.09	0.13	0.05	0.06	0.05
Total raw reads	1359250	372521	639801	1732919	4104491	526304	191591	505504	1180742	2169534

Table 1. Statistics of small RNA sequence reads. Categories are described as percentages of Total Raw Reads.

For example, miR156 was predicted to target the Squamosa Promoter binding factors; miR160, miR167, miR390, and miR393 were predicted to target the Auxin Response Factors (ARF) or other components of the auxin response pathway; and miR399 and miR828 were implicated in targeting the *MYB* family of transcription factors (Table 2). Although the majority targeted homologs of known *Arabidopsis* targets, there were several notable examples which were predicted as novel interactions in cotton. These include miR159 and Beta-ketoacyl-CoA synthase, miR162 and Alcohol dehydrogenase, and miR482-5p and sucrose synthase. Finally, one of the three novel miRNAs, referred to here as miR-nov3, was predicted to target gibberellin 3-hydroxylase. In total, 242 targets were predicted for the conserved miRNAs, many of which were likely redundant targets (unassembled, from different species of cotton, or from homeologous loci) deposited in CGI9. The normalized abundance of the miRNAs was such that the vast majority of miRNAs were lowly abundant in 0 DPA and +3 DPA relative to -3 DPA and Leaf tissues. The levels of several families, however, increased in ovule tissues, including miR156/157, miR167, miR168, and miR172 (Table 2).

To confirm whether these predictions were accurate, we experimentally tested whether the target mRNA strands were indeed cleaved at the predicted positions. In total, 12 targets were tested by 5' RACE. Of these, 10 were determined to be accurate prections, indicating a relatively low false positive rate of about 16.67% (data not shown). Nearly all of the targets were cleaved at the canonical cleavage position between nucleotides 10 and 11 relative to the miRNA 5' end (Pang *et al*., unpublished data). This evidence indicates that several miRNA-target interactions are indeed conserved between *Arabidopsis* and cotton, including those between miR165/166 and the HD ZIP III transription factors, miR172 and APETALA2, a transcription factor involved in floral development, and miR390 and TAS3, a trans-acting small interfering RNA (ta-siRNA)

