Development of Computational Tools and Resources for Cotton microRNA Analysis

by

Fuliang Xie

October, 2014

Director of Dissertation: Dr. Baohong Zhang

Major Department: Department of Biology

MicroRNAs (miRNAs) are an extensive class of small regulatory RNAs which regulate gene expression at the posttranscriptional levels. miRNAs target genes for mRNA cleavage or translation inhibition based on the complementary between the mRNAs and its corresponding miRNAs, and these miRNA target genes control development timing, organ development and response to environmental stress; thus miRNAs have been shown to play important roles in almost all biological and metabolic processes. Upland cotton (*Gossypium hirsutum L.*), one of the most important fiber producing crops, is widely planted in the world. Upland cotton originated from the reunion of two ancestral cotton genomes (A and D genomes) approximately 1-2 Myr ago, owning a complicated genome of allotetraploid (AADD, 2n=4x=52), with a haploid genome size estimated to be around 2.5 Gb. To date, about 80 miRNAs have been subsequentially identified in cotton by computational prediction or small RNA sequencing, many of which were also shown to be expressed differentially during fiber development. However, although miRNA-related research has become one of the hottest research in biology in the past decade and thousands of miRNAs have been identified, miRNA-related research in cotton is far beyond other plant species. One of the major reason is because of limited computational tools

and resources for cotton. In this dissertation project, we first developed a comprehensive computational tool named miRDeepFinder, which can be used for miRNA identification, target prediction and GO-/KEGG-based functional analysis for both model and non-model plant species. A case study with a small RNA sequencing data of Arabidopsis showed miRDeepFinder is an accurate and robust tool for plant miRNA analysis in deep sequencing, since 12 of 13 novel miRNAs in Arabidopsis identified by miRDeepFinder were further confirmed by qRT-PCR. miRDeepFinder also incorporated the popularly-used Cleaveland software package for analysis of degradome sequencing data. Although cotton genome is still not available, huge cotton ESTs could be a good data resource for identification of cotton miRNAs and their targets. To better utilize cotton ESTs for miRNA identification, we globally re-assembled all the cotton ESTs and developed it to a cotton EST database, in which cotton coding genes and miRNAs were deeply annotated using BLASTx, BLASTn, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) resources. A total of 28,432 unique contigs were assembled from all 268,786 cotton ESTs currently available, belonging into 5,461 groups with a maximum cluster size of 196 members. Using these contigs, we also performed EST-based investigations of comparative transcriptome similarity between cotton and other plant species, sequence polymorphisms, expressed miRNAs and their targets, and SSR analysis. A total of 27,956 indel mutants and 149,616 single nucleotide polymorphisms (SNPs) were identified from consensus contigs. In a comparison with six model plant species, cotton ESTs show the highest overall similarity to grape. We also identified 151 and 4,214 EST-simple sequence repeats (SSRs) from contigs and raw ESTs respectively. Finally, all results were integrated to a comprehensive web-based cotton EST database (www.leonxie.com), in order to make these data widely available, and to facilitate access to EST-related genetic information. Subsequently, 3 cotton small RNA

sequencing libraries treated by control, drought and salinity were sequenced. Based on miRDeepFinder, annotated cotton EST database, and cotton D genome of *Gossypium raimondii*, we identified 337 miRNAs with precursors in total, including 289 known miRNAs and 48 novel miRNAs. 155 of 337 miRNAs were found to be expressed differentially amongst the three treatments. Target prediction, GO-based functional classification, and KEGG-based functional enrichment uncovered many miRNAs and their stress-related targets might play roles in response to salinity and drought stresses. Using CitationRank-based literature mining, we sorted out the importance of genes related to stress of drought and salinity, respectively. It turned out NAC family, MYB family and MAPK family were ranked top under the context of drought and salinity, indicating their important roles for plant to combat stress of drought and salinity. To identify potential miRNAs and mRNA genes that significantly contribute to cotton fiber development, we constructed two libraries of 1-DPA (days post anthesis)-old leaf and ovule and sequenced them. A total of 128 pre-miRNAs, including 120 conserved and 8 novel pre-miRNAs were identified in cotton by miRDeepFinder. At least 40 miRNAs were either leaf or ovule-specific, whereas 62 miRNAs were shared in both leaf and ovule. Many transcription factors and other genes important for development of fiber were predicted to be miRNA targets. 22 predicted miRNA-target pairs were further validated by degradome sequencing analysis. In addition to miRNAs, we also identified 11 potential tasiRNAs-derived genes, many of which also might be involved in fiber development. miRNAs from cotton A and D genomes that reunioned together ~1-2 Myr ago might experience similar evolution pattern with coding genes. However, little is known about miRNA origin, expansion, loss, duplication, whether different derived miRNAs exchange with or affect each other, and how different genome-derived miRNAs and different genome-derived coding gene interact in cotton. To this, we systematically investigated miRNA

expansion, expression pattern, miRNA targets amongst three cotton species *Gossypium hirsutum* (AADD), *Gossypium arboreum* (AA), *Gossypium raimondii* (DD). The origin of miRNAs and coding genes were the first to be categorized in upland cotton. Our results also showed that cotton-specific miRNAs might undergo remarkably expansion and some highly conserved miRNAs were likely to be lost despite most of conserved miRNAs were remained after genome polyploidization. The comparison of miRNA expression during seedling and fiber at 5 developmental stages revealed that different genome-derived miRNAs and miRNA*s displayed asymmetric expression pattern, implicating their diverse function in upland cotton phenotype. Upon all the identified miRNAs identified in upland cotton above, we also globally investigated miRNA modification features in cotton. Besides the observation of some similar modification features with other plant species in cotton, we also found many interesting modification forms, such as modification balance between 5' and 3' end miRNAs. Comparison of isomiR expression shows differential miRNA modification amongst the 6 developmental stages in terms of selective modification form, development-dependent modification, and differential expression abundance. In contrast to previous reports, cytodine is more frequently truncated and tailed from the two ends of isomiRs in cotton, implying existence of a complex cytodine balance in isomiRs. Together, we developed a comprehensive computational tool and data resource for cotton miRNA research, and used these tools to investigate miRNA roles in cotton fiber development and response to abiotic stress. Cotton miRNA evolution and modification were also studied. Thus, our tools, data resources and research findings would contribute us to deciphering miRNA regulatory function and evolution in cotton.

DEVELOPMENT OF COMPUTATIONAL TOOLS AND RESOURCES FOR COTTON MICRORNA ANALYSIS

A Dissertation
Presented to

The Faculty of the Interdisciplinary Doctoral Program in Biological Sciences

Department of Biology, East Carolina University

Thomas Harriot College of Arts and Sciences

Submitted in Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Interdisciplinary Doctoral Program in Biological Sciences

By

Fuliang Xie

October 30, 2014

DEVELOPMENT OF COMPUTATIONAL TOOLS AND RESOURCES FOR

COTTON MICRORNA ANALYSIS

by

Fuliang Xie

APPROVED BY:

DIRECTOR OF DISSERTATION:_____
Baohong Zhang, Ph.D

COMMITTEE MEMBER:_____
Elizabeth Tweedie Ables, Ph.D

COMMITTEE MEMBER:_____
Xiaoping Pan, Ph.D

COMMITTEE MEMBER:_____
Yiping Qi, Ph.D

COMMITTEE MEMBER:_____
John Stiller, Ph.D

COMMITTEE MEMBER:_____
Qiang Wu, Ph.D

DIRECTOR, INTERDISCIPLINARY DOCTORAL PROGRAM IN BIOLOGICAL SCIENCES:

_____
Terry L. West, PhD

DEAN OF THE GRADUATE SCHOOL:

_____
Paul J. Gemperline, PhD

ACKNOWLEDGEMENTS

With the advent of the completion of my dissertation, numerous moved and warm moments, scenario, and words are increasingly sprung up into my brain, which enabled me build a strong heart in handling challenge and difficulty in life and study, and armed me with creative idea and academic knowledge. Therefore, I owe many thanks to the countless people around me over near five-year work and Ph.D study in ECU. First of all, I must thank my advisor, Dr. Baohong Zhang for the great opportunity of working and studying in his lab and in the field of biology research. My sincere appreciation to Dr. Zhang should date back to 2006 in which I received his patient and selfless guidance on the hot research filed of miRNAs by numerous emails. It was also because of Dr. Zhang, I successfully obtained my master degree and got a good job in Shanghai in 2007. Later on, he offered me a great opportunity to come and study in ECU regardless of my master degree. In my eyes, during the five years at ECU, he has token me as his student, his friend, and even one of his families, offering me countless help and care. He set up us a good example in how to perform academic research and how to be nice to other people, definitely benefiting my future work and life. I think I would always miss his wisdom and creativity in academic research, and his kindness to others in future. I would also like to extend my great gratitude to my committee members, Dr. Elizabeth Tweedie Ables, Dr. Yiping Qi, Dr. John Stiller, Dr. Qiang Wu, and Dr. Xiaoping Pan for valuable suggestions, inclusiveness, and humanized understanding and support.

I would also like to thank Department of biology and Interdisciplinary Doctoral Program in Biological Sciences (IDPBS) for providing me the opportunity to study here and finical support. I won't never forget so many nice people in ECU including Dr. Terry West, Dr. Marry Farewell, Dr. Ed Stellwag, Dr. Jeff McKinnon, Dr. Yong Zhu, Dr. Peng Xiao, Joyce Beatty, Kristen Andrews, and Barbara Beltran. Your help and support eased my life and study in USA.

I would like to thank the members of the Zhang lab and Pan lab, including Faten Taki, Dongliang Chen, Yanqiong Zhang, William (Brandon) Winfrey, Caitlin E. Burklew, and Taylor P Frazier for their help and discussion throughout the five years. It is my pleasure to be your friend and stay with you, since we have spent a wonderful time in the lab.

Finally and most importantly, I should express the greatest gratitude to my parents for cultivating me for so many years. My parents have sacrificed a lot for the education and good future of their three children. I can not imagine I could walk out from a small remote village and obtain a Ph.D degree without their support. No matter when, where, and how, my family is always my heart harbor that makes me stronger and more confident. The ultimate and greatest thanks to my wife, Yiyi, who makes everything I did and am doing meaningful. She is the core drive for me to keep moving on.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY DATA

LIST OF SYMBOLS OR ABBREVIATIONS

ABA: Abscisic acid

AGO: Argonaute protein

ANOVA: Analysis of Variance

AP1: Apetala1

AP2: Apetala2

APX: Ascorbate Peroxidase

ARF: Auxin Response Factor

ATP: Adenosine Triphosphate

ATPS: ATP Sulfurylase

bHLH: Basic helix-loop-helix–related protein

CAT: Catalase

CBC: Cap Binding Complex

cDNA: Complementary DNA

CFE1: Fiber expressed protein 1

Cg1: Corngrass1 gene

CMD: Cotton Marker Database

CPC: CAPRICE transcription factor

CPR1: NADPH:cytochrome P450 reductase 1

CSD: Cu/Zn superoxide dismutases

CT: Cycle Threshold

DCL1: Dicer-like 1

DHAR: Dehydroascorbate reductase

DPA: Day post anthesis

EF1α: Elongation Factor 1α

EST: Expressed sequence tag

ETC1: ENHANCER of TRY and CPC1

G. arboreum: Gossypium arboreum

G. hirsutum: Gossypium hirsutum

G. ramondii: Gossypium raimondii

GAPDH: Glyceraldehyde 3-Phosphate Dehydrogenase

GEO: Gene Expression Omnibus

GhCIPK6: Cotton CBL-interacting protein kinase gene 6

GhMPK2: Cotton group C MAP kinase gene 2

GL1: GLABROUS1

GO: Gene ontology

GOPX: Guaicol peroxidase

GPX: Glutathione peroxidase

GRF: Growth Regulating Factor

GSS: Genome survey sequence

GWAS: Genome-wide association studies

HD-ZIP: Homeobox leucine zipper protein

HEN1: HUA ENHANCER1

HSP: High-scoring segment pair

KEGG: Kyoto Encyclopedia of Genes and Genomes

MAS: Marker-assisted selection

miRBase: microRNA database

miRISC: miRNA-induced silencing complex

miRNA *: microRNA star

miRNA: microRNA

MS: Murashige and Skoog medium

MYB: MYB Transcription Factor

ORF: Open reading frames

PAZ: Piwi-Argonaute-Zwille

PPRs: Pentatricopeptide repeat gene transcripts

pre-miRNAs: Primary microRNA precursor

pri-miRNAs: Primary microRNA

qRT-PCR: Quantitative Reverse Transcription PCR

QTL: Quantitative trait locus

Rfam: Sanger RNA family database

RISC: RNA induced silencing complex

ROS: Reactive oxygen species

RPM: Reads per million

rRNA: Ribosomal RNA

SADR: Serious adverse drug reaction

SAM: Shoot apical meristem

siRNA: Small interfering RNA

snoRNA: Small nucleolar RNA

SNP: Single nucleotide polymorphisms

snRNA: Small nuclear ribonucleic acid

SOD: Superoxide dismutase

SPL: SQUAMOSA-promoter binding-like transcription factor

SSRs: Simple sequence repeats

Susy: Sucrose synthase

tasiRNA: Trans-acting siRNA

TCL1: TRICHOMELESS1

TCP: TCP transcription factor

TE: Transposable element

TIR1: TRANSPORT INHIBITOR RESPONSE 1

tRNA: Transfer RNA

TRY: TRIPTYCHON

TTG1: TRANSPARENT TESTA GLABRA1

UGT71C3: UDP-glucosyl transferase 71C3

UTG71C4: UDP-glucosyl transferase 71C4

WGD: Whole genome duplication

CHAPTER 1: INTRODUCTION

Cotton is one of leading economic crops in the world mainly because of its nature lint fiber, an important material for clothing, fine paper, and other purposes. Currently, increasing research is being performed on cotton to improve its fiber yields and quality, including studying related mechanisms of fiber development and cotton environmental adaption to salinity and drought. Cotton fiber development is a complex process and involves a large number of genes that function in different stages of development (Zhang et al., 2005). To date, many genes have been identified to play key roles in fiber development by different techniques, including cloning and microarray; those genes include cellulose synthase (*CelA*), MYB, 14-3-3, adenylyl cyclase associated protein (*CAP*), β-1,4-glucanase (*BG*), and sucrose synthase (Reyes and Chua, 2007; Wang et al., 2012; Wang et al., 2004). However, these genes are just a tip of iceberg, since they are merely known to be important in fiber development. How these genes regulate cotton fiber development is still not clear. Besides fiber development-related coding genes, a set of microRNAs (miRNAs) were detected to express differentially or specifically during fiber or ovule development (Kwak et al., 2009; Pang et al., 2009; Yin et al., 2012).

miRNA is a class of small non-coding RNAs in length of ~22 nt, regulating their target mRNAs either by degradation or protein translation repression. It is well known that the majority of miRNAs originate from non-coding regions including intron and intergenic regions on genome (Bartel, 2004). A great number of plant miRNAs were identified either from expression sequence tags (ESTs), genomic survey sequences (GSS), or genome. Comparative studies have shown that many miRNAs are highly evolutionarily conserved

from species to species (Zhang et al., 2006). This conservation mainly is at the level of mature miRNAs across different species. Despite miRNA-conservation-based bioinformatics approaches have speeded up the discovery of miRNAs, they have been of limited usage when it comes to many non-conserved miRNAs (Xie et al., 2012). Using different strategies, several miRNAs have been identified in cotton (Kwak et al., 2009; Pang et al., 2009; Qiu et al., 2007; Zhang et al., 2007). However, compared with hundreds of miRNAs from some model plant species, there are a lot of miRNAs in cotton to be identified due to only 80 mature miRNAs for cotton. One of the major reasons is due to the limited resources and tools.

Upland cotton (*Gossypium hirsutum L.)* is widely planted in the world, owning a complicated genome of allotetraploid (AADD, 2n=4x=52), with a haploid genome size estimated to be around 2.5 Gb (Hendrix and Stewart, 2005). Currently, cotton genome is still ongoing, largely because of its overall genetic and structural complexity, whose unavailability remarkably impedes cotton research related to fiber development. Fortunately, the draft D genome of the diploid cotton *Gossypium raimondii* and A genome of the diploid cotton *Gossypium arboreum* are available at present (Li et al., 2014; Lin et al., 2010; Paterson et al., 2012; Wang et al., 2012), offering an important molecular data resource for cotton research and breeding. Furthermore, there are several types of cotton genomic resources available, including bacterial artificial chromosomes (BACs), ESTs, linkage maps, and integrated genetic and physical maps (Chen et al., 2007). Udall and co-workers previously assembled cotton ESTs using a total of 185,198 sequence reads from 30 cDNA libraries (Udall et al., 2006). There currently are 300,270 EST reads available, which could be a good data resource for identifying cotton miRNAs.

In recent years, the emergence of next-generation sequencing technologies (also called deep sequencing), including Roche 454 GS System, Illumina Genome Analyzer, Applied Biosystems SOLiD System and Helicos Heliscope, have opened the door to identification and profiling of both known and novel miRNAs at unprecedented sensitivities. Expression profiling of small RNAs is highly complex and miRNA expression levels span a vast dynamic range (from tens of thousands to a few molecules per cell) (Friedlander et al., 2008). Fortunately, deep sequencing allows us not only to identify miRNAs, but also quantify their expression using high-throughput approaches. This can be accomplished with both known and novel miRNAs, conserved and non-conserved miRNAs, and at both low and high miRNA expression levels (Friedlander et al., 2008; Huang et al., 2010). Many novel miRNAs, especially tissue- or cell-specific sequences, have been uncovered by deep sequencing (Huang et al., 2010; Sunkar et al., 2008). The desirable characteristics of deep sequencing, including low cost and high efficiency, establish it as one of the most promising tools for miRNA research; however, this requires development of more extensive bioinformatics tools that can handle the huge and growing datasets produced by next-generation technologies.

To date, several computational algorithms and software tools have been developed to mine miRNAs from millions short reads generated by deep sequencing; these include miRDeep (Friedlander et al., 2008), miRExpress (Wang et al., 2009), miRAnalyzer (Hackenberg et al., 2009), and DSAP (Huang et al., 2010). However, at least 3 apparent shortcomings from these approaches or tools could be found not to be good for plant miRNA identification on basis of high-throughput sequencing: 1) almost all of available computational tools excise candidate precursors by extending a mapped locus of genome

by 200 nt or less on each side, whereas the lengths of plant miRNA precursor are rather diverse from 53 nt to 983 nt; 2) 2-nt overhangs at 3' of miRNA as a filter for miRNA identification in these tools is not always found for known miRNA in deep sequencing; 3) all of these tools depend on complete genome information.

Currently, miRNAs' function in cotton fiber-related development is little understood, mostly due to limited tools and resources for cotton miRNA-related research. Therefore, **the goal of this project** is first to develop comprehensive computational tools and resources for cotton genomic research, and then use these tools to investigate miRNA roles in cotton fiber development and response to abiotic stress. As a tetraploid species, upland cotton theoretically inherited two sets of miRNA system from A- and D-genome diploid species. Upland cotton miRNAs were likely to undergo some evolution events as its coding genes did, probably contributing to fiber or other phenotype changes. However, little is known about miRNA origin, expansion, loss, duplication, whether different derived miRNAs exchange with or affect each other, and how different genome-derived miRNAs and different genome-derived coding gene interact in cotton. Thus, based on the identified miRNAs, we are also interested in comparing miRNA evolution amongst A-, D- and AD-genome cotton species. Furthermore, to be a functional miRNA, plant miRNAs need to undergo a series of modifications including methylation, uridylation, adenylation, untemplated nucleotide addition, truncation, and tailing (Ji and Chen, 2012; Kim et al., 2010; Zhai et al., 2013). miRNA modification is found as an extensively existing phenomenon that could affect miRNA stability, miRNA diversity, or miRNA targeting specificity. However, little is known about the global miRNA truncation and modification

and their corresponded function. Therefore, we are also interested in investigating miRNA modification in cotton.

In the first project, we developed a software named miRDeepFinder, which is used for identifying and functionally analyzing plant miRNAs and their targets from small RNA datasets obtained from deep sequencing. miRDeepFinder allows user to do miRNA analysis for plant species with or without available genome information based on databases of GSS, EST, or genome. A case study on a small RNA dataset of Arabidopsis showed miRDeepFinder owns a good accuracy and a strong robustness in miRNA identification. 12 of 13 identified novel miRNAs in Arabidopsis were further confirmed by qRT-PCR. In addition, miRDeepFinder incorporated the popularly-used Cleaveland software package to further extend its function on miRNA target identification based on degradome sequencing data (Addo-Quaye et al., 2009).

Considering cotton genome is still not available, huge cotton ESTs (300,270 sequences) could be a good data resource for identification of cotton miRNAs and their targets. In the second project, we performed global assembly of cotton ESTs available from NCBI, and functional annotation using BLASTx, BLASTn, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) resources. Using the contigs obtained, we also performed EST-based investigations of comparative transcriptome similarity between cotton and other plant species, sequence polymorphisms, expressed miRNAs and their targets, and SSR analysis. Finally, we integrated these analytical data into a comprehensive web-based database so that EST-related information can be shared and queried publically.