Sequence (5'->3') *	miRNA	Total**	Leaf	-3 DPA	0 DPA	+3 DPA	No. targets	Target gene family description
UUGACAGAAGAUAGAGAGCAC	**156/157**	13.5	14.5	2.7	4.3	18.5	20	Squamosa promoter-binding factors, Ser/Thr protein phosphatase
UUUGGAUUGAAGGGAGCUCUA	**159**	7.9	7.4	12.8	6.6	7.8	4	Beta-ketoacyl-CoA synthase
UGCCUGGCUCCCUGUAUGCCA	**160**	4.0	11.4	0.0	1.2	0.0	5	Auxin response factor (ARF) family
UCGAUAAACCUCUGCAUCCAG	**162**	0.4	0.9	0.0	0.0	0.3	4	Allyl alcohol dehydrogenase
UGGAGAAGCAGGGCACGUGCA	**164**	5.7	9.6	0.7	3.1	4.6	2	NAC domain transcription factors
UCGGACCAGGCUUCAUUCCCC	**165/166**	2341.6	1709.4	4571.6	1769.7	2569.4	10	Class III HD-Zip proteins
UGAAGCUGCCAGCAUGAUCUCA	**167**	167.8	45.1	13.4	161.8	299.5	7	Auxin response factor (ARF) family, glycoprotease
UGCUUGGUGCAGAUCGGGAC	**168**	104.4	30.7	3.4	7.0	219.9	4	Argonaute 1, F-box proteins
CAGCCAAGGAUGACUUGCCGG	**169**	0.6	1.8	0.0	0.0	0.0	11	Heme activating protein (HAP2), CCAAT-binding transcription factors
UGAUUGAGCCGUGCCAAUAUC	**170/171**	19.2	55.7	2.0	0.4	1.3	8	Hairy meristem/Scarecrow-like 6 transcription factors
AGAAUCUUGAUGAUGCUGCAU	**172**	37.0	81.7	9.4	7.8	18.6	21	APETALA2, AHAP2-like factors, Target of EAT1 (TOE1)
UGGACUGAAGGGAGCUCCCUC	**319**	0.1	0.2	0.0	0.0	0.0	7	TCP family transcription factors
AAGCUCAGGAGGGAUAGCGCC	**390**	5.8	6.4	0.0	10.2	5.0	11	TAS3, leucine-rich repeat transmembrane protein kinase
UCCAAAGGGAUCGCAUUGAUUU	**393**	2.3	0.4	0.0	12.1	0.6	11	Transport Inhibitor Response 1 (TIR-1)
UUCCACAGCUUUCUUGAACUG	**396**	1.0	2.6	0.0	0.8	0.1	31	ATCHR12 transcriptional regulator, Growth regulating factors (GRF)
UCAUUGAGUGCAGCGUUGAUG	**397**	0.1	0.2	0.0	0.0	0.0	13	Laccase/Copper ion binding proteins, diphenol oxidase
UGCCAAAGGAGAUUUGCCCGG	**399**	0.8	2.0	0.0	0.0	0.3	2	MYB family transcription factor, TIR-1
AUGCACUGCCUCUUCCCUGGC	**408**	0.1	0.2	0.0	0.0	0.0	10	Blue copper proteins, uclacyanin 3
UGUGGGAGAGUUGGGCAAGAAU	**482-5p**	1.0	1.7	0.0	0.4	0.9	5	Sucrose synthase, Glucose-methanol-choline (GMC) oxidoreductase
UCUUGCCUACUCCACCCAUGCC	**472/482**	4.4	13.2	0.0	0.0	0.1	9	NBS-type resistance protein
UGCAUUUGCACCUGCACCUUC	**530**	1.4	4.0	0.0	0.0	0.1	5	C2H2 transcription factors, bHLH family protein
UGACAACGAGAGAGAGCACGU	**535**	8.6	19.3	0.7	1.2	4.6	4	Squamosa promoter-binding factors
UUAGAUGACCAUCAACAAACA	**827**	0.2	0.7	0.0	0.0	0.0	1	Unknown
UCUUGCUCAAAUGAGUAUUCUA	**828**	0.1	0.2	0.0	0.0	0.0	7	MYB family transcription factors
GUUUCACGUCGGGUUCACCA	**894**	19.3	16.4	33.6	29.7	14.7	4	Responsive to Dessication 20
UAUACCGUGCCCAUGACUGUAG	**nov1**	8.1	19.5	0.0	1.6	3.3	1	Serine/threonine protein phosphatase
ACUUUUGAACUGGAUUUGCCGA	**nov2**	4.1	3.5	0.0	6.3	4.6	6	Endosomal protein
UGGUGUGCAGGGGGUGGAAUA	**nov3**	0.6	1.7	0.0	0.0	0.1	5	Gibberellin 3-hydroxylase/anthocyanidin synthase

*Most abundant variant shown
**Abundance is normalized to Transcripts per Quarter Million (TPQ) and rounded to nearest tenth

Table 2. MicroRNAs detected by sequencing, their normalized tissue specific expression profiles, and predicted target gene families in cotton.

gene involved in a cascade that silences components of the auxin response (Pang *et al*., unpublished data).

After searching for precursor miRNAs in CGI9, 32 precursor loci were found, including 5 novel precursors (Table 3). Because CGI9 is composed of ESTs from several species,