In the third project, we constructed three cotton small RNA libraries treated by control, salinity and drought, and sequenced them by deep sequencing. Based on GSS and assembled EST of cotton, and cotton D genome (*G. raimondii)*, a total of 337 miRNAs with precursors were identified, including 289 known miRNAs and 48 novel miRNAs. miRNA expression profile comparison showed that 155 of 337 miRNAs expressed differentially amongst the three treatments. Target prediction, GO-based functional classification, and KEGG-based functional enrichment also show these miRNAs might play roles in response to salinity and drought stresses through targeting a series of stress-related genes. CitationRank-based literature mining was employed to sort out the importance of genes related to stress of drought and salinity, respectively. It turned out NAC family, MYB family and MAPK family were ranked top under the context of drought and salinity, indicating their important roles for plant to combat stress of drought and salinity.

In the fourth project, small RNAs were collected from 1-DPA (days post anthesis)-old leaf and ovule, and then sequenced. Using miRDeepFinder, a total of 128 pre-miRNAs, including 120 conserved and 8 novel pre-miRNAs were identified in cotton. At least 40 miRNAs were either leaf or ovule-specific, whereas 62 miRNAs were shared in both leaf and ovule. Many transcription factors and other genes important for development of fiber were predicted to be miRNA targets. In addition to miRNAs, we also identified 16 genes targeted by tasiRNAs, many of which also might be involved in fiber development.

In the fifth project, combining with two sequenced cotton genomes (*G. arboreum* (AA) and *G. raimondii* (DD)), we systematically investigated miRNA expansion, expression pattern, miRNA targets amongst three cotton species *G. hirsutum* (AADD), *G.*

arboreum (AA), and *G. raimondii* (DD). We were the first to categorize miRNAs and coding genes of upland cotton to different genome origin, A, D, or AD. Our results showed the overall miRNAs significantly expanded in upland cotton, whereas some highly conserved miRNAs maintain a relatively stable level. Different genome-derived miRNAs exhibit asymmetric expression pattern in the 6 developmental stages in upland cotton.

In the sixth project, we systematically investigated miRNA modification on both 5' and 3' end using small RNA sequencing data of upland cotton from 6 developmental stages, including truncation and addition. Global miRNA modification analysis in cotton revealed a series of features of cotton miRNA modification. For instance, 1-2-nt truncation and addition on both 5' and 3' ends of miRNAs consists of the major modification forms. The 5' and 3' end miRNA modification was almost equal in the 6 development stages. Truncation was more common than addition on both 5' and 3' end. In contrast to previous reports, cytodine is more frequently truncated and tailed from the two ends of isomiRs, implying existence of a complex cytodine balance in isomiRs.

Collectively, we developed a comprehensive computational tool for plant miRNA analysis in small RNA sequencing and constructed a deeply annotated cotton EST database for cotton research. Based on the tool and the cotton EST database, we systematically identified miRNAs and their targets, and investigated miRNA roles in response to abiotic stress and fiber development. Moreover, we were also the first to uncover a variety of features of miRNA evolution and modification in cotton, which might be associated with developing a cotton fiber phenotype. Thus, our study could contribute to understanding miRNA roles in cotton fiber development and response to abiotic stress.

Reference

Addo-Quaye, C., Miller, W., and Axtell, M.J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics *25*, 130-131.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Chen, Z.J., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X., Stelly, D.M., Rabinowicz, P.D., Town, C.D., *et al.* (2007). Toward sequencing cotton (Gossypium) genomes. Plant physiology *145*, 1303-1310.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nature biotechnology *26*, 407-415.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M., and Aransay, A.M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic acids research *37*, W68-76.

Hendrix, B., and Stewart, J.M. (2005). Estimation of the nuclear DNA content of gossypium species. Annals of botany *95*, 789-797.

Huang, P.J., Liu, Y.C., Lee, C.C., Lin, W.C., Gan, R.R., Lyu, P.C., and Tang, P. (2010). DSAP: deep-sequencing small RNA analysis pipeline. Nucleic acids research *38*, W385-391.

Ji, L., and Chen, X. (2012). Regulation of small RNA stability: methylation and beyond. Cell research *22*, 624-636.

Kim, Y.K., Heo, I., and Kim, V.N. (2010). Modifications of small RNAs and their associated proteins. Cell *143*, 703-709.

Kwak, P.B., Wang, Q.Q., Chen, X.S., Qiu, C.X., and Yang, Z.M. (2009). Enrichment of a set of microRNAs during the cotton fiber development. BMC genomics *10*, 457.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., Zou, C., *et al.* (2014). Genome sequence of the cultivated cotton Gossypium arboreum. Nature genetics *46*, 567-572.

Lin, L., Pierce, G.J., Bowers, J.E., Estill, J.C., Compton, R.O., Rainville, L.K., Kim, C., Lemke, C., Rong, J., Tang, H., *et al.* (2010). A draft physical map of a D-genome cotton species (Gossypium raimondii). BMC genomics *11*, 395.

Pang, M., Woodward, A.W., Agarwal, V., Guan, X., Ha, M., Ramachandran, V., Chen, X., Triplett, B.A., Stelly, D.M., and Chen, Z.J. (2009). Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). Genome biology *10*, R122.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., *et al.* (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature *492*, 423-427.

Qiu, C.X., Xie, F.L., Zhu, Y.Y., Guo, K., Huang, S.Q., Nie, L., and Yang, Z.M. (2007). Computational identification of microRNAs and their targets in Gossypium hirsutum expressed sequence tags. Gene *395*, 49-61.

Reyes, J.L., and Chua, N.H. (2007). ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. The Plant journal : for cell and molecular biology *49*, 592-606.

Sunkar, R., Zhou, X., Zheng, Y., Zhang, W., and Zhu, J.K. (2008). Identification of novel

and candidate miRNAs in rice by high throughput sequencing. BMC plant biology *8*, 25.

Udall, J.A., Swanson, J.M., Haller, K., Rapp, R.A., Sparks, M.E., Hatfield, J., Yu, Y., Wu, Y., Dowd, C., Arpat, A.B., *et al.* (2006). A global assembly of cotton ESTs. Genome research *16*, 441-450.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., *et al.* (2012). The draft genome of a diploid cotton Gossypium raimondii. Nature genetics *44*, 1098-1103.

Wang, S., Wang, J.W., Yu, N., Li, C.H., Luo, B., Gou, J.Y., Wang, L.J., and Chen, X.Y. (2004). Control of plant trichome development by a cotton fiber MYB gene. Plant Cell *16*, 2323-2334.

Wang, W.C., Lin, F.M., Chang, W.C., Lin, K.Y., Huang, H.D., and Lin, N.S. (2009). miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC bioinformatics *10*, 328.

Xie, F., Xiao, P., Chen, D., Xu, L., and Zhang, B. (2012). miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant molecular biology.

Yin, Z., Li, Y., Yu, J., Liu, Y., Li, C., Han, X., and Shen, F. (2012). Difference in miRNA expression profiles between two cotton cultivars with distinct salt sensitivity. Molecular biology reports *39*, 4961-4970.

Zhai, J., Zhao, Y., Simon, S.A., Huang, S., Petsch, K., Arikit, S., Pillay, M., Ji, L., Xie, M., Cao, X., *et al.* (2013). Plant microRNAs display differential 3' truncation and tailing modifications that are ARGONAUTE1 dependent and conserved across species. The Plant cell *25*, 2417-2428.

Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. (2006). Conservation

and divergence of plant microRNA genes. The Plant journal : for cell and molecular biology *46*, 243-259.

Zhang, B., Wang, Q., Wang, K., Pan, X., Liu, F., Guo, T., Cobb, G.P., and Anderson, T.A. (2007). Identification of cotton microRNAs and their targets. Gene *397*, 26-37.

Zhang, B.H., Pan, X.P., Wang, Q.L., Cobb, G.P., and Anderson, T.A. (2005). Identification and characterization of new plant microRNAs using EST analysis. Cell Res *15*, 336-360.

CHAPTER 2: miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs

Abstract

miRDeepFinder is a software package developed to identify and functionally analyze plant microRNAs (miRNAs) and their targets from small RNA datasets obtained from deep sequencing. The functions available in miRDeepFinder include pre-processing of raw data, identifying conserved miRNAs, mining and classifying novel miRNAs, miRNA expression profiling, predicting miRNA targets, and gene pathway and gene network analysis involving miRNAs. The fundamental design of miRDeepFinder is based on miRNA biogenesis and regulation and targeting, such as perfect or near perfect hairpin structures, different read abundances of miRNA and miRNA*, and targeting patterns of plant miRNAs. To test the accuracy and robustness of miRDeepFinder, we analyzed a small RNA deep sequencing dataset of *Arabidopsis thaliana* published in the GEO database of NCBI. Our test retrieved 128 of 131 (97.7%) known miRNAs that have a more than 3 read count in *Arabidopsis*. Because many known miRNAs are not associated with miRNA*s in small RNA datasets, miRDeepFinder was also designed to recover miRNA candidates without the presence of miRNA*. To mine as many miRNAs as possible, miRDeepFinder allows users to compare mature miRNAs and their miRNA*s with other small RNA datasets from the same species. In addition, Cleaveland software package was also incorporated into miRDeepFinder for miRNA target identification using degradome sequencing analysis. Using this new computational tool, we identified 13 novel miRNA candidates with miRNA*s from *Arabidopsis* and validated 12 of them experimentally.

Interestingly, of the 12 verified novel miRNAs, a miRNA named AC1 spans the exons of two genes (UTG71C4 and UGT71C3). Both the mature AC1 miRNA and its miRNA* were also found in four other small RNA datasets. We also developed a tool, "miRNA primer designer" to design primers for any type of miRNAs. miRDeepFinder provides a powerful tool for analyzing small RNA datasets from all species, with or without the availability of genome information. miRDeepFinder and miRNA primer designer are freely available at http://www.leonxie.com/DeepFinder.php and at http://www.leonxie.com/miRNAprimerDesigner.php, respectively. A program (called RefFinder: http://www.leonxie.com/referencegene.php) was also developed for assessing the reliable reference genes for gene expression analysis, including miRNAs.

Introduction

Since microRNAs were first discovered (miRNAs; lin-4 and let-7) in *Caenorhabditis elegans* (Lee et al., 1993; Reinhart et al., 2000), they have been established as an important class of endogenous non-coding regulatory molecules that are ~21 nt in length (Bartel, 2004). Long RNAs named primary miRNAs (pri-miRNAs) are initially transcribed by RNA polymerase II (pol II), then further processed into miRNA precursors (pre-miRNAs) by RNase III enzymes, including Drosha or Dicer-like protein (Kim, 2005). In animals, pre-miRNAs are exported from the nucleus to the cytoplasm by Exportin 5 and Ran-GTP. Pre-miRNAs then are cleaved into duplexes of miRNA/miRNA* by another RNase III enzyme, Dicer. To date, no Drosha homolog has been found in plants (Chen, 2008; Kim, 2005). Therefore, processing from pri-miRNAs into miRNA/miRNA* duplexes in plants is thought to be achieved in the nucleus only by a Dicer homolog, Dicer-

like 1 (DCL1) (Chen, 2008). Mature miRNAs, unbound from duplexes with miRNA*s, are loaded into the multi-protein RNA induced silencing complex (RISC) that contains the core protein Argonaute (AGO), whereas miRNA*s generally are degraded by an unidentified mechanism (Bartel, 2004). AGO proteins are characterized by Piwi-Argonaute-Zwille (PAZ) domains, which are the key components of RNA-silencing pathways. The critical catalytic residues in the AGO Piwi domain cleave target mRNAs in the middle of the complementary region between the mRNA and the corresponding miRNA (Chen, 2008; Ender and Meister, 2010; Herr, 2005). In addition to a cleavage-based mechanism of miRNA-mediated transcription regulation, miRNA-guided translational repression is carried out by preventing the circularization of mRNA needed for stimulation of translation (Pillai et al., 2007); however, the precise mechanisms of this translational repression are not yet clear (Ender and Meister, 2010). A great deal of evidence has demonstrated that miRNAs play crucial roles in post-transcriptional regulation involved in development, cell proliferation, apoptosis, inflammation, stress response, signal transduction, and metabolic processes (Ambros, 2004).

Recently, miRNAs have been shown to be extensive in animals, plants, and even viruses. According to the miRNA database miRBase (Release 18: November 2011) (Griffiths-Jones, 2004), there are currently a total of 18,226 miRNAs identified from 168 species.

Identification of miRNAs is a critical step for understanding the regulatory functions they carry out. Almost all currently known miRNAs were successfully identified using forward genetics, direct cloning, sequencing, and via bioinformatics methods (Bonnet et al., 2004; Chen, 2008; Grad et al., 2003; Qiu et al., 2007; Zhang et al., 2008).

Of those approaches, forward genetics requires creating loss-of-function or over-expression mutants of targeted miRNAs. Based on *Arabidopsis mir164a* and *mir164b* mutant plants, which express less miR164 and more NAC1 mRNA, miRNA 164 was shown to play an important role in the development of lateral roots by cleaving NAC1 mRNA and down-regulating auxin signals (Guo et al., 2005). Although forward genetics screens can identify miRNAs and their functions, for example, miRNA 164 (Guo et al., 2005), miRNA 172 (Aukerman and Sakai, 2003), and miR399 (Fujii et al., 2005), only a handful of miRNAs have been identified in this way. Clearly, genetic screening by itself is too time-consuming to meet the needs of identifying large numbers of miRNAs with both high accuracy and efficiency. Direct cloning appears to be a powerful tool for detecting more miRNAs by randomly selecting and sequencing clones from small RNA libraries; however, it is time-consuming and expensive (Pfeffer et al., 2005; Qiu et al., 2007).

Comparative studies have shown that many miRNAs are highly evolutionarily conserved from species to species. This conservation mainly is at the level of mature miRNAs across different species, although sometimes it also is evident in precursors and their secondary stem-loop structures (Griffiths-Jones, 2004). Many miRNAs from the same or different species can be grouped into distinct families with nuances in mature miRNA sequences. Furthermore, according to genome-wide mapping of miRNAs loci, most miRNAs come from intergenic regions of the genome, with a small portion also derived from intronic regions (Bonnet et al., 2004). Therefore, using the dominant features of miRNAs, including miRNAs biogenesis, origin, and conservation, a number of bioinformatics algorithms and tools have been developed to mine potential miRNAs in genomic and other sequence databases (Bartel, 2004; Bonnet et al., 2004; Qiu et al., 2007;

Zhang et al., 2008). In addition, bioinformatics approaches always have been useful for assisting in analysis of experimental data (Fu et al., 2005). Although bioinformatics approaches have speeded up discovery of miRNAs, they have been of limited utility when it comes to many non-conserved miRNAs or miRNAs with low expression abundance.

In recent years, the emergence of next-generation sequencing technologies (also called deep sequencing) including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, and others, have opened the door to identification and profiling of both known and novel miRNAs at unprecedented sensitivities. Expression profiling of small RNAs is highly complex and miRNA expression levels span a vast dynamic range (from tens of thousands to a few molecules per cell) (Friedlander et al., 2008). Fortunately, deep sequencing allows us not only to identify miRNAs, but also quantify their expression using high-throughput approaches. This can be accomplished with both known and novel miRNAs, conserved and non-conserved miRNAs, and at both low and high miRNA expression levels (Friedlander et al., 2008; Huang et al., 2010). Many novel miRNAs, especially tissue- or cell-specific sequences, have been uncovered by deep sequencing (Huang et al., 2010; Sunkar et al., 2008; Zhang et al., 2008). The desirable characteristics of deep sequencing, including low cost and high efficiency, established it as one of the most promising tools for miRNA research; however, this requires development of more extensive bioinformatics tools that can handle the huge and growing datasets produced by next-generation technologies.

To date, several computational algorithms and software tools have been developed to mine miRNAs from millions short reads generated by deep sequencing; these include miRDeep (Friedlander et al., 2008), miRExpress (Wang et al., 2009), miRAnalyzer

(Hackenberg et al., 2009), and DSAP (Huang et al., 2010). Almost all of these tools require available genome information or have been developed for use in a limited number of species. Moreover, according to the current miRBase, the lengths of plant miRNA precursors range from 53 nt to 983 nt. Of the 2,656 known plant pre-miRNAs, 473 (17.8%) are more than 200 nt in length. Unfortunately, almost all of available computational tools excise candidate precursors by extending a mapped locus of genome by 200 nt or less on each side (Friedlander et al., 2008; Hackenberg et al., 2009; Huang et al., 2010; Mathelier and Carbone, 2010). Clearly, miRNAs with precursor lengths of more than 200 nt can be skipped by mistake. The feature that a true miRNA should have a miRNA* in short reads, and that the two can form a duplex with 2 nt 3' overhangs, is one of the most important criteria that identifies novel miRNA (Meyers et al., 2008). However, our investigation of a short read library has shown that many known miRNAs do not have 3' overhangs in datasets generated by deep sequencing (data not shown). Therefore, it is necessary to develop a tool that can address all these issues. Here, we present miRDeepFinder, which is suitable for analyzing deep sequencing datasets. It offers two different versions for dealing with datasets both with and without genome information available. Users are able to set a variety of parameters, including sequence quality, short read abundance, length of short read, and length of extracted precursor from genome for hairpin structure analysis. Users also have the option to start from the dataset in recommended format, and a 3' overhang is not a mandatory criterion to filter miRNA candidates. To better understand regulatory functions and gene pathways involved by miRNAs, we integrated GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) analyses into miRDeepFinder. In addition, miRDeepFinder incorporated the popularly-used Cleaveland software package

to further extend its function on miRNA target identification based on degradome sequencing data (Addo-Quaye et al., 2009). Thus, miRDeepFinder not only can be used for identifying and classifying both conserved and novel miRNAs but also for gene function analysis and enrichment of gene pathway analysis.

Implementation

miRDeepFinder is programmed in Perl. Before running miRDeepFinder, several open-source software programs, including MySQL, RNAfold, EMBOSS package, BLAST from NCBI, and RepeatMasker need to be installed. More details about installation of these programs can be found in readme file of miRDeepFinder. miRDeepFinder has been tested successfully on our Linux server (Ubuntu, 10.04.1, 2.6.32-25-generic, x86_64; Perl, v5.10.1, x86_64; and MySQL, 5.1.41, x86_64). Although miRDeepFinder uses MySQL as a database management bench, it outputs user-friendly files in text and Excel formats.

Workflow

Pre-processing deep sequencing datasets.

miRDeepFinder can start from the original raw dataset generated by deep sequencing technologies like Illumina and Roche. Generally, a FASTQ file contains three kinds of data (identifier, sequence reads and sequencing quality values for each base). The size of FASTQ files usually reaches to gigabytes in size and sequences are ungrouped. We used the Perl script Adapter_trim.pl (http://centre.bioinformatics.zj.cn/mirtools/adaptortrim.php) from mirTools (Zhu et al., 2010) to filter low-quality reads (most likely from sequencing errors) and trim 3' and 5'

adaptor sequences. We further modified the program by adding rules to filter short reads. Our filters remove the following: 1) the short reads with more than 80% A, C, G, or T; 2) reads shorter than 16 nt; 3) reads containing the minimum homopolymers 7A, 8C, 6G, 7T; 4) reads with 10 repeats of a dimer, 6 repeats of a trimer, or 5 repeat of a tetramer; 5) sequences with only A+ C and no G or T; 6) sequences with only G and T; 7) sequences matching rRNAs, tRNAs, snRNAs and noRNAs in RFam (Rfam 10.0, Sanger) (Friedlander et al., 2008; Gardner et al., 2009; Griffiths-Jones, 2004). In addition, unique sequences with a copy number of more than 3 and length between 18 nt and 28 nt were retained for later analysis (Friedlander et al., 2008). Finally, the output is formatted as a tab-delimited file, which contains only the unique sequence reads and their corresponding number of copies. This is the same format of trimmed datasets downloaded from GEO database of NCBI. Therefore, users also can initiate miRDeepFinder from trimmed datasets of small RNAs from NCBI.

Expression analysis of conserved miRNAs

Many mature miRNAs have been shown to be highly conserved (Bartel, 2004). Some short reads with the same or similar sequences (no more than 3 mismatches) as known miRNAs in the miRNA database could be functional miRNAs. However, for those species without a completely sequenced genome, there may be no corresponding precursors for these short reads in existing nucleotide databases. To investigate conserved miRNA expression, we aligned sorted short reads against all known plant miRNAs using WATER (EMBOSS 6.1.0.1, http://emboss.sourceforge.net/) (Smith and Waterman, 1981) with a criterion of less than three mismatches (Figure 2-1). If a short read aligned to several

known miRNAs with different mismatches, the read with the smallest mismatch was classified as the miRNA, allowing us to discriminate conserved miRNAs from whole short read dataset, then mark and store them in MySQL for further analysis.