miRNA	miRNA sequence (5'->3')	CGI9 ID	miR Pos	Orientation	Species*	Mature Strand	miR:miRStar
157a	UUGACAGAAGAUAGAGAGCAC	TC108116	963	antisense	Gr	5prime	210:0
157b	UUGACAGAAGAGAGUGAGCAC	TC120651	1151	antisense	Gr	5prime	99:0
156	UUGACAGAAGAGAGUGAGCAC	TC124575	308	antisense	Gh	5prime	54:4
162a	UCGAUAAACCUCUGCAUCCAG	TC101648	275	sense	Gh	3prime	6:1
164	UGGAGAAGCAGGGCACGUGCA	DR461140	297	antisense	Gh	5prime	77:0
166b	UCGGACCAGGCUUCAUUCCCC	DW502146	309	sense	Gh	3prime	17514:2
167	UGAAGCUGCCAGCAUGAUCUA	**	201	sense	Gh	5prime	27:0
171	UGAUUGAGCCGUGCCAAUAUC	DW507416	190	sense	Gh	3prime	269:0
172	AGAAUCCUGAUGAUGCUGCAG	TC124126	642	antisense	Gh	3prime	90:1
390b	AAGCUCAGGAGGGAUAGCGCC	TC87445	76	sense	Gh	5prime	47:0
390a	AAGCUCAGGAGGGAUAGCGCC	DW238152	71	sense	Gh	5prime	47:0
393	UCCAAAGGGAUCGCAUUGAUCU	TC118931	84	sense	Gh	5prime	7:0
394	UUGGCAUUCUGUCCACCUCC	DW517361	113	sense	Gh	5prime	0:0
394	UUGGCAUUCUGUCCACCUCC	ES802173	457	sense	Gh	5prime	0:0
396a	UUCCACAGCUUUCUUGAACUG	TC124879	195	sense	Gh	5prime	12:0
398	UGUGUUCUCAGGUCACCCCUU	CO070687	112	sense	Gr	3prime	0:0
398	UGUGUUCUCAGGUCACCCCUU	DW498056	93	sense	Gh	3prime	0:0
399	UGCCAAAGGAGAGUUGGCCUU	DW510913	56	sense	Gh	3prime	0:0
399	UGCCAAAGGAGAUUUGCCCUG	DW509341	97	sense	Gh	3prime	2:0
479	CGUGAUAUUGGUUCGGCUCAUC	ES809290	42	sense	Gh	5prime	0:0
482-5'	UGUGGGAGAGUUGGGCAAGAAU	DW517596	175	sense	Gh	3prime	24:1
482-3'	UCUUUCCAAUUCCUCCCAUUCC	TC106817	1126	antisense	Gr	3prime	1:1
482-3'	UCUUUCCUACUCCUCCCAUACC	DR457519	248	sense	Gh	3prime	3:1
482-3'	UCUUGCCUACUCCACCCAUGCC	DT527030	137	sense	Gh	3prime	40:5
827	UUAGAUGACCAUCAACAAACA	TC100384	472	antisense	Gh	3prime	3:1
827	UUAGAUGACCAUCAACAAACA	TC107389	180	sense	Gh	3prime	3:1
827	UUAGAUGACCAUCAACAAACA	EX166072	187	sense	Gh	3prime	3:1
novel1	UAUACCGUGCCCAUGACUGUAG	TC109941	110	sense	Ga	3prime	80:1
novel2a	ACUUUUGAACUGGAUUUGCCGA	AI054573	123	sense	Gh	5prime	39:1
novel2b	UCUUUUGAACUGGAUUUGCCGA	EV497941	291	sense	Gh	5prime	11:1
novel2c	UCUUUUGAACUGGAUUUGCCGA	TC94314	491	antisense	Gh	5prime	11:1
novel3	UGGUGUGCAGGGGGUGGAAUA	DW514754	161	sense	Gh	3prime	8:5

Table 3. Conserved and Novel Cotton miRNA precursors identified from Cotton Gene Index 9 (CGI9); *Gr – Gossypium raimondii, Gh – Gossypium hirsutum, Ga – Gossypium arboreum; **Precursor was sequenced in this study

precursors were found that corresponded to three cotton species: *Gossypium hirsutum*, *Gossypium raimondii*, and *Gossypium arboreum*. For 16 of the cotton precursors, a miRNA* sequence was detected at a low abundance. These miRNA* sequenced existed on the strand immediately opposite of the mature miRNA with a 2-bp overhang, consistent with what is expected given a *DCL1*-mediated cleavage event on a RNA hairpin precursor (Reinhart *et al.* 2002, Meyers *et al.* 2008).

Nearly all of conserved mature miRNAs originated from the same miRNA end of the hairpin stemloop (5prime or 3prime) as those deposited in miRBase from other species such as *Populus trichocarpa* (cottonwood) and *Arabidopsis* (http://microrna.sanger.ac.uk/; Table 3, Figure 2). After visualizing the structures using the UNAFold package, it was found that there was tremendous heterogeneity in the size of the hairpins, even within the same miRNA family. Several motifs, such as bulges or mismatches, that existed in the secondary structures were conserved with homologous sequences from other species (Figure 2). In the miR156 family, for example, we see the conservation of a 5'UU-3'C bulge proximal to the loop of the hairpin in both the *Gossypium hirsutum* and *Arabidopsis* precursors (Figure 2a). While the majority of known miRNAs are processed from the same miRNA strand across species, we find one miRNA whose mature miRNA strand varies. In cotton, miR482 is processed from either strand depending upon the precursor, while in *Populus*, and all other species known to have miR482, it arises solely from the 3' end (Figure 2b). We can see greater conservation of structural features in precursors in which the miR482 is processed from the 3' end, including 4-5 bulged nucleic acids in the 3' end, which does not exist in the GhmiR482 precursor whose mature strand is processed from the 5' end (Figure 2b). Because the orientation of this miRNA changed, it must have also evolved a novel set of targets, and was predicted to target sucrose synthase (Table 3).