Identifying novel miRNAs from a deep sequencing database

To recover known miRNAs and identify novel miRNAs from a small RNA deep sequencing dataset, both conserved and non-conserved reads are mapped back to a genome, EST database or the NCBI Genome Survey Sequence database by megablast (BLAST 2.2.19, http://www.ncbi.nlm.nih.gov/) (Figure 2-1). Only perfect alignments were retained (full length, 100% identity) (Friedlander et al., 2008). The sites hit these databases are extended by 700 nt on each side to examine precursor candidates. "The extending length" in miRDeepFinder is an adaptive parameter that can be adjusted to meet diverse needs of the users.

Using genome annotation data, miRDeepFinder can analyze the genome annotation to discard precursor candidates that are located in coding-regions. Precursor candidates are further folded into RNA secondary hairpin structures by RNAfold (Vienna RNA Package 1.8.4, http://www.tbi.univie.ac.at/*ivo/RNA/index.html) (Hofacker, 2003; Zuker and Stiegler, 1981). Criteria we reported previously (Qiu et al., 2007; Zhang et al., 2006; Zhang et al., 2008) are employed to evaluate whether a predicted secondary hairpin structure is perfect. Briefly, these are 1) the candidate pre-miRNA can be folded into a perfect stem-loop hairpin secondary structure with the miRNA located in one arm of the stem at either the 5' or 3' end; 2) no more than six nucleotide mismatches are allowed between the miRNA and its opposite miRNA* sequence in the secondary structure; 3) no loops or

breaks occur in the miRNA/miRNA* duplex; 4) the minimum length of the precursor is 45 nt.

Candidate pre-miRNAs are searched against each other by BLAST to remove repeated sequences with an *E* value cut-off of 1e-20. If the analysis is performed on EST or GSS data, the original precursors sequences are used in further BLASTX searches against the NCBI plant reference protein database (http://www.ncbi.nlm.nih.gov/) using an *E* value cut-off of 1e-25 to remove any potential protein-encoding genes (Xie et al., 2010). Finally, miRDeepFinder aligns identified novel miRNAs against each other by WATER; those with fewer than 3 mismatches are grouped into a family.

After a miRNA is identified, miRDeepFinder extracts the miRNA* sequence according to the rule of a 3' overhang of 2 nts in the miRNA/miRNA* duplex. After that, miRDeepFinder continues to search for miRNA*s in the short reads dataset. If a miRNA* is found, its read count is recorded. Because many known miRNAs may not have miRNA* sequences in a given dataset, miRDeepFinder allows the user to find the miRNA* sequence in other small RNA deep-sequencing datasets from the same species that were submitted previously to the GEO database in NCBI. Thus, users can use miRDeepFinder to test whether identified miRNAs and their miRNA* co-exist in other datasets. Clearly, if both a miRNA and its miRNA* sequence simultaneously occur in multiple short read datasets, it is very likely to represent a *bona fide* miRNA.

Predicting miRNA targets

Most plant miRNAs regulate their targets by perfect or near-perfect complement base pairing (Schwab et al., 2005). By investigating the transcriptome of plants that over-

express different miRNAs, Schwab and co-workers derived a set of empirical parameters for target recognition (Schwab et al., 2005). miRDeepFinder first adopts Target-align (Xie and Zhang, 2010) to search miRNA targets against a cDNA or EST database (Figure 2-2). The rules for target recognition are described as follows: 1) no more than four mismatches are allowed between the mature miRNA and its potential target; (2) no more than two mismatches are allowed at nucleotide positions 1-9; 3) no more than two consecutive mismatches are allowed; and 4) no mismatches are allowed at positions 10 and 11 (Xie et al., 2010). To avoid hitting a repeated target, candidates are used in BLAST searches against each other to remove repeated target sequences with an e-value cut-off of 1e-20. If an EST is used to predict a target, target candidates are compared continuously against the reference protein database of all plants, using BLASTX, to remove non-coding sequences.

In relative to animals, plant miRNAs tend to act on their targets in way of perfect or near-perfect base complementarity, which in general causes AGO-catalyzed target cleavage. Degradoem sequencing is a modified 5'-rapid amplification of cDNA ends (RACE) with high-throughput deep sequencing method and is widely used to analyze patterns of RNA degradation, as well as miRNA target cleavage site. To make full use of existing degradome sequencing data for facilitating miRNA target prediction, the widely-used CleaveLand software package (CleaveLand 3.0.1, http://axtell-lab-psu.weebly.com/1/post/2011/08/cleaveland3.html) is incorporated into miRDeepFinder (Addo-Quaye et al., 2009). To further favor users to operate CleaveLand, miRDeepFinder also contained two required third-party softwares, targetFinder (targetFinder 1.6, http://carringtonlab.org/resources/targetfinder) (Fahlgren et al., 2007) and FASTA (FASTA35, http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml). The whole

degradome analysis could be done automatically in batch. Considering CleaveLand uses targetFinder as its initial miRNA prediction tool, we meanwhile added targetFinder as one more miRNA target search engine besides of our miRNA target identification tool, Target-align.

Classifying target function and enriching target-involved pathways

Multiple miRNAs can affect the same biological process; when this occurs, many complex miRNA-miRNA synergistic networks can occur (Xu et al., 2010). Thus, using bioinformatics to help classify miRNA functions and enrich pathway that involve miRNAs can strengthen our understanding of miRNA function substantially. Toward this purpose, we incorporated the GO database (ftp://ftp.geneontology.org/godatabase/archive/) (Ashburner et al., 2000) into miRDeepFinder. miRDeepFinder-identified targets are compared against the GO protein database using BLASTX with an *E* value cut-off of 1e-20. Only the highest hit is retained as evidence for classifying the potential function of the miRNA in question. Meanwhile, to further investigate regulation pathways, we constructed a MySQL scheme to manage the KEGG pathway database (ftp://ftp.genome.jp/pub/kegg/) (Kanehisa and Goto, 2000). Using KEGG information annotated in the GO protein database, we can detect corresponding pathways using a combined-search against the GO and reconstructed KEGG pathway database.

Testing miRDeepFinder with a small RNA dataset

To test miRDeepFinder's performance, the small RNA dataset GSM442933 from *Arabidopsis thaliana* was downloaded from NCBI and used as our reference dataset.

According to miRDeepFinder's workflow, both known and novel miRNAs from *A. thaliana* were analyzed. Novel miRNAs and their miRNA*s predicted from the dataset GSM442933, were tested to determine whether they occurred in other 11 small RNA datasets (GSE18302, GSM313212, GSM442932, GSM442933, GSM442934, GSM442935, GSM456944, GSM456945, GSM518429, GSM518430, and GSM518431) from *A. thaliana*. Only a miRNA found along with its miRNA* more than 4 times in the 11 datasets was kept for later experimental validation. A total of 102 known miRNAs in *A. thaliana* with known functions were employed to test miRDeepFinder's miRNA's target prediction and downstream GO and KEGG analyses.

Validating novel miRNAs in Arabidopsis using stem-loop RT and qRT-PCR

We conducted stem-loop RT followed by TaqMan PCR analysis (Chen et al., 2005) to validate novel miRNA candidates. Ten day old seedlings grown in culture medium were harvested and immediately frozen in liquid nitrogen. Total RNA was extracted from seedlings using the mirVanaTM miRNA Isolation Kit (Ambion, Austin, TX, USA) according to the manufacturer's protocol. Although stem-loop RT is highly sensitive and is not influenced by genomic DNA contamination (Chen et al., 2005), small RNA still was isolated from total RNA using the flashPAGE Fractionator (Ambion, Austin, TX, USA) and flashPAGE™ Reaction Clean-Up Kit (Ambion, Austin, TX, USA). The small RNA was qualified by Nanodrop ND-1000 (Nanodrop technologies, Wilmington, DE, USA) and then stored at -80°C for later use. According to rules for designing stem-loop RT primers reported by Chen and co-workers (Chen et al., 2005), a small tool named miRNA primer designer was developed in Csharp (Window version) and in PHP (Web version) for

improved design of primers based on mature miRNA sequences. The tool is available from the website (http://www.leonxie.com/miRNAprimerDesigner.php). Thirteen miRNA-specific primers for reverse transcription PCR (RT-PCR) (Table 2-1) and 13 miRNA-specific primers for qRT-PCR (Table 2-2) were designed. All qRT-PCR reactions were run on an Applied Biosystems 7300 Sequence Detection System (Foster City, CA, USA) according to the manufacturer's instructions.

Result and discussion

Running miRDeepFinder on a small RNA dataset of Arabidopsis thaliana

To test miRDeepFinder, we downloaded the small RNA dataset GSM442933 (Hsieh et al., 2009) as reference data. GSM442933 contains a total of 701,277 unique short reads ranging in length from 16 to 27 nt and corresponding to about 3.5 million reads. Most of reads range from 18 nt to 24 nt and account for 93.4% (data not shown).

Many miRNAs have been shown to be highly conserved (Bartel, 2004). Conserved miRNAs, whether from the same or different species, are considered to have no more than 3 mismatches in mature miRNA sequences (Griffiths-Jones, 2004). In order to investigate known and potential conserved miRNAs in a small RNA library, miRDeepFinder defines short reads having no more than 3 mismatches with known miRNAs as conserved reads. Under default settings (reads of 17~35 nt), 9,679 short reads in *A. thaliana* were aligned against a total of 1,250 unique known miRNAs from plants. A total of 631 reads were identified as conserved miRNAs corresponding to 182 distinct miRNAs (Supplementary 2-6). Of the 182 known miRNAs found, 92 were from *A. thaliana* and 27 from *Oryza sativa*. The 631 conserved reads contained 122 known miRNA families. The twelve most abundant

miRNA families were miR-2911 (35 members), miR-2916 (34 members), miR-166 (30 members), miR-158 (29 members), miR-167 (26 members), miR-169 (26 members), miR-822 (23 members), miR-156 (22 members), miR-396 (21 members), miR-161 (20 members), miR-168 (20 members), and miR-165 (16 members) (Supplementary 2-2). Statistical analysis of read counts by miRNA families indicated the following expression levels in decreasing order: miR-158 (257,432 reads), miR-166 (105,524 reads), miR-165 (25,367 reads), miR-156 (22,932 reads), miR-2911 (20,587 reads), miR-161 (17,859 reads), miR-167 (14,203 reads), miR-168 (5,594 reads), miR-173 (4,833 reads), and miR-822 (4,371 reads) (Supplementary 2-2). In addition, to better understand the distribution of conserved reads in other species, miRDeepFinder offers the user the option to report the results of clustering of conserved short reads in other species (Supplementary 2-6).

Reads with lengths of 17~35 nt, and with at least a read count of 3, were mapped to the genome of *A. thaliana* resulting in more than 305,685 complete hits (100% identity). After analyzing hairpin structures of precursor candidates and removing sequences that overlapped coding genes and repeated sequences, a total of 1,545 miRNA candidates were detected (Supplementary 2-1); 110 of these reads had the same sequences known mature miRNAs from *Arabidopsis*. Of these 110 reads, 87 reads with more than 3 of read count correspond to 131 known miRNAs in *Arabidopsis* and 23 reads whose count is less than 3 correspond to 32 known miRNAs in *Arabidopsis*. Among 1,545 total miRNA candidates, 129 miRNAs were identified and matched at least 128 known miRNAs in *Arabidopsis*, accounting for 97.71% of the 131 known *Arabidopsis* miRNAs.

Although a miRNA and its miRNA* typically form a duplex with two nucleotide 3' overhangs (Meyers et al., 2008), only 67 of 131 (51.15%) known miRNAs were found

to have miRNA*s in the small RNA dataset. Meanwhile, only 32 of 1,416 (2.26%) miRNA candidates had a corresponding miRNA* present. Based on the criterion that a small RNA can be considered a functional miRNA only if its miRNA* also exists in a short read library, it is not possible to recover some known miRNAs, resulting in a high chance that some important miRNAs will not be identified. To address this issue, miRDeepFinder identified potential miRNA candidates even when no miRNA* is present. In addition, it is difficult to rule out a miRNA candidate when no miRNA* is found, because miRNAs have the feature of tissue-specificity (Bartel, 2004) and sequencing errors can result in no miRNA* being present in short reads library due to their low expression abundance. miRDeepFinder provides users the option of combining other small RNA datasets from NCBI to test whether an miRNA and its typical miRNA* co-exist in these datasets. We searched for the 1,416 novel miRNA candidates and their miRNA*s in the other 11 small RNA datasets. Only miRNAs and their miRNA*s that co-existed in at least 4 of the 12 datasets were kept for later experimental validation. In this experiment, 13 novel miRNA candidates were investigated.

qRT-PCR validation of newly identified miRNAs

According to results of qRT-PCR, expression was validated for 12 out of 13 novel miRNA candidates in 10-days-old seedlings of *Arabidopsis*. Based on the fact that the 12 miRNAs didn't occur in all small RNA datasets in *Arabidopsis*, we believe that these 12 miRNAs likely are expressed in a tissue- or development-specific manner. More interestingly, one of the 12 validated miRNAs spans the regions of two coding genes in the *Arabidopsis* genome and was named AC1 (Figure 2-3). Currently, nearly all of

bioinformatics approaches that analyze potential miRNA from a small RNA library discard short reads that map to coding regions in the genome. Although we also used the same rule in miRDeepFinder, out of curiosity we went ahead and analyzed short reads mapping to coding region of the *Arabidopsis* genome. Based on current knowledge, miRNAs are widely considered to be derived from non-coding genomic regions in plants. The precursor of AC1 is 655 nts in length (loci: 2,227,245 - 2,227,899) spanning two exons (locus: 2,225,963 - 2,227,402 and locus: 2,227,403 - 2,227,566) of UGT71C4 (UDP-GLUCOSYL TRANSFERASE 71C4) and one exon (locus: 2,227,594 - 2,229,319) of UGT71C3 (UDP-GLUCOSYL TRANSFERASE 71C3) (Figure 2-3B and 2-3C). Moreover, AC1 has an opposite strand direction with UGT71C4 and the same strand direction with UGT71C3. We also tested whether AC1 is conserved in other species; however, we found no homolog in any other plant, suggesting a relatively recent origin of AC1, and a potential new mechanism for the evolution of new miRNAs.

Target analysis

Our recently developed miRNA target prediction tool, Target-align (Xie and Zhang, 2010), also is integrated into miRDeepFinder for improved analysis of miRNA function. A total of 48 miRNAs, from 102 validated miRNA-target pairs in *Arabidopsis,* were employed to search against whole the *Arabidopsis* transcriptome (TAIR annotation version 9, http://www.arabidopsis.org/). The parameters for predicting miRNA targets were set as in our previous report (Xie and Zhang, 2010) with some minor modifications. We were able to detect a total of 1,317 miRNA-target pairs, which included 98 of 102 (96.07%) validated and 1,094 (false positives: 83.07%) invalidated miRNA-target pairs

(Supplementary 2-3). Compared with other miRNA target tools, like Srn target (Moxon et al., 2008), miRU (Zhang, 2005), and TAPIR (Bonnet et al., 2010), Target-align displays good specificity and lower rate of false positives.

Function classification and pathway enrichment

A total of 1,068 predicted target genes including 884 gene families (Supplementary 2-3) were used as BLASTX queries against the Gene Ontology (GO) protein database; 996 of 1,068 (93.26%) targets were confirmed exactly to the *Arabidopsis* GO annotation. Moreover, of the 1,068 targets, 41 had no overlap with the GO database and 32 were found to associate with GO annotations from other species. Based on a comparison with GO analysis results in the *Arabidopsis* Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/), a majority of the 73 targets that had no *Arabidopsis* GO annotation were marked with an unknown function classification or were not present at all in the GO annotation database.

In GO annotation, 1,541 molecular functions, 1,733 biological processes, and 881 cellular components were classified based on miRDeepFinder analysis; these GO annotations are incorporated into miRDeepFinder (Supplementary 2-4). All 48 tested miRNAs were involved in classification of molecular function, biological process, and cellular component. According to classification of molecular function and biological process, most predicted targets are involved in regulation of transcription, metabolism, stress response, and signal transduction (Supplementary 2-4). The result is consistent with current knowledge of miRNA functions (Ambros, 2004; Bartel, 2004). We built a relationship database scheme to manage pathway information from KEGG. Meanwhile,

based on GO protein annotations, we could track KEGG information attached to GO annotation of proteins. Finally, a combined search against the GO and KEGG databases was made to enrich pathways. The result showed that 27 miRNAs were enriched in 42 pathways from *Arabidopsis* (Supplementary 2-5). These pathways are involved in amino acid and nucleotide metabolism, starch and sucrose metabolism, oxidative phosphorylation, sulfur metabolism, ubiquitin mediated proteolysis, and plant-pathogen interactions. Apparently, functional classification and pathway enrichment can contribute to a better understanding of miRNA function in the gene regulation network.

Conclusion

miRDeepFinder provides a complete, systematic, and adaptive solution for analyzing small RNA datasets from deep sequence, from raw sequence processing to downstream deep functional analysis. Our results show that miRDeepFinder displays high sensitivity (97.71% accurate) in recovering known miRNAs. Moreover, to mine miRNAs more reliably, miRDeepFinder users can compare identified miRNAs with multiple small RNA datasets. Based on this approach, we successfully validated 12 novel miRNAs in *Arabidopsis* using stem-loop PCR and qRT-PCR. Our previously reported program Target-align also was incorporated into miRDeepFinder. According to testing of 120 validated miRNA-target pairs against the whole *Arabidopsis* transcriptome, miRDeepFinder demonstrated a 96.7% accuracy rate for target prediction. Furthermore, based on CleaveLand, miRDeepFinder allows users to fully utilize the RNA degradome data available from public databases for identification of miRNA target cleavage sites. In addition to Target-align, miRDeepFinder offers users one more miRNA target search

engine, targetFinder, which depends on FASTA alignment between miRNAs and target sequence candidates. A reconstructed KEEG database offers a good platform to make pathway enrichment with the GO database. Currently, the majority of tools available for analyzing small RNAs from deep sequencing data are based only on complete genomes. miRDeepFinder can be used with almost all data sources including complete genomes, EST, and GSS. Users are allowed to set nearly all parameters through the entire process of analyzing small RNA datasets. In conclusion, miRDeepFinder is a versatile new tool for analyzing small RNA datasets generated by next-generation sequencing technology.

Supplementary information

Supplementary 2-1: 13 newly identified miRNAs from Arabidopsis small RNA dataset http://link.springer.com/content/esm/art:10.1007/s11103-012-9885-2/file/MediaObjects/11103_2012_9885_MOESM1_ESM.xls

Supplementary 2-2: miRDeepFinder identified miRNAs and their reads from the deep sequencing datasets http://link.springer.com/content/esm/art:10.1007/s11103-012-9885-2/file/MediaObjects/11103_2012_9885_MOESM2_ESM.xls

Supplementary 2-3: miRDeepFinder identified miRNAs and their targets http://link.springer.com/content/esm/art:10.1007/s11103-012-9885-2/file/MediaObjects/11103_2012_9885_MOESM3_ESM.xls

Supplementary 2-4: GO analysis http://link.springer.com/content/esm/art:10.1007/s11103-012-9885-2/file/MediaObjects/11103_2012_9885_MOESM4_ESM.xls

Supplementary 2-5: KEGG analysis

http://link.springer.com/content/esm/art:10.1007/s11103-012-9885-2/file/MediaObjects/11103_2012_9885_MOESM5_ESM.xls

Supplementary 2-6: A total of 631 reads were identified as conserved miRNAs corresponding to 182 distinct miRNAs http://link.springer.com/content/esm/art:10.1007/s11103-012-9885-2/file/MediaObjects/11103_2012_9885_MOESM6_ESM.txt

Reference

Addo-Quaye, C., Miller, W., and Axtell, M.J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics *25*, 130-131.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350-355.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. The Plant cell *15*, 2730-2741.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. Bioinformatics *26*, 1566-1568.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. (2004). Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. Proc Natl Acad Sci U S A *101*, 11511-11516.

Chen, C., Ridzon, D.A., Broomer, A.J., Zhou, Z., Lee, D.H., Nguyen, J.T., Barbisin, M., Xu, N.L., Mahuvakar, V.R., Andersen, M.R.*, et al.* (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic acids research *33*, e179.

Chen, X. (2008). MicroRNA metabolism in plants. Curr Top Microbiol Immunol *320*, 117-136.

Ender, C., and Meister, G. (2010). Argonaute proteins at a glance. J Cell Sci *123*, 1819-1823.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L.*, et al.* (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PloS one *2*, e219.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol *26*, 407-415.

Fu, H., Tie, Y., Xu, C., Zhang, Z., Zhu, J., Shi, Y., Jiang, H., Sun, Z., and Zheng, X. (2005). Identification of human fetal liver miRNAs by a novel method. FEBS Lett *579*, 3849-3854.

Fujii, H., Chiou, T.J., Lin, S.I., Aung, K., and Zhu, J.K. (2005). A miRNA involved in

phosphate-starvation response in Arabidopsis. Curr Biol *15*, 2038-2043.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R.*, et al.* (2009). Rfam: updates to the RNA families database. Nucleic Acids Res *37*, D136-140.

Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G., and Kim, J. (2003). Computational and experimental identification of C. elegans microRNAs. Mol Cell *11*, 1253-1263.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res *32*, D109-111.

Guo, H.S., Xie, Q., Fei, J.F., and Chua, N.H. (2005). MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. The Plant cell *17*, 1376-1386.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M., and Aransay, A.M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res *37*, W68-76.

Herr, A.J. (2005). Pathways through the small RNA world of plants. FEBS Lett *579*, 5879-5888.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Research *31*, 3429-3431.

Hsieh, L.C., Lin, S.I., Shih, A.C., Chen, J.W., Lin, W.Y., Tseng, C.Y., Li, W.H., and Chiou, T.J. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiol *151*, 2120-2132.

Huang, P.J., Liu, Y.C., Lee, C.C., Lin, W.C., Gan, R.R., Lyu, P.C., and Tang, P. (2010). DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res *38 Suppl*, W385-

391.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.

Kim, V.N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol *6*, 376-385.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Mathelier, A., and Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics *26*, 2226-2234.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J.*, et al.* (2008). Criteria for annotation of plant MicroRNAs. The Plant cell *20*, 3186-3190.

Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D.J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics *24*, 2252-2253.

Pfeffer, S., Lagos-Quintana, M., and Tuschl, T. (2005). Cloning of small RNA molecules. Curr Protoc Mol Biol *Chapter 26*, Unit 26 24.

Pillai, R.S., Bhattacharyya, S.N., and Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? Trends Cell Biol *17*, 118-126.

Qiu, C.X., Xie, F.L., Zhu, Y.Y., Guo, K., Huang, S.Q., Nie, L., and Yang, Z.M. (2007). Computational identification of microRNAs and their targets in Gossypium hirsutum expressed sequence tags. Gene *395*, 49-61.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E.,

Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901-906.

Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of microRNAs on the plant transcriptome. Dev Cell *8*, 517-527.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Sunkar, R., Zhou, X., Zheng, Y., Zhang, W., and Zhu, J.K. (2008). Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC plant biology *8*, 25.

Wang, W.C., Lin, F.M., Chang, W.C., Lin, K.Y., Huang, H.D., and Lin, N.S. (2009). miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinformatics *10*, 328.

Xie, F., Frazier, T.P., and Zhang, B. (2010). Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (Panicum virgatum). Planta *232*, 417-434.

Xie, F., and Zhang, B. (2010). Target-align: a tool for plant microRNA target identification. Bioinformatics *26*, 3002-3003.

Xu, J., Li, C.X., Li, Y.S., Lv, J.Y., Ma, Y., Shao, T.T., Xu, L.D., Wang, Y.Y., Du, L., Zhang, Y.P*., et al.* (2010). MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. Nucleic Acids Res.

Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. (2006). Conservation and divergence of plant microRNA genes. Plant J *46*, 243-259.

Zhang, B., Pan, X., and Stellwag, E.J. (2008). Identification of soybean microRNAs and their targets. Planta *229*, 161-182.

Zhang, Y. (2005). miRU: an automated plant miRNA target prediction server. Nucleic Acids Res *33*, W701-704.

Zhu, E., Zhao, F., Xu, G., Hou, H., Zhou, L., Li, X., Sun, Z., and Wu, J. (2010). mirTools: microRNA profiling and discovery based on high-throughput sequencing. Nucleic Acids Res *38 Suppl*, W392-397.

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res *9*, 133-148.

Table 2-1. Reverse transcription primers of 13 novel miRNA candidates

miRNA	Mature sequence	RT-Primer sequence
	UUGUACAAAUUUA	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN1	AGUGUACG	TCGCACTGGATACGACCGTACA
	AGCGGUUAAGACG	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN2	GACAAAUCUGU	TCGCACTGGATACGACACAGAT
	AGCUAAGGAUUUG	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN3	CAUUCUCA	TCGCACTGGATACGACTGAGAA
	ACAUAUGAUCUGC	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN4	AUCUUUGCAUU	TCGCACTGGATACGACAATGCA
AN5	UUCGCACAAUUGG	GTCGTATCCAGTGCAGGGTCCGAGGTAT

	UCAUCGCG	TCGCACTGGATACGACCGCGAT
	GCCCAGCUUGAAA	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN6	AUCGGACG	TCGCACTGGATACGACCGTCCG
	CGGGAGAAGUGCG	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN7	GUGCGGUUCA	TCGCACTGGATACGACTGAACC
	CGGGCUUGGCAGA	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN8	AUCAGCGGGGA	TCGCACTGGATACGACTCCCCG
	GACGAAUUCUCGG	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN9	ACCCGGUCGAC	TCGCACTGGATACGACGTCGAC
	CAGAAAUGCAAUC	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN10	GGCCUGACUAU	TCGCACTGGATACGACATAGTC
	AAUGGAUCUGGCC	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN11	AAAGUUGAGGG	TCGCACTGGATACGACCCCTCA
	AUUCGACAAAGUG	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AN12	AAGGGUUU	TCGCACTGGATACGACAAACCC
	CCAUCAAAGAUCG	GTCGTATCCAGTGCAGGGTCCGAGGTAT
AC1	UACGGCUC	TCGCACTGGATACGACGAGCCG

Table 2-2. miRNA-specific forward primers of 13 novel miRNA candidates for qRT-PCR

miRNA	miRNA-specific primers
AN1	GCGGCGGTTGTACAAATTTAAG

AN2	GCGGCGGAGCGGTTAAGACGGAC
AN3	GCGGCGGAGCTAAGGATTTGC
AN4	GCGGCGGACATATGATCTGCATC
AN5	GCGGCGGTTCGCACAATTGGTC
AN6	GCGGCGGGCCCAGCTTGAAAATC
AN7	GCGGCGGCGGGAGAAGTGCGGTG
AN8	GCGGCGGCGGGCTTGGCAGAATC
AN9	GCGGCGGGACGAATTCTCGGAC
AN10	GCGGCGGCAGAAATGCAATCGG
AN11	GCGGCGGAATGGATCTGGCCAAAG
AN12	GCGGCGGATTCGACAAAGTGAAG
AC1	GCGGCGGCCATCAAAGATCGTAC

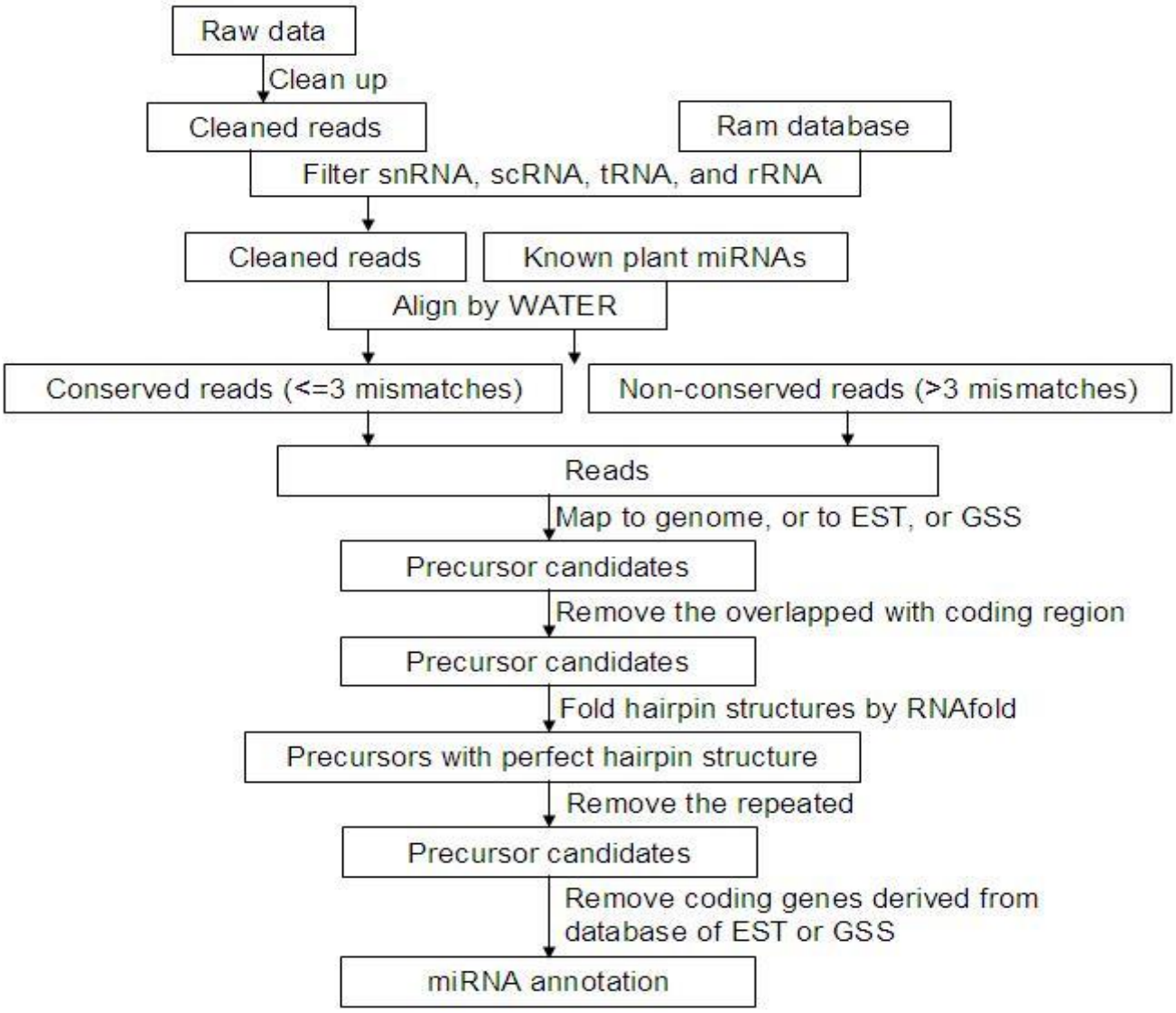


Figure 2-1. miRDeepFinder miRNA identification pipelines from small RNA dataset obtained from deep sequencing.

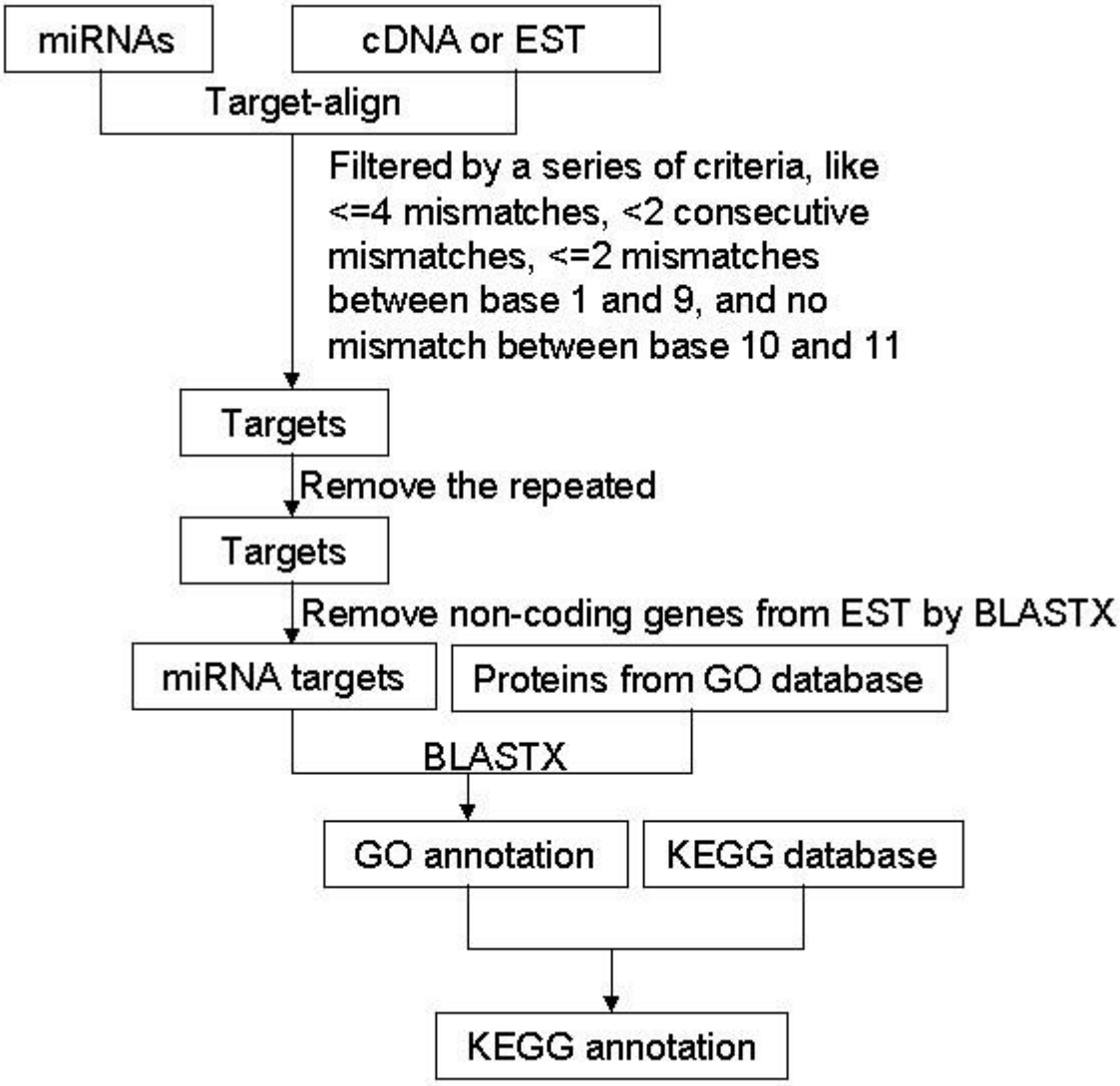
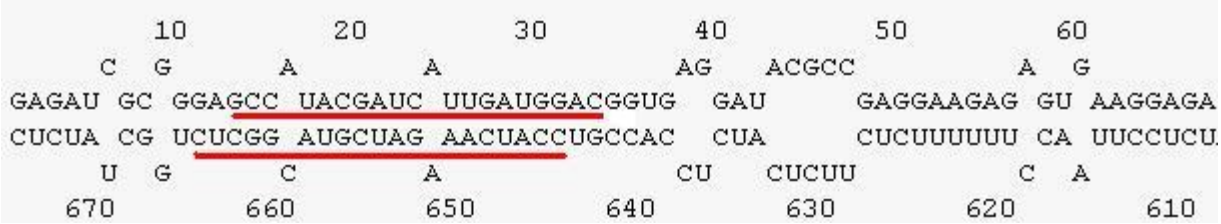


Figure 2-2. miReepFinder miRNA targets identification and their function annotation.

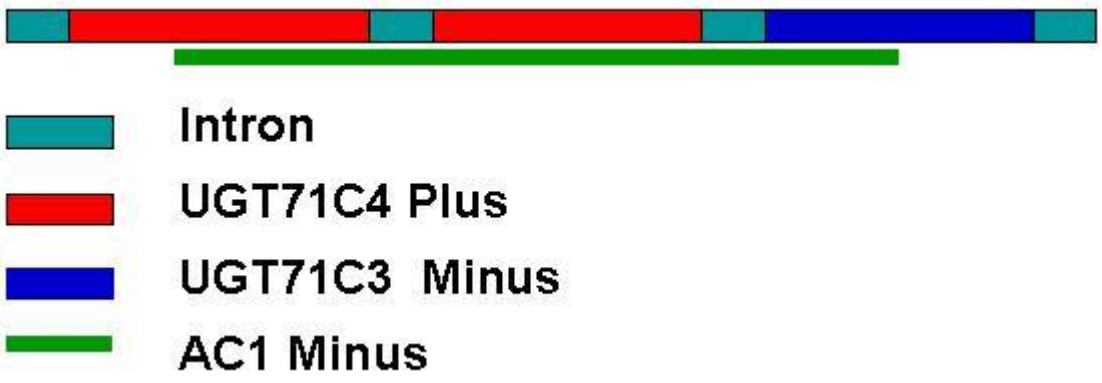
A.



B.

miRNA	Mature sequence
AC1 miRNA	CCATCAAAGATCGTACGGCTC
AN2 miRNA*	GCCATACGATCATTGATGGAC
Chromosome	1
miRNA start	2,227,245
miRNA end	2,227,265
Precursor start	2,227,245
Precursor end	2,227,899
Strand type	Minus
Precursor length	655 nt

C.



D.

Dataset	AC1 miRNA	AC1 miRNA*
GSM442932	167	5
GSM442933	233	6
GSM442934	14	1
GSM442935	12	1

Figure 2-3. miRNA AC1 and its miRNA*: A. Stem-loop hairpin structure of AC1 precursor; B. information on AC1 and its miRNA*; C. location of AC1 and its miRNA* in chromosome 1; D. read count distribution of AC1 and its miRNA* in four small RNA datasets.

CHAPTER 3: Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database

Abstract

A total of 28,432 unique contigs (25,371 in consensus contigs and 3,061 as singletons) were assembled from all 268,786 cotton ESTs currently available. Several in silico approaches [comparative genomics, Blast, Gene Ontology (GO) analysis, and pathway enrichment by Kyoto Encyclopedia of Genes and Genomes (KEGG)] were employed to investigate global functions of the cotton transcriptome. Cotton EST contigs were clustered into 5,461 groups with a maximum cluster size of 196 members. A total of 27,956 indel mutants and 149,616 single nucleotide polymorphisms (SNPs) were identified from consensus contigs. Interestingly, many contigs with significantly high frequencies of indels or SNPs encode transcription factors and protein kinases. In a comparison with six model plant species, cotton ESTs show the highest overall similarity to grape. A total of 87 cotton miRNAs were identified; 59 of these have not been reported previously from experimental or bioinformatics investigations. We also predicted 3,260 genes as miRNAs targets, which are associated with multiple biological functions, including stress response, metabolism, hormone signal transduction and fiber development. We identified 151 and 4,214 EST-simple sequence repeats (SSRs) from contigs and raw ESTs respectively. To make these data widely available, and to facilitate access to EST-related genetic information, we integrated our results into a comprehensive, fully downloadable web-based cotton EST database (www.leonxie.com).

Introduction

Cotton is among most important crops for natural textile fiber oilseed and is planted widely in 70 developed and developing countries, including the U.S., China, India, and Australia (IAC, 1996; Zhang et al., 2007). Although there are more than 50 species in the genus *Gossypium*, only four of them are cultivated; these are upland cotton (*Gossypium hirsutum L.*), sea-island cotton (*Gossypium barbadense*), Asian cotton (*Gossypium arboreum*), and Arabian cotton (*Gossypium herbaceum*). Upland cotton is, by far, the most widely planted, accounting for more than 95% of the annual cotton crop worldwide.

There are approximately 45 diploid (2n=2x=26) and five tetraploid (2n=4x=52) *Gossypium* species. Upland cotton has a complex allotetraploid genome (AADD, 2n=4x=52) (Chen et al., 2007b), with a haploid genome size estimated to be around 2.5 Gb (Hendrix and Stewart, 2005). Decoding the cotton genome is a crucial foundation for enhancing research on fiber development, quality, yield, and other important agronomic traits. Although some progress has been made on cotton genetics and agronomic improvement, sequencing of the complete cotton genome is still ongoing, largely because of its overall genetic and structural complexity (Chen et al., 2007b).

Currently, there are several types of cotton genomic resources available, including bacterial artificial chromosomes (BACs), expressed sequence tags (ESTs), linkage maps, and integrated genetic and physical maps (Chen et al., 2007b). To date, a total of 268,786 ESTs have been deposited in the public database GenBank. This large number of ESTs provides at least three obvious advantages: 1) broad EST coverage is a key landmark for future genome analysis and assembly (Seki et al., 1997); 2) ESTs can contribute to more efficient gene discovery and identification, especially from species with unavailable

genome sequences (Hattori et al., 2005); 3) ESTs provide information about gene expression, including tissue- and developmentally specific differences, as well as temporal responses to environmental changes (Zhang et al., 2007). Udall and co-workers previously assembled cotton ESTs using a total of 185,198 sequence reads from 30 cDNA libraries (Udall et al., 2006); however, it now is necessary to re-assemble cotton ESTs because there currently are 268,786 EST reads available. Furthermore, careful investigation of the likely functions of these assembled ESTs will be more important for enhancing cotton molecular genetics, for example, identifying useful new genetic markers.