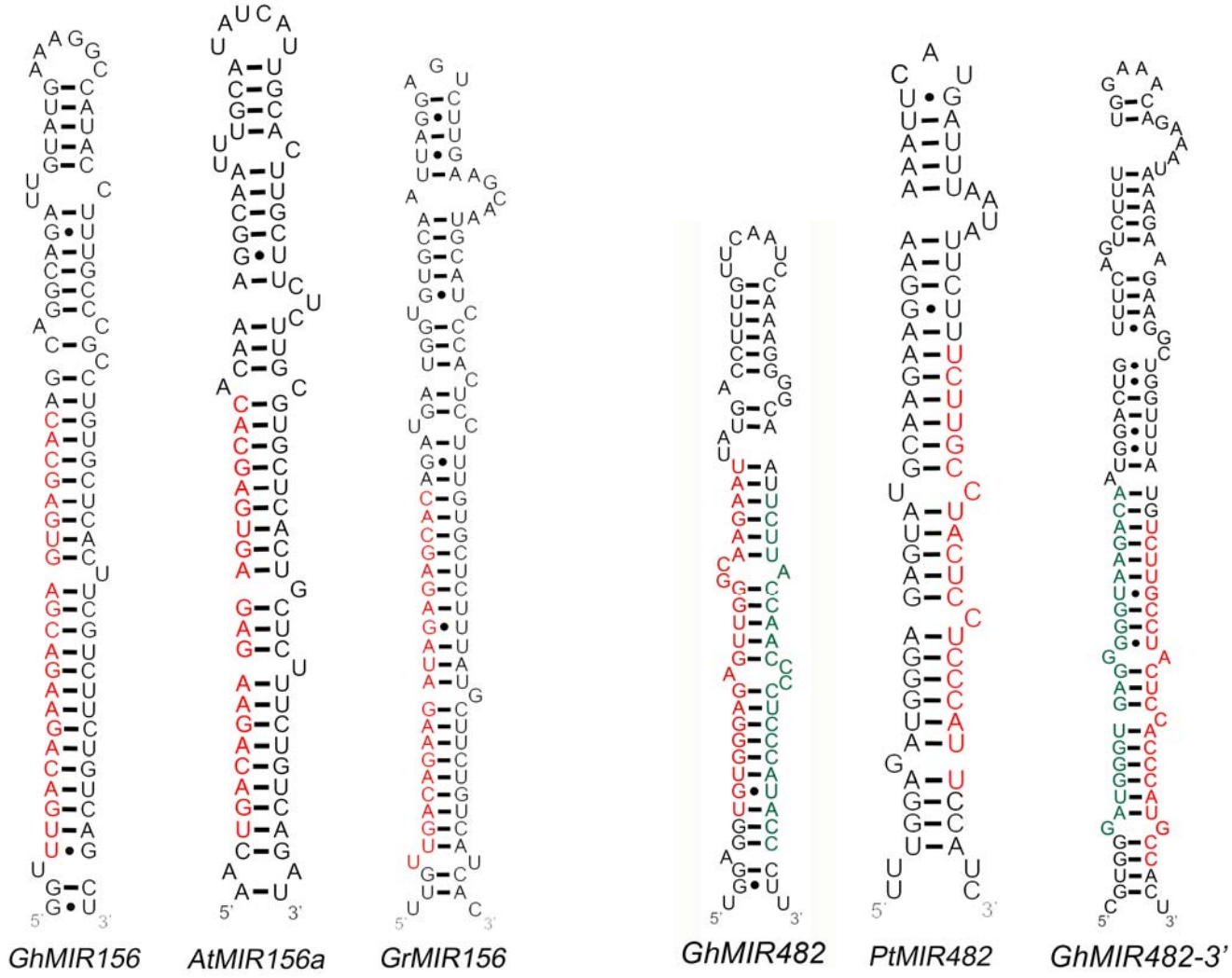

Figure 2. Secondary structures of miRNA precursors from the a) miR156 and b) miR482 families. Red sequence indicates the location of the mature miRNA, green indicates miRNA* sequence. Conservation of structural features can be observed when comparing with the At (*Arabidopsis*) and Pt (*Populus trichocarpa*) miRNA precursors with their Gh (*Gossypium hirsutum*) and Gr (*Gossypium raimondii*) counterparts.

Finally, we summarize multiple forms of evidence this study provides for the existence of miRNA families, including sequence homology to known miRNAs, the detection of a precursor in CGI9, microarray or Northern blot verification of miRNA expression, the successful validation of the cleavage of predicted targets, or the detection of a miRNA* transcript (Table 4).

	Embryophyte	Dicots			Monocots			Evidence
miRNA	***Physcomitrella***	***Arabidopsis***	***Populus***	***Vitis***	***Oryza***	***Sorghum***	***Zea***	
156/157	3	12	11	9	12	5	11	**H, P, M, S**
159		3	6	3	6	2	4	**H, N, M, T**
160	9	3	8	6	6	5	6	**H, M, T**
162		2	3	1	2		1	**H, P, M, S**
164		3	6	4	6	3	4	**H, N, P, M, T**
165/166	13	9	17	8	14	7	13	**H, N, P, M, T, S**
167	1	4	8	5	10	7	9	**H, N, P, M, T**
168		2	2	1	2	1	2	**H, N, M, T**
169		14	32	25	17	9	11	**H**
170/171	2	4	14	9	9	6	11	**H, P, M**
172		5	9	4	4	5	5	**H, N, P, M, T, S**
319	5	3	9	5	2	1	4	**H**
390	3	2	4	1	1			**H, P, M, T**
393		2	4	2	2	1	1	**H, P, M**
394		2	2	3	1	2	2	**P**
396		2	7	4	6	3	4	**H, P**
397		2	3	2	2			**H, M**
398		3	3	3	2			**P, M**
399		6	12	9	11	9	6	**H, P**
408	2	1	1	1	1		1	**H, M**
479			1	1				**P**
472/482		1	4	1				**H, P, S**
530			2		1			**H**
535	4			5	1			**H, M**
827		1	1		3			**H, P, S**
828		1		2				**H, T**
894	1							**H**

Table 4. Conservation of miRNA families across species, and the evidence provided in this study for their existence in cotton. H: Homology; P: Precursors detected; M: Microarrays; N: Northern validated; T: Targets identified; S: miRNA* sequenced Numbers represent the number of precursor loci deposited in miRbase 13.0.