One example of such genetic markers is simple sequence repeats (SSRs), also termed microsatellites, which are tandem repeats of two-to-six base-pair nucleotide motifs. They vary in length among different genotypes and offer a rich source of allelic polymorphisms. In contrast, SSR flanking sequences are often relatively conserved among genomes, making it possible to develop genetic markers for molecular breeding selection and genotype identification (Pearson and Sinden, 1998; Sanchez de la Hoz et al., 1996; Zeng et al., 2010). Compared with other types of molecular markers, SSRs have a number of advantages including co-dominant inheritance, high abundance, a generally random distribution across the genome, high information content, and reproducibility (Zeng et al., 2010). There are two classes of SSRs, those located in non-coding genomic regions and those found in ESTs. EST-SSRs generally are more conserved within and across related species and show higher transferability because more variable intron or intergenic sequences are absent from ESTs (Varshney et al., 2005). Additionally, it is more likely that EST-SSRs are tightly linked to specific gene functions and perhaps some even play a direct role in controlling important agronomic traits (Bozhko et al., 2003). Therefore, EST-SSRs

are good tools to facilitate marker-assisted selection (MAS) for breeding. To date, EST-SSRs have been used to screen cotton fiber-related loci from EST libraries generated from the cultivated diploid species *Gossypium arboreum L. cv* AKA8401 (Park et al., 2005).

Although it is possible to find polymorphic loci using EST-SSR markers, alone they are not sufficient for uncovering the underlying genetics of highly complex traits, such as disease resistance, yield, and quality, because of their low density of coverage across the genome. Furthermore, there are limited polymorphic SSR markers available to help in discriminating between closely related species (Wang et al., 2008). Single nucleotide polymorphisms (SNPs) are the most abundant type of DNA polymorphism in genomes. SNPs are alternative nucleotides present at a given, defined genetic location at a frequency exceeding 1% in a given population. Theoretically, each SNP can have four alleles, but bi-allelic variation has been shown to be the most frequent (Krawczak, 1999). SNPs are considered to be the major genetic source of phenotypic variability that differentiates individuals within given species (Nicolae et al., 2010). They have been applied extensively to genome-wide association studies (GWAS) of complex traits (Nicolae et al., 2010), fine mapping of QTLs (Zhang et al., 2009), and linkage disequilibrium-based association mapping (Schneider et al., 2007). Because ESTs are rich in current public databases, it is possible for EST-derived SNPs to be a low-cost and efficient resource for investigating genome-level variability before a draft cotton genome becomes available (Li et al., 2009; Wang et al., 2008).

MicroRNAs (miRNAs) are short non-coding RNA molecules that regulate protein-encoding gene expression at post-transcriptional levels. The main mechanisms of miRNA action are 1) promoting degradation and 2) inhibiting translation of their target mRNAs

(Bartel, 2004). Recently, several investigations have shown that translational inhibition is widespread in the plant kingdom (Bartel, 2004; Brodersen et al., 2008). In plants, primary miRNAs (pri-miRNA) are transcribed by RNA polymerase II from intergenic or intron regions and then folded into pre-miRNA hairpins. DICER-LIKE 1 (DCL1) directs conversion of pri-miRNAs to pre-miRNAs, and their processing into mature miRNAs. These steps mostly are carried out in the nucleus. Mature miRNA duplexes are stabilized by the S-adenosyl methionine-dependent methyltransferase Hua Enhancer 1 (HEN1) and are exported to the cytoplasm with the assistance of the plant homolog of exportin-5, HASTY (Voinnet, 2009). Mature miRNAs are generated by unbinding mature miRNA duplexes and then are loaded into the miRNA-induced silencing complex (miRISC). Integrated miRISC acts on a target message by perfect or near-perfect complementary base-pairing (Voinnet, 2009). In both plants and animals, many miRNA families are highly conserved through hundreds of million of years of evolution (Bartel, 2004). To date, miRNAs have been identified successfully from plant EST and GSS databases based on sequence conservation and characteristic miRNA features (Xie et al., 2007; Zhang et al., 2006a; Zhang et al., 2007). EST databases also provide evidence on temporal and developmental patterns of miRNA expression. ESTs are considered to be a reliable data source for prediction of miRNAs as well their targets, especially in those species without complete genome information (Xie et al., 2007; Zhang et al., 2006a; Zhang et al., 2007).

In this study, we performed global assembly of cotton ESTs available from NCBI, and functional annotation using BLASTx, BLASTn, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) resources. Using the contigs obtained, we also performed EST-based investigations of comparative transcriptome similarity between

cotton and other plant species, sequence polymorphisms, expressed miRNAs and their targets, and SSR analysis. Finally, we integrated these analytical data into a comprehensive web-based database so that EST-related information can be shared and queried publically.

Result and discussion

EST assembly

A total of 268,786 cotton ESTs were collected from NCBI; they have been obtained from different tissues, including fiber, ovule, anther, boll, callus, cotyledon, embryo, leaf, root, stem, seedling, and cultured cells (Table 2-1). The largest fraction of cotton ESTs is from fiber, with 114,167 sequences or 42.48% of all ESTs available. These ESTs were isolated from different treatments, including cold, cycloheximide, drought, aging, and *Fusariumoxysporum f. sp. vasinfectum* and *Xanthomonascampestris pv. Malvacearum* infections. After pre-processing raw sequences, a total of 235,328 clean ESTs were assembled into 28,432 unique genes (contigs) including 25,371 consensus contigs and 3,061 singletons. Contig lengths ranged from 101 to 4,080 nt (Figure 3-1). Consensus assemblies shared a similar sequence size distribution with singletons, except that few of the latter were found among longer length contigs. Most assembled contigs fell in the ranges from 500 nt to 900 nt (46.44%) or 900 nt to 1300 nt (26.76%) in length (Figure 3-1).

Annotation

Because a complete cotton genome is unavailable, it is difficult to determine precise CDS and protein sequences. Gene functions were annotated in two ways: BLASTx against

all plant reference proteins data and BLASTn against all plant reference nucleotide data Most ESTs were inferred to be homologous with at least one protein-coding gene counterpart in another plant species, including *Arabidopsis*, rice, maize or grape. However, 6,441 sequences (22.64% of assembled EST contigs and singletons) by BLASTx and 7,992 contigs by BLASTn (Table 2-2). In total, 4,043 contigs (14.22%) could not be annotated through BLAST searches. In addition, more than 60% of ESTs shared the same or similar annotation amongst BLASTx and BLASTn search results.

The 28,432 assembled cotton contigs were further annotated by BLASTx against the GO protein database, using an E-value cutoff of 1e-20, with 22,400 cotton ESTs finding a protein homolog. A total of 372 unique cellular component classes were identified for 13,657 ESTs (Figure 3-3A). According to annotation classification of GO database, the largest cellular component found for cotton ESTs was from cell part (6,810 contigs, 55%) and the smallest was from virion part (7 sequences, ~0%). We infer that ESTs associated with the virion part could result from contamination by virus mRNAs. A total of 13,964 ESTs were associated with 1,628 GO categories for biological processes. The majority of biological processes identified are involved in responses to stimuli (18%) and cellular process (17%) (Figure 3-3B). Furthermore, 15,378 ESTs were classified as involved in 1,407 molecular functions. The major molecular functions were associated with binding (57%) and catalytic (32%) activities (Figure 3-3C). Based on KEEG annotations from GO proteins, we made pathway enrichment analysis for cotton ESTs. This revealed 3,176 contigs to be involved in 271 different pathways (Supplementary 3-1).

Using BLASTn cutoffs for E-value (≤1e-30) and sequence identity (≥90%), a total of 5,461 gene clusters were identified from the entire set of 28,432 assembled cotton ESTs.

The sizes of clusters varied from two to 196 members with an average size of 3.62 (Figure 3-4). The majority of clusters (3,358 / 59.8%) had 2 members.

Genomic comparisons with other model plants

Based on comparisons with reference protein databases from six model species, *Arabidopsis thaliana, Chlamydomonas reinhardtii, Medicago truncatula, Oryza sativa, Vitis vinifera, Zea mays*, cotton contigs were shown to be the most similar overall to *Vitis,* followed by *Arabidposis* (Figure 3-5); like cotton, both of these species are dicots. Using a BLASTx E-value cutoff of 1e-30, 18,613 of 22,699 (82.0%) sequences from *Vitis* were found to be homologous with 19,688 of 28,432 (69.2%) cotton ESTs (Figure 3-5C), whereas 17,471 of 26,379 (66.2%) sequences from *Arabidposis* were similar to 18,529 of 28,432 (65.1%) cotton contigs (Figure 3-5D). Amongst the six model species, *Chlamydomonas* was identified as having the least overall similarity (31.4%) to cotton. These data generally agree with current views of plant evolution; however, the highest overall similarity of cotton sequences to *Vitis* is somewhat surprising. Molecular phylogenetic analyses place the Malvaceae (cotton) and Brassicaceae (*Arabidopsis*) as sister families, with the Vitaceae (*Vitis*) a more distant outgroup (Wang et al., 2009). The greater similarity between cotton and *Vitis* suggests that they retain somewhat more similar genome contents and sequence conservation from the common ancestor of all three taxa, than does *Arabidopsis*.

miRNAs and their targets in cotton

Because of the limited nucleotide sequence resources available, miRNA-related

research in cotton has lagged far behind other plant species. Currently, only 34 cotton miRNAs have been identified and deposited into the miRBase database (Griffiths-Jones et al., 2008). In this study, we used a total of 2,454 known plant miRNAs deposited in miRBase (Release 15) (Griffiths-Jones et al., 2008) as a reference set, and identified 87 miRNAs among cotton EST contigs and raw ESTs (Table 3-3). Of these, 59 were identified for the first time in cotton.

Of the 87 miRNAs identified, 33 were from our newly assembled contigs and 54 came directly from raw EST reads (Table 3-3). The length of the cotton miRNAs varied from 18 to 24 nt, with average of 20.3 ± 1.4 nt (Figure 3-6A). The most abundant cotton miRNAs were 21 nt in length. These results are similar to miRNA lengths reported previously in plants (Zhang et al., 2006c). The 87 miRNAs from cotton clustered into 57 families. The size of miRNA families in cotton varied from one to six sequence members (Table 3-3); 44 of 57 (77.2%) families had only one member (e.g., miR159, miR162, miR166, miR171, miR172, miR390, miR393, and miR395), whereas 13 (22.8%) had multiple members (e.g., miR156, miR164, miR394, miR398, miR399, miR414, and miR482) (Figure 2-6B). The largest miRNA families, including miRNA156, miRNA414, and miRNA1533, each with six members. Thirty-two of 87 miRNAs in cotton were obtained from the antisense strand of our original contig or EST, and the other 55 came from the sense strand (Table 3-3). miRNAs are located at either the 5' or 3' end of the hairpin arm. Our results show 50 of 87 miRNAs to be located at the 3' end and 37 at the 5' end.

Given that miRNAs target the transcripts of protein-encoding genes, a total of 18,621 ESTs, with E-values of less than 1e-25 in BLASTx searches against the plant

protein database, were selected as a subject dataset for target prediction. Based on a discrete set of criteria (see experimental procedures), 87 miRNAs identified in cotton were found to target a total of 3,260 protein-encoding genes (Supplementary 3-2). Our target prediction suggests that cotton miRNAs regulate the expression of many types of genes associated with diverse biological and metabolic processes, including metabolic pathways, hormone signal transduction, stress response, and fiber development. As in previous investigations, validated miRNA-target pairs also were identified in cotton, including miR156-squamosa promoter-binding protein (SBP) (Schwab et al., 2005), miR164-NAC domain protein (NAC) (Guo et al., 2005), miR398- Cu/Zn superoxide dismutase (Sunkar et al., 2006), miR172-AP2 domain-containing transcription factor (Aukerman and Sakai, 2003), and miR393-transport inhibitor response 1 (Schwab et al., 2005). In addition, because cotton is one of most important fiber crops, we also carefully examined targets associated with fiber development or fiber yield. Amongst the potential miRNA targets identified in cotton, there were at least 23 genes tightly associated with fiber development (Table 3-4). These targets control cellulose synthesis (miR156g and contig16368), fiber development (miR414b and contig7645), and glucose metabolism (miR529a and contig16806).

Sequence polymorphisms

We detected a total of 149,614 putative SNPs in 14,516 cotton contigs and 27,956 putative insertions/deletions (indels) in 8,674 contigs. Both SNPs and indels were detected in a total of 8,118 contigs. Our results show that SNPs occur once every 215 nt in cotton ESTs and indels occur once every 1,111 nt. The maximum frequencies of SNP and indels were 0.122 and 0.069 respectively. We generated a standard normal distribution to analyze

the frequencies of SNPs/indels among contigs, and determine which contigs had a significantly high number of SNPs at $P<0.05$ (significant) and $P<0.01$ (highly significant). We found 1,933 contigs to contain significant SNP frequencies, with 802 of these contigs at high significance. A significant frequency of indels was found for 1,089 contigs, 735 of which were highly significant. Currently, the genome of cotton is incompletely sequenced; in its absence, however, the large resource of ESTs available allow for identification of large numbers of SNPs (Wang et al., 2008). The apparently high frequency of SNPs and indels we observed in cotton ESTs could be due in part to sequencing errors. To address this issue, we followed the criteria of Wang and co-workers (Wang et al., 2008) to remove pseudo-SNPs and pseudo-indels as much as possible. Without experimental validation, however, it is difficult to determine whether a given SNP or an indel in cotton represents a real polymorphism. Nevertheless, we suggest that the high average frequency of SNPs we observed could, indeed, reflect real genetic variation resulting from the complicated genetic background present in large cotton EST libraries. However, because of the nature of cotton EST data in the NCBI database, it is not 100% sure that these SNPs are really SNPs or caused by sequencing errors. As deep sequencing technology become available, more study may be performed to investigate this issue.

Aside from those that could not be assigned a presumed function, many cotton EST contigs with significant rate of SNPs and indels are associated with transcription factors, energy metabolism, stress response, signal transduction, and protein kinases (see supplementary 3-3). A previous investigation showed that high SNP frequency (0.013) occurred in R2R3-MYB transcription factors from cotton (An et al., 2008). In this study, we also detected two contigs (contig2733 and contig15263) annotated to encode MYB

transcription factors that have significantly high SNP frequencies. Therefore, it is possible that the high diversity of SNPs and indels in the cotton transcriptome could be related to functional adaptations to environmental stress.

Simple sequence repeats

Because of their relative abundance and ease of generation, SSRs are among the most powerful of molecular markers, and have been applied widely in molecular-assisted selection (MAS) for plant breeding programs (Kantartzi et al., 2009). SSR markers derived from expressed sequence tags (EST-SSRs) originate from transcribed regions of the genome and are likely to be even more transferable across lines, populations and species than random genomic SSRs (Park et al., 2005). In this study, we analyzed SSRs in both cotton contigs and raw ESTs. We identified a total of 151 SSRs from cotton contigs and 4,214 from raw ESTs (see Supplementary 3-4). Among SSRs from contigs, the most abundant repeat types were trinucleotides (130, 86.09%) followed by dinucleotides (21, 13.91%). The dominant sequence repeat in contigs was AAG / CTT (10, 6.62%) followed by TGA / TCA (9, 5.96%). Trinucleotide repeats also were the most common among SSRs from raw ESTs (2,961, 70.27%) again followed by dinucleotides (829, 19.67%) along with a sizeable fraction of tetranucleotides (424, 10.06%). Dominant repeat types in raw ESTs were GAA / TTC (159, 3.77%) and GAT / ATC (159, 3.77%). Amongst the 151 SSR markers found, only 43 come from the contigs annotated with known functions. Potentially, these markers could be exploited for use in marker-assist breeding selection. Of these SSRs, 51 from contigs and 1,663 from raw ESTs have not been reported previously in cotton.

In further investigate the potential of these SSR repeats as genetic markers, we

employed eprimer3 (primer 3) to design primer pairs for each SSR under a series of primer-designing parameters (see Experimental procedures). We were able to find viable primer pairs for 121 of 151 contig SSRs and 3,092 of 4,214 raw EST SSRs (all these primers can be downloaded from the cotton EST website www.leonxie.com).

Web-based database for cotton ESTs

To facilitate further investigation and application of cotton genome-related research, we constructed a web-based, searchable and downloadable database for managing cotton ESTs data, along with related deep sequence analyses including assembly, annotation, miRNAs, SNP and indels, and SSRs (Figure 3-2). This database can be accessed freely through a web interface (www.l eonxie.com). Raw ESTs, as well as annotation and assembly data can be queried using different strategies, such as gene accession, gene ID, and function (Figure 3-7). We also incorporated the Cotton Marker Database (CMD) into our web-server and built connections with raw EST, assembled contigs, and SSR databases. In this way, users can quickly access marker information from cotton ESTs or access marker-related ESTs through CMD markers. We have attempted to develop a seamless connection among all of these cotton EST datasets and resources. For instance, when investigating a contig, users can visit its related information, including functional annotation, miRNA, SSR, SNP, GO, and KEGG; alternatively that contig can be accessed from any one of the related resources as a starting point. To improve the efficiency of BLAST analyses of cotton ESTs, we also built a local WWW-BLAST server permitting directed and advanced BLAST options. Raw cotton ESTs, assembled contigs, consensus assemblies, singletons, all reference protein databases from plants, and all reference plant

nucleotide databases are incorporated within our local WWW-BLAST server as potential query targets. Furthermore, EST data and related analytical tools and results, all can be freely accessed and downloaded.

Conclusions

We have developed a specific and dedicated workbench for assembling cotton ESTs and for performing genome-wide analyses of the cotton transcriptome. In addition to raw ESTs and assembled contigs, additional EST-related information, including miRNAs, SNPs, and SSRs has been integrated into this database. A friendly web-interface allows users to access and download these data as batch files or via directed searches based on specific interests and needs. Moreover, now that this platform for cotton EST data has been established, it will be very convenient to add new cotton ESTs and annotated resources to our database in future. Therefore, this cotton EST database can contribute significantly to advancing research on cotton ESTs and global genome-wide analyses.

Experimental procedures

Dataset

A total of 268,786 cotton ESTs (*Gossypium hirsutum L.*) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). These ESTs were obtained from at least 90 EST libraries and samples treated under at least eight different abiotic and biotic conditions.

Data pre-processing

A majority of raw EST sequences potentially contain various contaminating

elements, such as sequencing primers, vector sequence, sequences from other species, and sequencing errors. In addition, poly A / T tail and low complexity sequences are inevitably present in some raw ESTs. Thus, a critical first step is to remove these contaminated sequences before performing more deep analysis. In this study, we first cleaned original cotton ESTs by Seqclean (Chen et al., 2007a) (ftp://ftp.tigr.org/pub/software/tgi/seqclean/) from TIGR under default parameters. Seqclean is a versatile tool for removing sequences from vectors, mitochondria, ribosomal RNAs, sequencing primers, polyA / T tails, low complexity sequences, and sequences with lengths under 100 nt (Chen et al., 2007a). After processing with SeqClean (Figure 3-2), we employed RepeatMasker (version 3.2.9, http://www.repeatmasker.org/) to mask repeated elements based on Repbase (Repbase 15.04, http://www.girinst.org/) (Jurka, 1998). Finally, a total of 235,328 cleaned ESTs were kept for further assembly.

EST clustering and assembling

The cleaned EST sequences were clustered and assembled into contigs (consensus and singletons) by TGICL (ftp://ftp.tigr.org/pub/software/tgi/tgicl/) (Pertea et al., 2003), which could partition the input dataset into small groups of sequences (clusters) using Megablast and assemble each cluster by using the cap3 program (Huang and Madan, 1999) into contigs. The resulted data was further performed an ortholog search against the published assembled data of Gossypium's ESTs (http://www.agcol.arizona.edu/cgi-bin/pave/Cotton/index.cgi) (Udall et al., 2006) using Orthomcl (Version 2.0, http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=orthomcl#Software) under the cutoff of E-value of 1e-25 and identify of 95%.

Functional annotation

In order to investigate putative functions of cotton ESTs, we performed BLASTx (Altschul et al., 1997) against reference protein databases from all plants using an E-value cutoff of 1e-20, and BLASTn against reference nucleotide acid databases from all plants at an E-value cutoff of 1e-25. Only the best high-scoring segment pair (HSP) was kept for annotation. We also tried to annotate possible open reading frames (ORFs) of contigs and further infer their protein sequences by GETORF from Emboss tools package (http://emboss.sourceforge.net/). The longest ORF was considered to be the candidate CDS sequence, and its translation the presumed protein sequence as well.

To better understand the functional classification of ESTs, contigs were used as queries in BLASTx using Gene Ontology (GO) analysis (Ashburner et al., 2000). Cellular component, biological process, and molecular function were classified for these contigs. We performed further pathway enrichment according to GO annotations for Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000).

Cluster analysis

Each individual contig was queried against the complete assembled EST data set using BLASTn. All contigs hit by the query with an E-value of less than 1e-30 and an identity of more than 90% were defined as a cluster.

Overall genomic sequence similarity

Using different BLASTx E-value cutoffs (E≤1e-10, E≤1e-30, E≤1e-50, and E≤1e-

100), we investigated sequence similarity between the cotton contigs we obtained and reference cDNA databases from several model species; these included *Arabidposis thaliana* (TAIR9, ftp://ftp.arabidopsis.org/Sequences/blast_datasets/TAIR9_blastsets/), *Chlamydomonas reinhardtii* (Chlre4, http://genome.jgi-psf.org/chlamy/chlamy.download.ftp.html), *Medicago truncatula* (Mt3.0 release, http://www.medicago.org/genome/downloads.php), *Vitis vinifera* (ftp://ftp.ncbi.nih.gov/genomes/Vitis_vinifera/Assembled_chromosomes/), *Zea mays* (http://www.plantgdb.org/ZmGDB/cgi-bin/downloadGDB.pl), and *Oryza Sativa* (version 6.1, ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.1/all.dir/).

Sequence polymorphism analysis

Based on assembly results of consensus contigs, SNP and indel polymorphisms were analyzed. A perl script was developed to detect SNPs and indels under several criteria as described by Wang and co-workers (Wang et al., 2008). Briefly, 1) a mismatch identified within contigs containing more than four individual EST reads was definable as a SNP or an indel; 2) variation among sequences was considered to be a bona fide SNP or indel polymorphism when it was found at least twice within contigs assembled by 5-6 ESTs; 3) at least three times within contigs assembled by 7-8 ESTs; 4) at least four times within contigs assembled by 9-12 ESTs; 5) and at least five times within contigs assembled by 13 or more ESTs.

Identification of miRNAs and their targets

MicroRNAs (miRNAs) are known as a class of none-coding endogenous small RNA molecules with lengths of ~21 nt. Investigations increasingly show that miRNAs regulate target mRNAs either by inducing their degradation or by inhibiting translation (Bartel, 2004). To date, miRNAs have been predicted successfully from various EST (Zhang et al., 2006b) and GSS databases (Zhang et al., 2006a). Especially for those species without complete genome information, an EST database is considered to be an ideal data source for predicting miRNAs their targets as well (Venne et al., 2008; Xie et al., 2007). In our analysis, low complexity sequences, sequences with lengths of less than 100 nt, and sequences with repeated elements were removed in data pre-processing; EST contigs generated and raw ESTs then were combined as the subject dataset. We employed all known plant miRNAs from miRBase (Release 15: April 2010, http://www.mirbase.org/) (Griffiths-Jones et al., 2008) as a reference set and performed homology searches against the subject dataset using methods reported previously (Xie et al., 2010; Zhang et al., 2005). Cotton miRNA targets also were predicted according to method in previous reports (Xie et al., 2010).

SSR detection and primer design

In order to locate simple sequence repeats (SSRs) in cotton ESTs, we performed SSR analyses on cotton contigs and raw ESTs using a software SSR Finder from GRAMENE (ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl). The parameters were designed for identifying perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of 6, 5, 4, 4, and 4 repeats respectively (Zeng et al., 2010). Eprimer3 from

EMBOSS bioinformatics software packages (http://emboss.sourceforge.net/) (Rychlik, 1995) was used to design flanking primers for detected microsatellites. The major parameters for primer design were set as following: PCR products ranging from 100 to 300 nt; primer lengths ranging from 18 to 24 nt with an optimum of 20 nt, 60ºC optimal annealing temperature, and GC content from 40%~65% with an optimum of 50% (Zeng et al., 2010).

Construction of a web-based cotton EST database

In order to share our integrated data and analytical results on cotton ESTs, including raw ESTs, assembled EST contigs, predicted miRNAs, sequence polymorphisms, and SSRs and primers, we integrated the information from each step of our investigation into a web-based cotton EST database, using open-source software (Apache, PHP, and MySQL), and constructed interfaces among the data types (Figure 3-2). Furthermore, to facilitate access to potentially useful markers from cotton raw ESTs and assembled contigs, we incorporated current data (SSR and QTL) from the Cotton Marker Database (CMD) (http://www.cottonmarker.org/) into our EST database. Our new web-based cotton EST database provides users with a friendly interface to query or download data. It is freely available at the website www.leonxie.com.

Supplementary information

Supplementary 3-1: Pathway analysis by KEGG http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0026980.s001

Supplementary 3-2: Predicted miRNA targets http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0026980.s002

Supplementary 3-3: Cotton EST contigs with significant SNPs and indels http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0026980.s003

Supplementary 3-4: Identified SSR markers with designed primers http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0026980.s004

Reference

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

An, C., Saha, S., Jenkins, J.N., Ma, D.P., Scheffler, B.E., Kohel, R.J., Yu, J.Z., and Stelly, D.M. (2008). Cotton (Gossypium spp.) R2R3-MYB transcription factors SNP identification, phylogenomic characterization, chromosome localization, and linkage mapping. Theor Appl Genet *116*, 1015-1026.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene Ontology: tool for the unification of biology. Nature Genet *25*, 25-29.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. The Plant cell *15*, 2730-2741.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bozhko, M., Riegel, R., Schubert, R., and Muller-Starck, G. (2003). A cyclophilin gene marker confirming geographical differentiation of Norway spruce populations and indicating viability response on excess soil-born salinity. Mol Ecol *12*, 3147-3155.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science *320*, 1185-1190.

Chen, Y.A., Lin, C.C., Wang, C.D., Wu, H.B., and Hwang, P.I. (2007a). An optimized

procedure greatly improves EST vector contamination removal. BMC Genomics *8*, 11.

Chen, Z.J., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X., Stelly, D.M., Rabinowicz, P.D., Town, C.D., *et al.* (2007b). Toward sequencing cotton (Gossypium) genomes. Plant physiology *145*, 1303-1310.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Research *36*, D154-D158.

Guo, H.S., Xie, Q., Fei, J.F., and Chua, N.H. (2005). MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. The Plant cell *17*, 1376-1386.

Hattori, J., Ouellet, T., and Tinker, N.A. (2005). Wheat EST sequence assembly facilitates comparison of gene contents among plant species and discovery of novel genes. Genome *48*, 197-206.

Hendrix, B., and Stewart, J.M. (2005). Estimation of the nuclear DNA content of gossypium species. Ann Bot *95*, 789-797.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res *9*, 868-877.

IAC (1996). Cotton: Review of World Situation, Monogram by International Advisory Committee. Washington, D.C.

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol *8*, 333-337.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *28*, 27-30.

Kantartzi, S.K., Ulloa, M., Sacks, E., and Stewart, J.M. (2009). Assessing genetic diversity

in Gossypium arboreum L. cultivars using genomic and EST-derived microsatellites. Genetica *136*, 141-147.

Krawczak, M. (1999). Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis *20*, 1676-1681.

Li, F., Kitashiba, H., Inaba, K., and Nishio, T. (2009). A Brassica rapa linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits. DNA Res *16*, 311-323.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet *6*, e1000888.

Park, Y.H., Alabady, M.S., Ulloa, M., Sickler, B., Wilkins, T.A., Yu, J., Stelly, D.M., Kohel, R.J., el-Shihy, O.M., and Cantrell, R.G. (2005). Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Mol Genet Genomics *274*, 428-441.

Pearson, C.E., and Sinden, R.R. (1998). Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. Curr Opin Struct Biol *8*, 321-330.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., *et al.* (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics *19*, 651-652.

Rychlik, W. (1995). Selection of primers for polymerase chain reaction. Mol Biotechnol *3*, 129-134.

Sanchez de la Hoz, M.P., Davila, J.A., Loarce, Y., and Ferrer, E. (1996). Simple sequence

repeat primers used in polymerase chain reaction amplifications to study genetic diversity in barley. Genome *39*, 112-117.

Schneider, K., Kulosa, D., Soerensen, T.R., Mohring, S., Heine, M., Durstewitz, G., Polley, A., Weber, E., Jamsari, Lein, J.*, et al.* (2007). Analysis of DNA polymorphisms in sugar beet (Beta vulgaris L.) and development of an SNP-based map of expressed genes. Theor Appl Genet *115*, 601-615.

Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of microRNAs on the plant transcriptome. Dev Cell *8*, 517-527.

Seki, M., Hayashida, N., Kato, N., Yohda, M., and Shinozaki, K. (1997). Rapid construction of a transcription map for a cosmid contig of Arabidopsis thaliana genome using a novel cDNA selection method. Plant J *12*, 481-487.

Sunkar, R., Kapoor, A., and Zhu, J.K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. Plant Cell *18*, 2051-2065.

Udall, J.A., Swanson, J.M., Haller, K., Rapp, R.A., Sparks, M.E., Hatfield, J., Yu, Y., Wu, Y., Dowd, C., Arpat, A.B.*, et al.* (2006). A global assembly of cotton ESTs. Genome Res *16*, 441-450.

Varshney, R.K., Graner, A., and Sorrells, M.E. (2005). Genic microsatellite markers in plants: features and applications. Trends Biotechnol *23*, 48-55.

Venne, L.S., Anderson, T.A., Zhang, B., Smith, L.M., and McMurry, S.T. (2008). Organochlorine pesticide concentrations in sediment and amphibian tissue in playa wetlands in the Southern High Plains, USA. Bulletin of Environmental Contamination and Toxicology *80*, 497-501.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell *136*, 669-687.

Wang, H.C., Moore, M.J., Soltis, P.S., Bell, C.D., Brockington, S.F., Alexandre, R., Davis, C.C., Latvis, M., Manchester, S.R., and Soltis, D.E. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A *106*, 3853-3858.

Wang, S., Sha, Z., Sonstegard, T.S., Liu, H., Xu, P., Somridhivej, B., Peatman, E., Kucuktas, H., and Liu, Z. (2008). Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics *9*, 450.

Xie, F., Frazier, T.P., and Zhang, B. (2010). Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (Panicum virgatum). Planta *232*, 417-434.

Xie, F.L., Huang, S.Q., Guo, K., Xiang, A.L., Zhu, Y.Y., Nie, L., and Yang, Z.M. (2007). Computational identification of novel microRNAs and targets in Brassica napus. FEBS Lett *581*, 1464-1474.

Zeng, S., Xiao, G., Guo, J., Fei, Z., Xu, Y., Roe, B.A., and Wang, Y. (2010). Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, Epimedium sagittatum (Sieb. Et Zucc.) Maxim. BMC Genomics *11*, 94.

Zhang, B., Pan, X., and Anderson, T.A. (2006a). Identification of 188 conserved maize microRNAs and their targets. FEBS Lett *580*, 3753-3762.

Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. (2006b). Conservation and divergence of plant microRNA genes. Plant J *46*, 243-259.

Zhang, B.H., Pan, X.P., Cox, S.B., Cobb, G.P., and Anderson, T.A. (2006c). Evidence that miRNAs are different from other RNAs. Cellular and Molecular Life Sciences *63*, 246-

254.

Zhang, B.H., Pan, X.P., Wang, Q.L., Cobb, G.P., and Anderson, T.A. (2005). Identification and characterization of new plant microRNAs using EST analysis. Cell Research *15*, 336-360.

Zhang, B.H., Wang, Q.L., Wang, K.B., Pan, X.P., Liu, F., Guo, T.L., Cobb, G.P., and Anderson, T.A. (2007). Identification of cotton microRNAs and their targets. Gene *397*, 26-37.

Zhang, J., Lu, Y., Yuan, Y., Zhang, X., Geng, J., Chen, Y., Cloutier, S., McVetty, P.B., and Li, G. (2009). Map-based cloning and characterization of a gene controlling hairiness and seed coat color traits in Brassica rapa. Plant Mol Biol *69*, 553-563.

Table 3-1. Distribution of sources of raw cotton ESTs from different tissues

EST library	Count of EST
Anther	51
Boll	5,387
Callus	242
Cell	4
Cotyledon	2,444
Embryo	509
Fiber	114,167
Fiber/Embryo	113
Fiber/Ovule	16,861
Leaf	6,675
Meristematic	44,615
Ovule	53,499
Protoplast	210
Root	6,003
Hypocotyl tissues	1,014
Seedling	2,468
Stem	14,482
Other	42

Table 3-2. Coding and non-coding contigs inferred by BLASTx and BLASTn

Method	Coding count	Coding %	Non-coding Count	Non-coding %
BLASTx	21,991	77.35	6,441	22.65
BLASTn	20,510	72.14	7,922	27.86
Common	16,124	56.71	4,043	14.22

Table 3-3. 87 miRNAs identified in cotton ESTs

miRNA	Family	Mature sequence	LM*	Strand	Location	GC%	MFE	MFEI	EST Id	Data Type#
ghr-miR156d	156	UGACAGAAGAGAGUGAGCAC	20	-	5'	51.81	54	1.26	contig21398	Predicted
ghr-miR156e	156	UGAAGAAAGACAGAGCAU	18	-	5'	39.14	94.3	0.58	contig18605	Predicted
ghr-miR156f	156	UGAAGAAGAAAGAGAGCAU	19	+	5'	36.62	24.9	0.96	EV488115	Predicted
ghr-miR156g	156	UGAAGAAGAAAGAGAGAAG	19	+	3'	33.8	16	0.67	DW508826	Predicted
ghr-miR156h	156	UGAAGAAUAGAGCGAUCAC	19	+	3'	51.28	121.63	0.55	EV491219	Predicted
ghr-miR156i	156	UGAAGACCAGAGUGAGCAC	19	-	5'	41.47	79.5	0.64	AJ513999	Predicted
ghr-miR159	159	UUUGGAUUGGAGGGAGCUCUA	21	+	3'	47.02	72.7	0.92	ES824206	Predicted
ghr-miR162a	162	UCGAUAAACCUCUGCAUCCAG	21	+	3'	42.86	35.4	0.91	DW493971	Predicted
ghr-miR164	164	UGGAGAAGCAGGGCACGUGCA	21	-	5'	50.77	38.3	1.16	DR461140	Validated
ghr-miR164b	164	UGGAGAACAUGGGCACAUGGU	21	+	5'	37.52	138.1	0.72	contig25636	Predicted
ghr-miR164d	164	UGGAAAGCGGGCAGUGAG	18	-	3'	56.26	174.4	0.66	AJ514172	Predicted
ghr-miR166b	166	UCGGACCAGGCUUCAUUCCCC	21	+	3'	43.54	61.49	0.96	DW502146	Predicted
ghr-miR169	169	AAGCCAAGAAUGAAUUGCCUG	21	-	5'	51.47	65.5	0.62	DW509134	Predicted
ghr-miR171	171	AGAUUGAGCCGCGCCAAUAUC	21	+	3'	43.53	37.8	1.02	DW507416	Predicted
ghr-miR172	172	AGAAUCCUGAUGAUGCUGCAG	21	+	3'	34.74	38.21	1.16	ES839084	Validated
ghr-miR390a,c	390	AAGCUCAGGAGGGAUAGCGCC	21	+	3'	42.86	40.2	0.96	contig17644	Predicted
ghr-miR393	393	UCCAAAGGGAUCGCAUUGAUCU	22	+	5'	38.66	45	0.98	ES827656	Validated
ghr-miR394a	394	UUGGCAUUCUGUCCACCUCC	20	+	5'	48.19	35	0.88	ES802173	Validated
ghr-miR394b	394	UUGGCAUUCUGUCCACCUCC	20	+	5'	40.21	28.52	0.73	DW517361	Validated
ghr-miR395	395	CUGAAGUGUUUGGGGGAACUC	21	+	3'	52.94	55	1.02	DW501342	Predicted
ghr-miR396a,b	396	UUCCACAGCUUUCUUGAACUG	21	+	5'	40	43.3	0.94	contig21626	Predicted
ghr-miR398	398	UGUGUUCUCAGGUCACCCCUU	21	+	3'	50.75	32.1	0.94	DW498056	Validated
ghr-miR398b	398	UGUUUAUCAGGCACCCCUU	19	+	5'	49.15	12	0.41	contig28115	Predicted
ghr-miR399c	399	UGCCAAAGGAGAGUUGGCCUU	21	+	3'	47.3	31.7	0.91	DW510913	Validated
ghr-miR399d	399	UGCCAAAGGAGAUUUGCCCUG	21	+	3'	41.56	39.1	1.22	DW509341	Validated
ghr-miR399e	399	UGCCAAAGGUGCUGCUCUU	19	-	3'	57.35	28	0.72	contig21507	Predicted
ghr-miR408	408	UGCUCGCCUCAUCCUCUCU	19	+	5'	43.84	115.99	0.65	DR454452	Predicted
ghr-miR413	413	CUGGUUUCACUUGCUCUGAAC	21	+	3'	43.38	45.52	0.77	DW504189	Predicted
ghr-miR414a	414	GCAUCUUCAUCUUCAUCUUCA	21	+	3'	37.43	183.79	0.59	contig20173	Predicted
ghr-miR414b	414	UCAUCUUCUUCAUCAUCUUCG	21	-	5'	49.63	97	0.72	contig17531	Predicted
ghr-miR414c	414	UCAUCAUCAUCAUCACCUUCA	21	+	3'	46.51	29.9	0.75	contig20222	Predicted
ghr-miR414d	414	CCAUCUUCAUCAUCAUCAUCA	21	-	5'	48.82	76.7	0.62	ES799840	Predicted

ghr-miR414e	414	UCUCCUUCAUCAUCAUCGUCA	21	-	3'	44.33	14.7	0.34	DW502456	Predicted
ghr-miR414f	414	UCAUUUUCAUCAUCAUCGUCA	21	-	5'	42.74	48.85	0.47	ES835113	Predicted
ghr-miR414g	444	UGCAGUUGUUGUCUAUGCCU	20	-	5'	42.64	32.1	0.58	AJ513351	Predicted
ghr-miR479	479	CGUGAUAUUGGUUCGGCUCAUC	22	+	5'	37.88	32.6	1.3	ES809290	Validated
ghr-miR482a	482	UCUUUCCUACUCCUCCCAUACC	22	+	3'	40	33.5	0.99	DR457519	Validated
ghr-miR482b	482	UCUUGCCUACUCCACCCAUGCC	22	+	3'	46.94	43.9	0.95	DT527030	Validated
ghr-miR482c	482	CCUCCUCCUCUCCAUUGC	18	+	3'	50.26	70.7	0.72	ES808713	Predicted
ghr-miR482d	482	UCUUCUUCUUCCUCCCAUC	19	-	3'	52.44	32.7	0.76	DT464811	Predicted
ghr-miR528	528	UGGAAGGGNGCAUGCAUGGAG	21	+	3'	34.41	43.7	0.68	DN804697	Predicted
ghr-miR529a	529	AGAAGGAGAGAGUCAACUU	19	+	3'	39.22	11.8	0.59	contig4544	Predicted
ghr-miR529b	529	UUUUCCCCUCUCUCUUCUUC	20	+	5'	42.06	33.86	0.64	contig26549	Predicted
ghr-miR529c	529	CUGUACUCGCUCUCUUCAUC	20	-	3'	48.44	114.3	0.61	DT046423	Predicted
ghr-miR530	530	UGCAUUUGCAAUCUGCUCCUA	21	+	3'	41.27	20.9	0.8	contig16357	Predicted
ghr-miR808	808	AUGAAUGUGGGAAAUGCUAGAA	22	-	3'	29.79	56.9	2.03	EX172412	Predicted
ghr-miR827a,b,c	827	UUAGAUGACCAUCAACAAACA	21	+	3'	37.4	39.2	0.85	contig22556	Validated
ghr-miR835	835	UUCUUCAUUGUUCUUUCUC	19	+	5'	36.78	57.94	0.6	DW506095	Predicted
ghr-miR838a	838	UUUUCUUCUCCUUCUUUACA	20	+	3'	42.7	27.2	0.72	DW516621	Predicted
ghr-miR838b	838	UUUUCUUCUACUUCUAGCAUU	21	-	5'	44.26	54.4	0.67	DW476363	Predicted
ghr-miR847a	847	UCACUCCUUUCCUUGAUG	18	-	3'	32.94	17.5	0.63	contig27404	Predicted
ghr-miR847b	847	UCACUCUCUUCUUUUGUUG	19	-	3'	36.21	13.65	0.65	contig23150	Predicted
ghr-miR855	855	AGGAAAAGAAAGGAAAAGGAA	21	-	3'	42.76	118.7	0.64	CO499070	Predicted
ghr-miR1132a	1132	GAUUAGGGACGGAAGGAG	18	+	5'	47.26	69.4	0.73	contig11460	Predicted
ghr-miR1132b	1132	CAUUAUGGCCAGAAGGAG	18	-	5'	49.8	85.4	0.67	contig26869	Predicted
ghr-miR1134	1134	UAACAACAACAAGAAGAAGGAGCU	24	+	5'	40.63	46.8	0.6	contig18889	Predicted
ghr-miR1144	1144	UGGAACCGUGGCAGGAGGAG	20	-	3'	62.96	76.6	0.75	contig5195	Predicted
ghr-miR1161	1161	UACUGGAGUUCUCAAGAAA	19	-	3'	32.73	14.6	0.81	DV849247	Predicted
ghr-miR1444	1444	UCCACAUUGGGUAAUGGUC	19	+	3'	33.67	68.1	1.03	contig21923	Predicted
ghr-miR1507	1507	UCUCUUCCAUGCAUCUUCUGA	21	-	3'	40.45	28.5	0.79	DT048287	Predicted
ghr-miR1509	1509	UUAAUGUAAAAAUACGGUG	19	-	3'	22.67	8.4	0.49	contig12637	Predicted
ghr-miR1533a	1533	AUAAUAAAAAGAAAAGGA	18	+	5'	27.05	25.6	0.78	contig21520	Predicted
ghr-miR1533b	1533	CUAAUAAUAAUAAUAAUGU	19	+	3'	20.69	5.87	0.49	contig15142	Predicted
ghr-miR1533c	1533	AGAUUAAAAAUAAUAAUGU	19	+	3'	30.3	11.9	0.6	DR453981	Predicted
ghr-miR1533d	1533	AAAAUAAAAAUAAAAGGA	18	+	3'	10.61	6.36	0.91	DT561626	Predicted
ghr-miR1533e	1533	AUAAUUAAAAAUAAUAAUUU	20	+	5'	28.11	53.4	0.68	AI055426	Predicted
ghr-miR1533f	1533	AAAUUAAAAAUAAUAAUAA	19	-	3'	34.23	45.41	0.89	CD486467	Predicted