Nine miRNAs (miR156/157, miR160, miR165/166, miR167, miR170/171, miR319, miR390, miR408, miR535) appeared to be deeply conserved across all plants, including in *Physcomitrella patens*, an extant moss that was one of the first plants to populate the land. Twelve other families seemed to be specific to the flowering plants, while three were apparently dicot-specific

(miR479, miR472/482, miR828). Lastly, one miRNA was conserved in both cotton and moss (miR894), although this miRNA was detected only through its sequence homology (Table 4).

DISCUSSION

The Role of 24-bp small RNAs in Genome Maintenance

In this study we find a large fraction of 24-bp small RNAs, especially in the 0 DPA and +3 DPA stages, during the process of fiber initiaton. This enormous percentage was significantly greater than that described in *Arabidopsis* (Kasschau *et al.* 2007), and occured in a tissue-specific fashion. It was further found that a much greater portion of small RNAs mapped to the genome relative to a cotton transcriptome database. This suggests that many of the 24-bp small RNAs map to intergenic regions in the cotton genome. This finding is consistent with the notion that many 24-bp are in fact short interfering RNAs (siRNAs) derived from repetitive elements, transposons, and retrotransposons. These siRNAs have been shown to have diverse roles within the plant nucleus, including methylation of genomic regions through complementarity, mediated by the recruitment of mehtyltransferases by *ARGONAUTE4* (Zilberman *et al.* 2003). This methylated DNA is oftentimes silenced. Thus, we postulate that the large fraction of 24-bp small RNAs are likely siRNAs involved in silencing much of the repetitive fraction of the genome, which composes nearly 40-65% of the the genome. Much of the variation in the size of the cotton genome is due to the "selfish elements" such as the *gypsy* and *copia* retrotranspon families, which have amplified and dispersed through the genome (Hawkins *et al.* 2006). Many 24-bp siRNAs were derived from these retrotranspons (data not shown), suggesting that these siRNAs may be an endogenous silencing mechanism for retrotransposable elements that pose the threat of being activated in certain tissues and amplifying through the cotton genome.

Biological Roles for Conserved and Novel miRNAs in Cotton

We demonstrate that a majority of conserved miRNAs in cotton retain targets that are homologs of known targets in *Arabidopsis* (Sunkar and Zhu 2004, Rajagopalan *et al.* 2006). However, it is possible that several miRNAs have evolved novel targets in cotton, especially miR159, miR162, and miR482. Four conserved miRNAs (miR160, miR167, miR390, and miR393) were predicted to target proteins that aid in the response to auxin. These miRNA sequences were highly abundant in ovule tissue, with miR167 increasing over the early stages of ovule development. This suggests that auxin response may be important in the early stages (-3 DPA) of fiber development, but decreasingly influential in the later stages, consistent with *in vitro* studies investigating the growth of fibers after treatment of ovules with auxin (Gokani and Thaker 2002). Besides their role in the auxin response, it was found that one potentially novel miRNA could be involved in gibberellin regulation, a known phytohormonal signal that regulates early fiber development (Yang *et al.* 2006).

Common Features of miRNAs Across Plant Species

While comparing cotton precursor structures to those of other species, it was found that several contained conserved structural features, such as certain mismatches and bulges in the hairpin. This inter-species conservation of structural motifs in the secondary structures of precursor miRNAs has been suggested to work as a positioning mechanism to guide *DCL1* to cleave a particular region of the hairpin into a mature strand, and specify whether the mature strand be on the 5' or 3' end (Jones-Rhoades *et al.* 2006). Thus, identifying precursors in a diverse group of species allows one to perform comparative studies to identify conserved structural features, which could aid in finding miRNA precursors using *ab initio* techniques.