ghr-miR1535a	1535	CGUUUUUGUGGUGAUGGUCU	20	-	3'	41.92	121.4	0.63	contig21820	Predicted
ghr-miR1535b	1535	CUUGUUUGUGAUGUGUGU	18	-	5'	36.62	148.8	0.72	contig21907	Predicted
ghr-miR1854	1854	UGGGCCAUUUGUAGAUUGGA	20	+	5'	32.73	11.36	0.63	DT459810	Predicted
ghr-miR1857	1857	UGGUUUUUCUUGGAGAUGAAG	21	+	3'	41.64	83.44	0.68	ES792140	Predicted
ghr-miR1860	1860	AUCUGAGAAGCUAGGUUUUCUUU	23	+	3'	28.28	37.8	0.68	DW494072	Predicted
ghr-miR1862	1862	ACAAGGUUGGUAUAUUUUAGGACG	24	+	3'	40.32	22.6	0.9	EX172412	Predicted
ghr-miR1869	1869	UGAGAACAAUAGGAUGGGAGAUA	23	-	3'	39.19	18.86	0.65	contig14048	Predicted
ghr-miR1884	1884	AAUGUAUGACGCUGUUGACUUUUC	24	+	5'	23.83	45.2	0.98	EX172380	Predicted
ghr-miR2529	2592	AAAUCUUGAAUCAUGUGUU	19	-	3'	44.82	184.51	0.47	contig14636	Predicted
ghr-miR2595	2595	UCCAUUUUCUUCUUUCUUCU	20	+	5'	39.04	94.12	0.72	contig19425	Predicted
ghr-miR2635	2635	AUUAUUGUCAAGUGUCUUG	19	+	5'	25.76	8.45	0.5	contig4047	Predicted
ghr-miR2645	2645	UUUAUAGAAUGAGCAUAUAC	20	-	3'	30.97	25.6	0.73	AJ513108	Predicted
ghr-miR2673	2673	CCUCUUCCUCUUCCUCUUCUUC	22	-	5'	38.99	69.6	0.47	ES825617	Predicted
ghr-miR2868	2868	UUGAUUUUGGUAGAAGAAA	19	+	5'	35.19	24	0.63	contig17454	Predicted
ghr-miR2876	2876	UUCCUCUAUGGACACUGUUUC	21	+	5'	42.03	177.72	0.58	contig24591	Predicted
ghr-miR2938	2938	GAGCUUUGAGAGGGUUCCGG	20	-	3'	52.33	26.6	0.59	CD485951	Predicted
ghr-miR2948-5p	2948	UGUGGGAGAGUUGGGCAAGAAU	22	+	5'	45.83	30.9	0.94	DW517596	Validated
ghr-miR2949a,b,c	2949	UCUUUUGAACUGGAUUUGCCGA	22	+	5'	43.04	27.3	0.8	contig9309	Validated
ghr-miR2950	2950	UGGUGUGCAGGGGGUGGAAUA	21	+	3'	49.35	43.1	1.13	DW514754	Validated
ghr-miR3476	3476	UGAACUGGGUUUGUUGGCUGC	21	+	5'	37.23	38	1.09	DW497660	Validated

* Length of mature miRNA sequence

Validated means that the miRNA was confirmed by experimental methods (deep sequencing, qRT-PCR or direct cloning)

Table 3-4. Potential targets of cotton miRNAs associated with fiber development

MiRNA	Family	Target	Function	Type
ghr-miR156g	156	contig16368	Cellulose synthase	Fiber development
ghr-miR156g	156	contig18138	Glycosyl transferase, CAZy family GT43	Fiber development
ghr-miR156g	156	contig4371	Glycosyltransferase QUASIMODO1	Fiber development
ghr-miR156f	156	contig13757	Glycosyltransferase, CAZy family GT8	Fiber development
ghr-miR156g	156	contig17691	Glycosyltransferase, CAZy family GT8	Fiber development
ghr-miR156f	156	contig8831	Sugar transporter	Fiber development
ghr-miR156f	156	contig1543	UDP-glucuronate 5-epimerase	Fiber development
ghr-miR414b	414	contig7645	Similar to fiber protein Fb2	Fiber development
ghr-miR414e	414	contig22187	Sugar transporter	Fiber development
ghr-miR529a	529	contig16806	Glycosyl hydrolase family 17 protein	Fiber development
ghr-miR529b	529	contig23483	Glycosyl hydrolase family 17 protein	Fiber development
ghr-miR529a	529	contig19551	Glycosyltransferase, CAZy family GT8	Fiber development
ghr-miR529b	529	contig8845	Sugar transporter, putative	Fiber development
ghr-miR1533e	1533	contig22176	Glycosyltransferase, CAZy family GT47	Fiber development
ghr-miR1533e	1533	contig9681	UDP-glucose 4-epimerase	Fiber development
ghr-miR1533d	1533	contig20591	UGT73C6 (UDP-glucosyl transferase 73C6)s	Fiber development
ghr-miR1533d	1533	contig2536	Xyloglucan endotransglucosylase/hydrolase protein 22 precursor	Fiber development
ghr-miR1533b	1533	contig71	Xyloglucan endotransglucosylase/hydrolase protein 9 precursor	Fiber development
ghr-miR1535b	1535	contig21984	Sucrose synthase	Fiber development
ghr-miR2595	2595	contig8413	Glycosyl transferase family 2 protein	Fiber development
ghr-miR2595	2595	contig9765	Sugar transporter	Fiber development
ghr-miR2595	2595	contig24807	Xylulose kinase	Fiber development
ghr-miR2635	2635	contig2406	Xylose isomerase	Fiber development

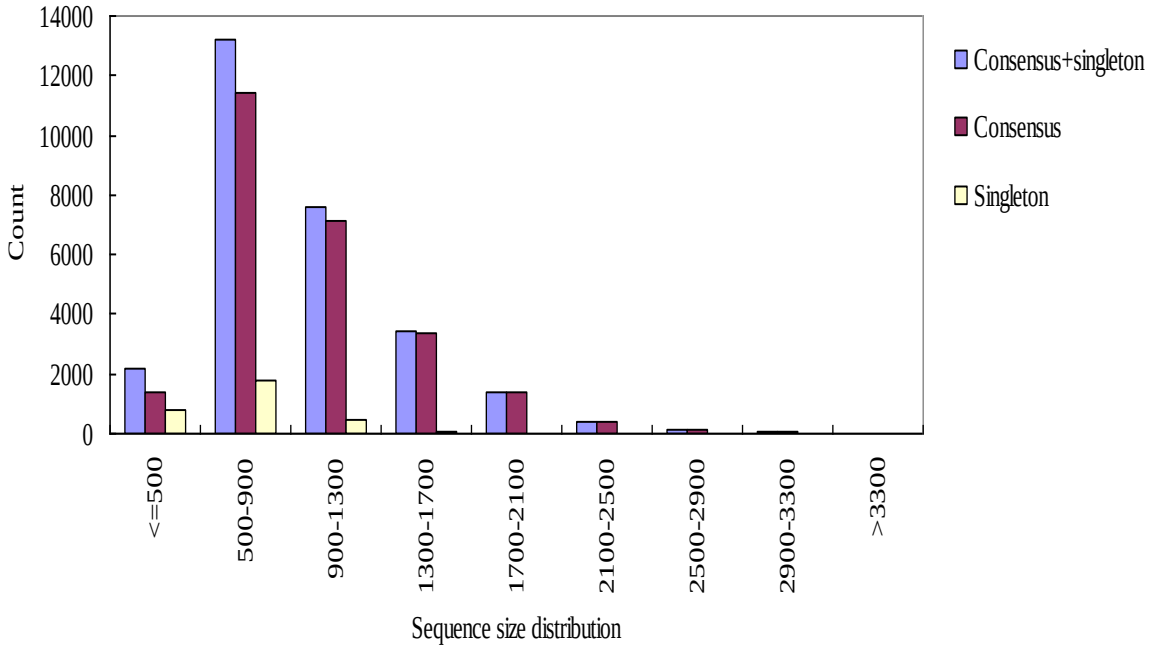


Figure 3-1. Sequence size distribution of consensus contigs and singletons in cotton.

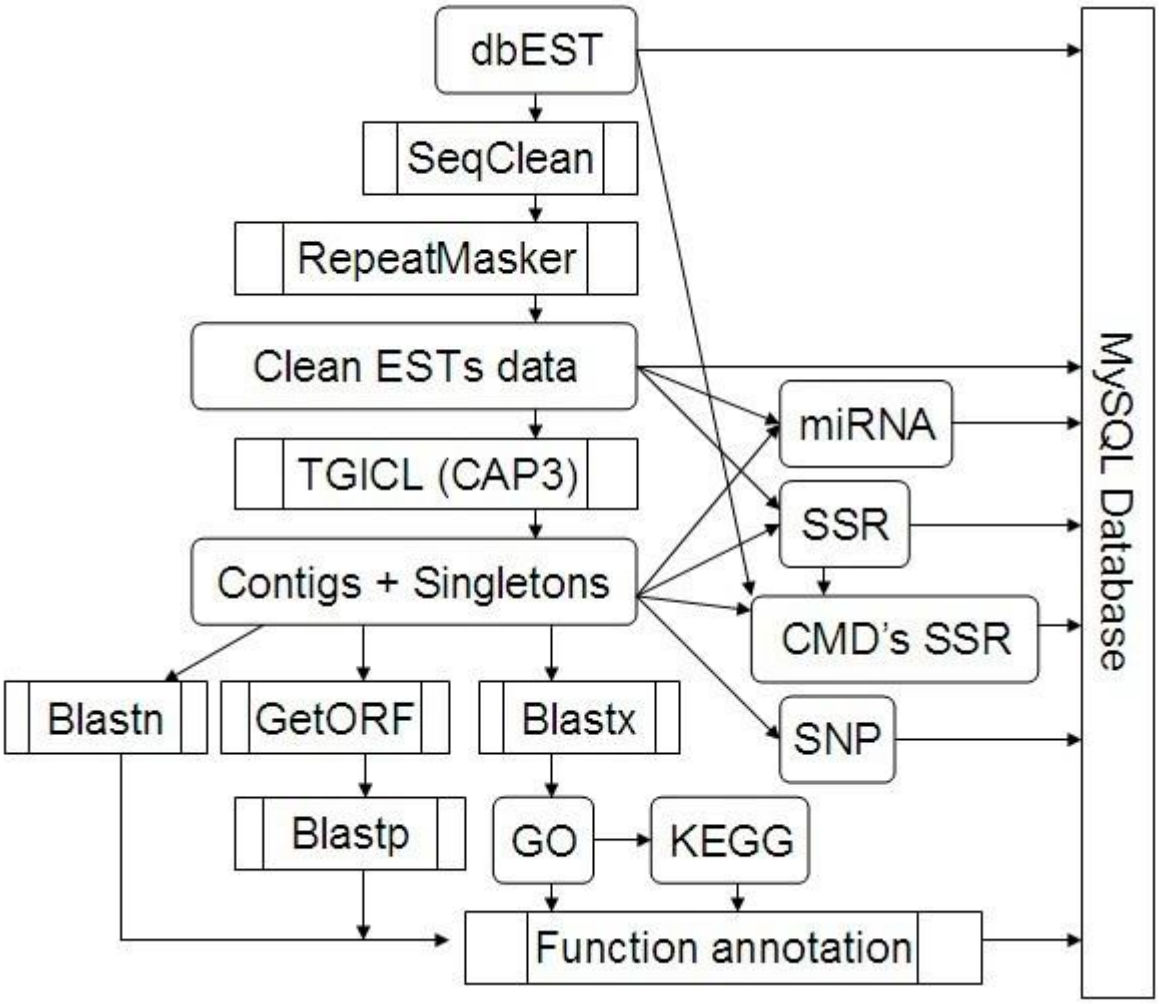
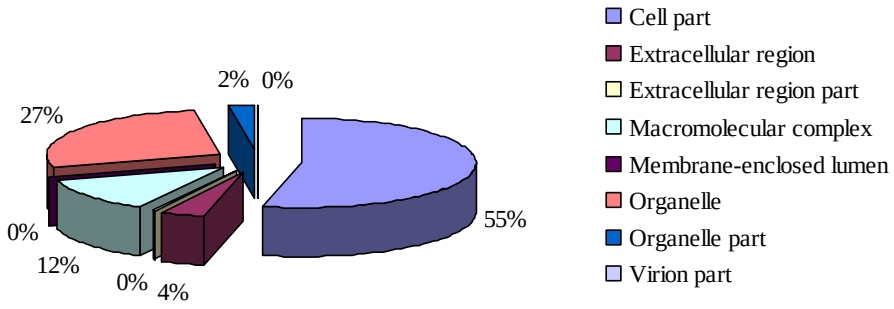
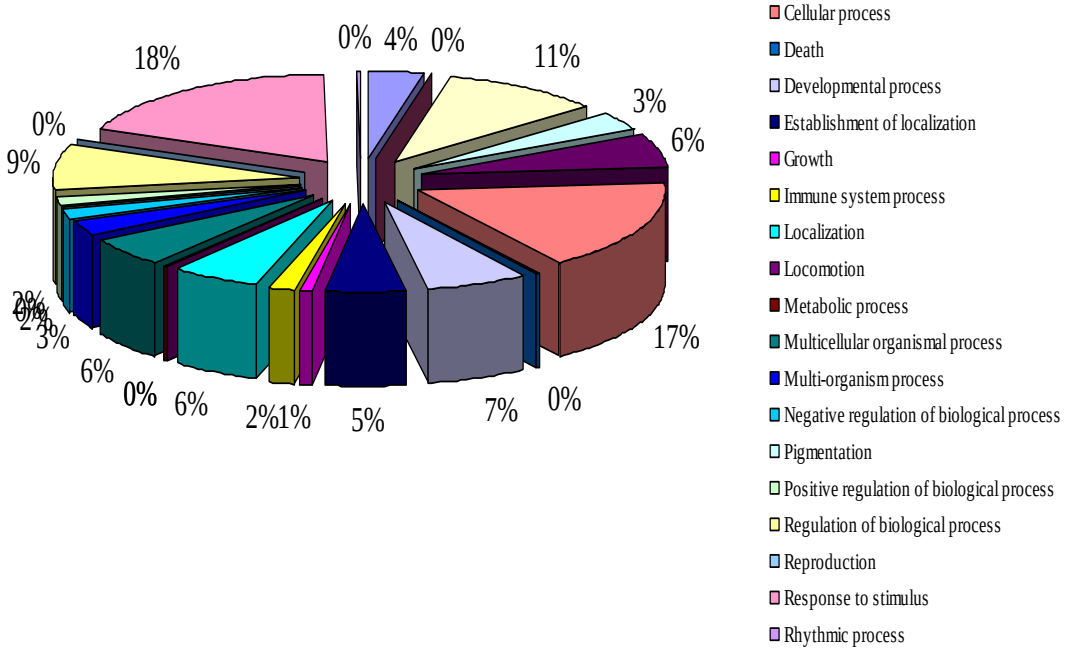


Figure 3-2. Schematic pipeline for cotton EST assembly, data analysis and database development.

A


2%
0%
27%
55%
0%
12%
0%
4%
Cell part
Extracellular region
Extracellular region part
Macromolecular complex
Membrane-enclosed lumen
Organelle
Organelle part
Virion part

B


0%
4%
0%
18%
11%
3%
0%
6%
9%
17%
3%
6%
0%
6%
2%
1%
5%
7%
0%
Anatomical structure formation
Biological adhesion
Biological regulation
Cellular component biogenesis
Cellular component organization
Cellular process
Death
Developmental process
Establishment of localization
Growth
Immune system process
Localization
Locomotion
Metabolic process
Multicellular organismal process
Multi-organism process
Negative regulation of biological process
Pigmentation
Positive regulation of biological process
Regulation of biological process
Reproduction
Response to stimulus
Rhythmic process

C

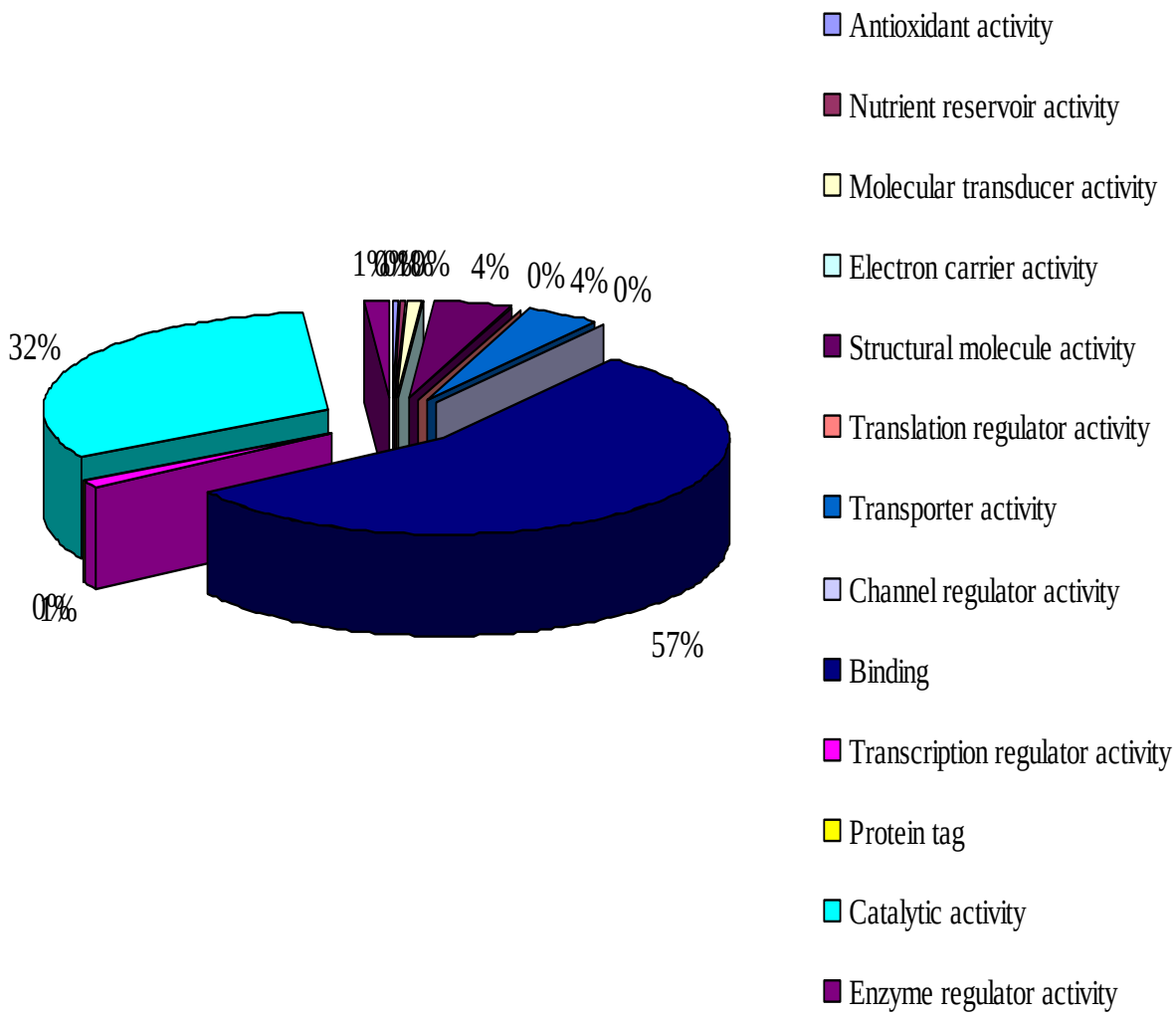
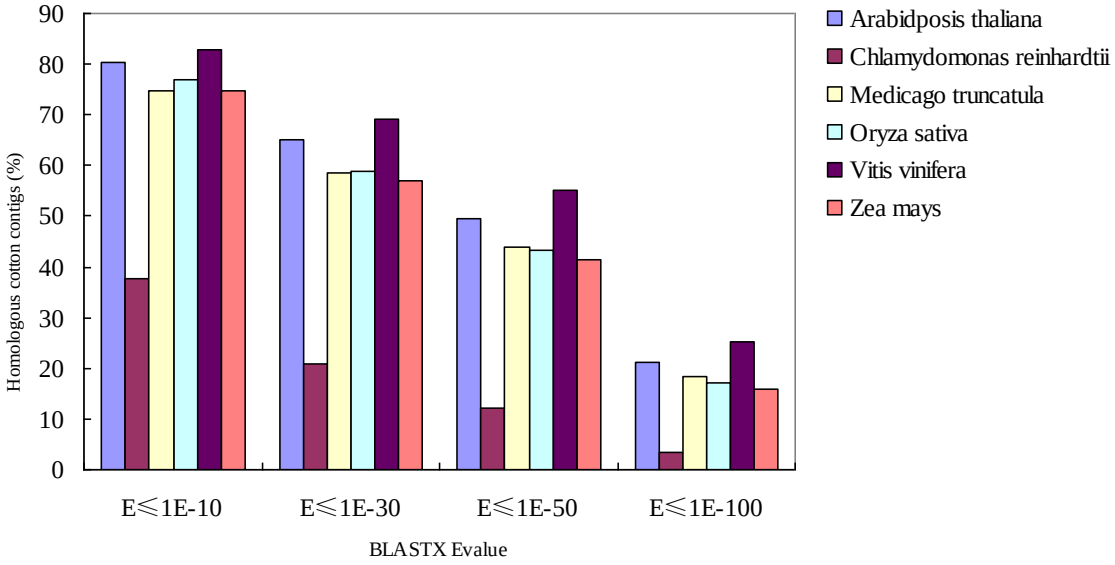