Besides the conservation of structural features in precursors, several miRNAs themselves were deeply conserved across all plant lineages, suggesting that they play critical roles in plant development. Most notably, nine miRNAs were deeply conserved across many flowering plant and basal land plant lineages. These miRNAs are thought to be ancient and have been conserved over 500 million years of evolution (Axtell *et al.* 2007). In addition, we provide evidence for a miRNA (miR894) that has been conserved in moss and cotton, but has not been detected in other flowering plant species. If this is indeed an authentic miRNA, this pattern of conservation suggests that it may have been retained in certain species (such as cotton) for specialized purposes. Many other miRNAs were conserved in both monocots and dicots, or solely in the dicots. This lineage-specific conservation of miRNAs may provide some insight into their regulatory roles. For example, miR172, a flowering-plant specific miRNA, is known to be a crucial regulator of the floral homeotic transcription factor APETALA2 (Dugas and Bartel 2004, Jones-Rhoades *et al.* 2006). Future work will aim to rationalize the conservation patterns of other miRNA families, including those that have evolved as novel miRNAs in the dicots.

ACKNOWLEDGEMENTS

I would like to thank Dr. Andrew Woodward for generating the small RNA sequences and Dr. David Pang for validating the predicted miRNA targets in this study. I am also greatful to the members of the Chen laboratory for helpful suggestions and methodological advice, including Dr. Misook Ha for insight into computational methods employed in small RNA analysis. I would especially like to thank Dr. Jeff Chen, my supervising professor, for overlooking the directions of this research and providing me with the opportunity to work in his laboratory.

REFERENCES

Allen, E., Xie, Z., Gustafson, A.M., and Carrington, J.C. (2005). MicroRNA-directed phasing during *trans*-acting siRNA biogenesis in plants. Cell 121: 207–221.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Axtell MJ, Snyder JA, and Bartel DP. (2007). Common functions for diverse small RNAs of land plants. Plant Cell 19: 1750–1769.

Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Chen X, Stelly DM, Rabinowicz PD, Town CD, Arioli T, Brubaker C, Cantrell RG, Lacape JM, Ulloa M, Chee P, Gingle AR, Haigler CH, Percy R, Saha S, Wilkins T, Wright RJ, Van Deynze A, Zhu Y, Yu S, Guo W, Abdurakhmonov I, Katageri I, Ananda-Kumar P, Rahman M, Yusuf Zafar Y, Yu JZ, Kohel RJ, Wendel JF, and Paterson AH. (2007). Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol. 145:1303-1310.

Dugas DV, Bartel B. (2004). MicroRNA regulation of gene expression in plants. Curr Opin Plant Biol. 7(5):512-520.

Gokani SJ, Thaker VS. (2002). Physiological and biochemical changes associated with cotton fiber development: IX. Role of IAA and PAA. Field Crops Res. 77(2):127-136.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. 16(10):1252-1261.

Jones-Rhoades MW, and Bartel DP. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol. Cell 14: 787–799.

Jones-Rhoades MW, Bartel DP, Bartel B. (2006) MicroRNAs and their regulatory roles in plants. Annu Rev Plant Biol 57:19-53.

Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC.(2007). Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. PLoS Biol. 5(3):e57.

Llave C, Xie Z, Kasschau KD, Carrington JC. (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science 297(5589):2053-2056.

Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, Griffiths-Jones S, Jacobsen SE, Mallory AC, Martienssen RA, Poethig RS, Qi Y, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu JK. (2008) Criteria for annotation of plant MicroRNAs. Plant Cell 20: 3186-3190.

Rajagopalan R, Vaucheret H, Trejo J, and Bartel DP. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev. 20: 3407–3425.

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. (2002). MicroRNAs in plants. Genes Dev. 16(13):1616-1626.

Sunkar R, Zhu JK (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. Plant Cell 16(8):2001-19.

Yang SS, Cheung F, Lee JJ, Ha M, Wei NE, Sze S-H, Stelly DM, Thaxton P, Triplett B, Town CD, Chen ZJ. (2006). Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. Plant Journal 47:761-775.

Zilberman D, Cao X, Jacobsen SE. (2003). *ARGONAUTE4* Control of Locus-Specific siRNA Accumulation and DNA and Histone Methylation. Science 299(5607):716-719.