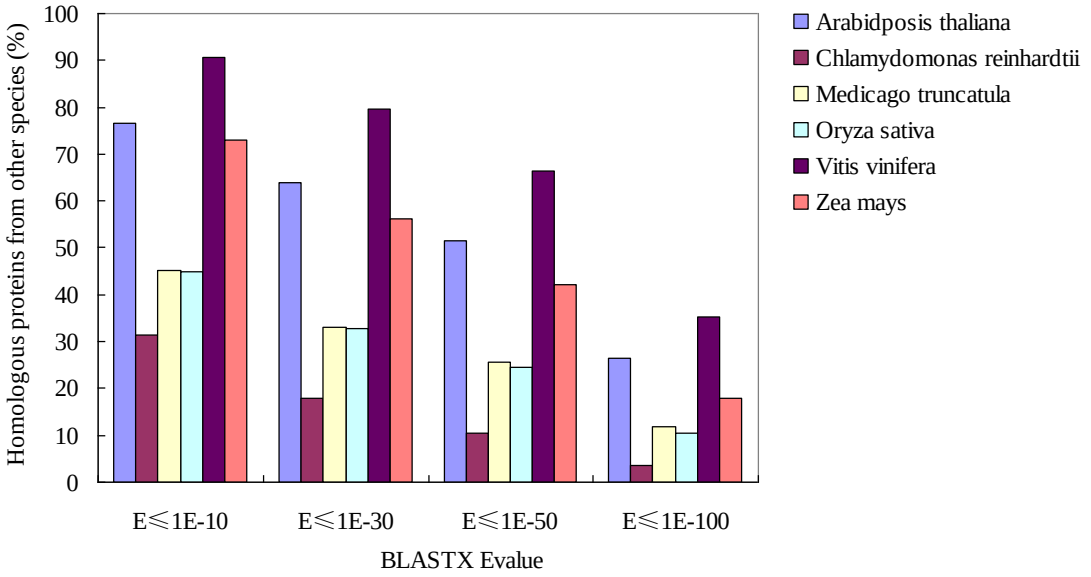
Figure 3-3. Gene Ontology (GO) analysis of 28,432 cotton annotated contigs. The three GO categories are presented: cellular component (a), biological process (b), and molecular function (c).

Figure 3-4. Cluster size distribution of cotton contigs

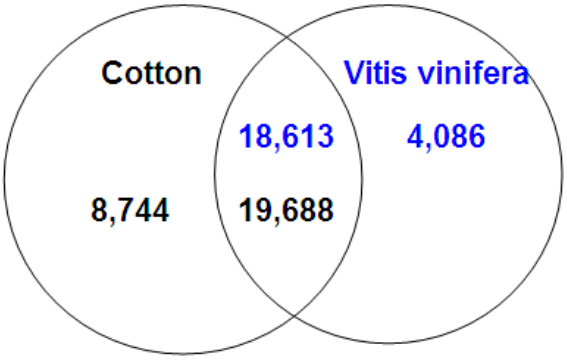
A


Homologous cotton contigs (%)
90
80
70
60
50
40
30
20
10
0
E≤1E-10
E≤1E-30
E≤1E-50
E≤1E-100
BLASTX Evalue
Arabidposis thaliana
Chlamydomonas reinhardtii
Medicago truncatula
Oryza sativa
Vitis vinifera
Zea mays

B


Homologous proteins from other species (%)
100
90
80
70
60
50
40
30
20
10
0
E≤1E-10
E≤1E-30
E≤1E-50
E≤1E-100
BLASTX Evalue
Arabidposis thaliana
Chlamydomonas reinhardtii
Medicago truncatula
Oryza sativa
Vitis vinifera
Zea mays

C.



D.

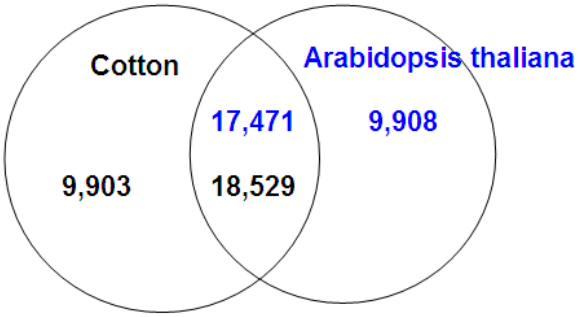
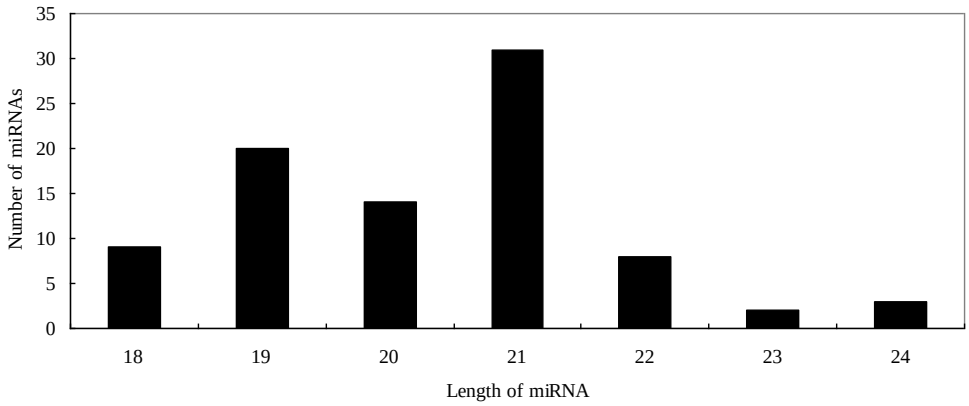


Figure 3-5. Homologous genomic comparison using several blast E-value cutoffs. A. Distribution of percent cotton contigs finding a hit in each genome. B. Distribution of cotton homologous proteins identified in other plant species. C. Comparison of number of homologs identified between cotton and *Vitis vinifera* with a BLASTx E-value cutoff of 1e-30. D. The same comparison between cotton and *Arabidopsis thaliana*.

A.



B.

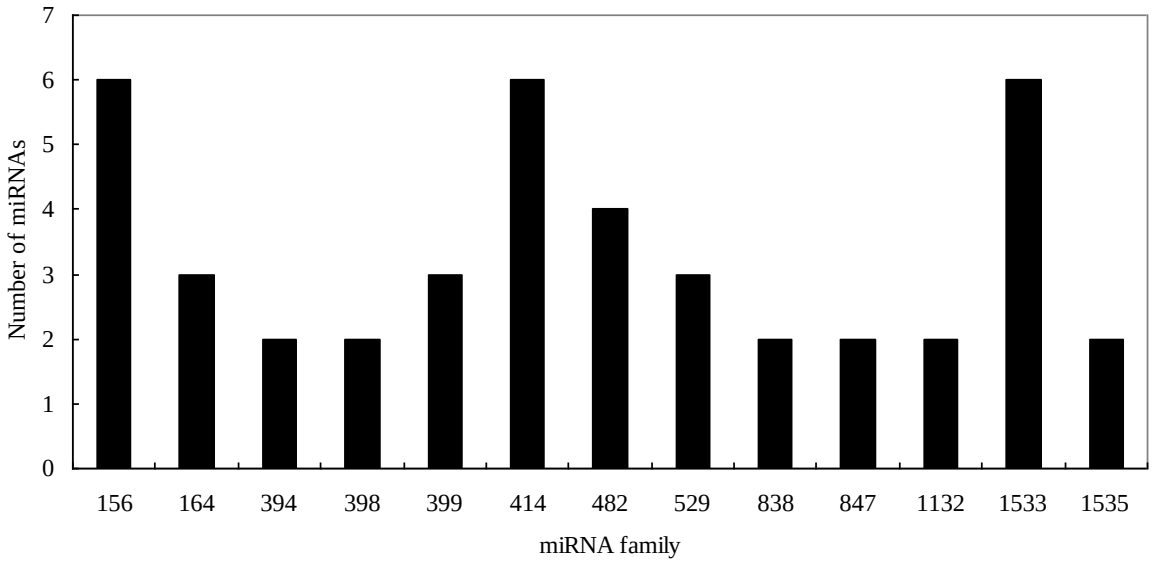
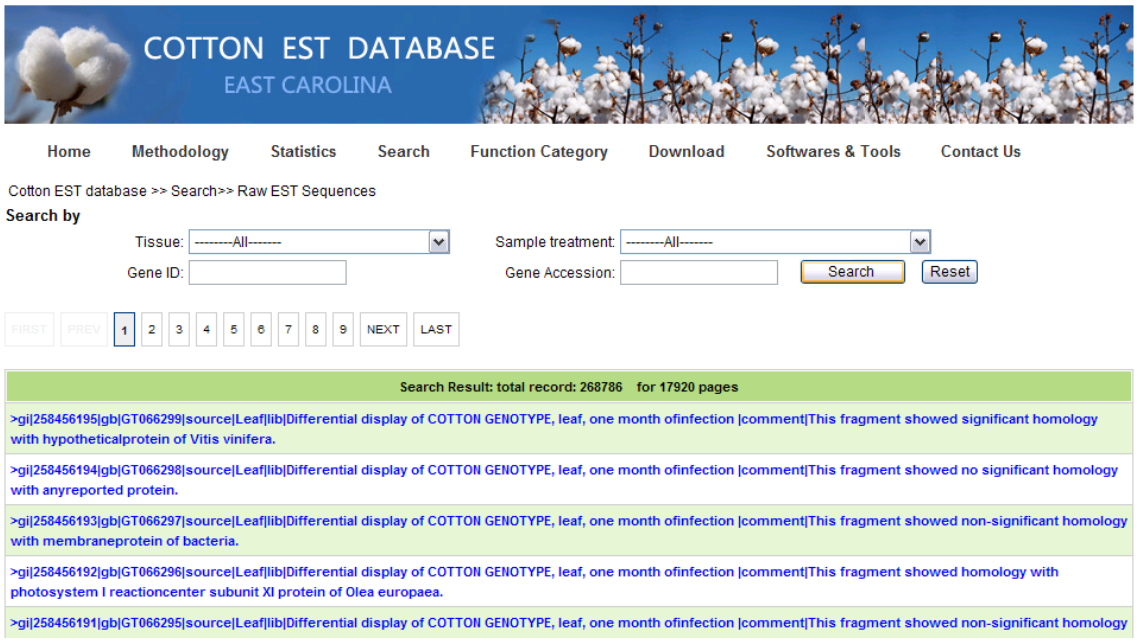


Figure 3-6. A. Distribution of length of miRNAs in cotton. B. Size distribution of cotton miRNA families with more than one member.

A.



B

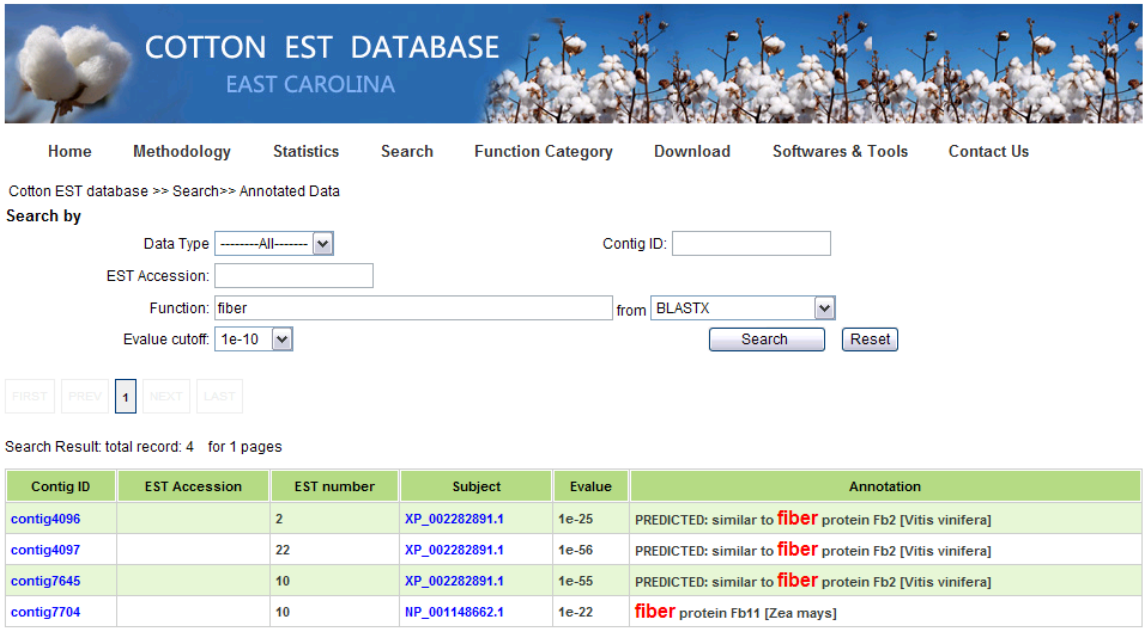


Figure 3-7. Interface of cotton EST database for querying raw ESTs (A), and assembled contigs (B).

CHAPTER 4: Deep sequencing deciphers important miRNA roles in response to drought and salinity stress as well as fiber development in cotton

Abstract

Drought and salinity are two major environmental factors adversely affecting plant growth and productivity. However, the regulatory mechanism is unknown. In this study, we investigated the potential roles of small regulatory miRNAs in cotton response to those stresses. Using the next-generation deep sequencing, a total of 337 miRNAs with precursors were identified, including 289 known miRNAs and 48 novel miRNAs. Of these miRNAs, 155 miRNAs expressed differentially. Target prediction, GO-based functional classification, and KEGG-based functional enrichment show these miRNAs might play roles in response to salinity and drought stresses through targeting a series of stress-related genes. Degradome sequencing analysis showed that at least 55 predicted target genes were further validated to be regulated by 60 miRNAs. CitationRank-based literature mining was employed to sort out the importance of genes related to stress of drought and salinity. NAC, MYB and MAPK family were ranked top under the context of drought and salinity, indicating their important roles for plant to combat drought and salinity stress. According to our target prediction, a series of cotton miRNAs are associated with these top-ranked genes, including miR164, miR172, miR396, miR1520, miR6158, ghr-n24, ghr-n56, and ghr-n59. Interestingly, 163 cotton miRNAs were also identified to target 210 genes that are important in fiber development. These results would contribute to cotton stress-resistant breeding as well understanding fiber development.

Introduction

Drought and high salinity are two of most severe and wide-range abiotic factors that inhibit plant growth and development and ultimately negatively affect plant yields or quality (Krasensky and Jonak, 2012). During long-term evolution process, plant has developed a series of regulatory mechanisms to cope with these unfavorable conditions at different levels, including cellular, physiological, biochemical and molecular processes (Covarrubias and Reyes, 2010). A great number of efforts have been made to identify factors involved in the response to drought and salinity stresses. For instance, it is well established that hormone-mediated signaling cross-talk in plant participates in response against drought and salinity stress, such as ABA, salicylic acid and ethylene (Huang et al., 2012). These studies suggested that gene expression regulation is a crucial strategy for plant to combat the stress of drought and salinity at the post-transcriptional levels (Covarrubias and Reyes, 2010). One of most important players for gene expression regulation is microRNAs (miRNAs), a class of non-coding small RNA molecules in length of ~21 nt. miRNAs are well-known to negatively modulate their targets expression either by mRNA cleavage or translation inhibition, based on perfect or near perfect complementary nucleotide binding to their target mRNAs (Ambros, 2004). Asides from roles in development and metabolism, a series of miRNAs have been shown to participate in abiotic and biotic stress response (Dugas and Bartel, 2004). For example, miR394 is a conserved miRNA that has been identified in a series of plant species, such as *Arabidopsis* (Jones-Rhoades and Bartel, 2004), rice, cotton (Zhang et al., 2007), and *Brassica napus* (Zhao et al., 2012). Recent studies indicated that miR394 is a versatile miRNA that is involved in multiple stress response. miR394 was found to be upregulated by sulfate

deficiency, cadmium, and iron deficiency (Huang et al., 2010). *Arabidopsis* miR398 was identified to detoxify superoxide radicals by directing the cleavage of its two targets, Cu/Zn superoxide dismutases (cytosolic CSD1 and chloroplastic CSD2) (Sunkar et al., 2006). Currently, plant miRNA families related to abiotic stress is estimated up to 40, many of which are associated with salt and drought stress response (Covarrubias and Reyes, 2010; Sunkar, 2010; Wang et al., 2013).

Cotton is one of leading economic crops in the world mainly because of its nature lint fiber, an important material for clothing, fine paper, and other purposes. Currently, increasing research is being performed on cotton to improve its fiber yields and quality, including cotton fiber development mechanism and cotton environmental adaption to salinity and drought (Lu et al., 2013; Trivedi et al., 2012; Zhao et al., 2013). Although cotton is a relatively drought-tolerant and salt-tolerant crop, exposure of cotton to high salinity and excessive water deficit could lead to a series of metabolic disorders in terms of osmotic effects (dehydration), nutritional imbalance, and toxicity of ions, which have a considerable negative impact on cotton growth and lint yield, especially at cotton critical growth stages (Dong, 2012; Mahajan and Tuteja, 2005). To date, based on transcriptome and transgenic analysis, a great number of cotton genes have been identified to display aberrant expression in response to stress of salinity and drought either in cotton or in other species. For instance, overexpression of cotton CBL-interacting protein kinase gene (GhCIPK6) in transgenic *Arabidopsis* resulted in improved tolerance to salt, drought and ABA stress, indicating GhCIPK6 might be as a positive regulator to fight salt and drought stress in cotton (He et al., 2013). Transgenic tobacco overexpressing cotton group C MAP kinase gene (GhMPK2) had a lower rate of water loss and exhibited enhanced tolerance to salt and drought, implicating

GhMPK2 might positively regulate salt and drought tolerance in tobacco and cotton (Zhang et al., 2011). Microarray-based transcriptome analysis uncovered some salt/drought-mediated signal transduction pathways in cotton, where a number of candidate genes express differentially and might be potential makers of tolerance to salt and drought stress, such as WRKY, ERF, transmembrane nitrate transporter, pyruvate decarboxylase, and sucrose synthase (Ranjan et al., 2012; Yao et al., 2011). However, the mechanism controlling cotton response to abiotic stress is still unclear although lots of progress has been recently made on cotton genome sequencing. Therefore, identification of salt-responsive and drought-responsive genes in cotton lags to other model species, like *Arabidopsis* and rice. Also, the regulatory mechanism mediated by these responsive genes is still poorly understood.

miRNAs may play an role during cotton respond to drought and salinity stress. Using both computational and deep sequencing technology, some conserved and new miRNAs have been recently identified in cotton (Chen et al., 2013; Gong et al., 2013; Pang et al., 2009; Qiu et al., 2007; Yang et al., 2013; Zhang et al., 2007; Zhang et al., 2013). Some of these miRNAs and their predicted targets were also found to express differentially in terms of dose-dependent and tissue-dependent under salinity and drought conditions, such as miR156-SPL2, miR162-DCL1, miR159-TCP3, miR395-APS1 and miR396-GRF1 (Wang et al., 2013; Yin et al., 2012). These findings suggest that cotton miRNAs play important roles in response to stress of salt and drought. Understanding how miRNAs participate in gene regulation in stress of salt and drought could allow us to improve cotton tolerance and adaption to salt and drought, further result in improving fiber yields and quality. However, cotton miRNAs identification is also hindered and no genome-wide

identification and functional analysis on cotton miRNAs during exposure to drought and salinity stress. Cotton miRNAs are largely unknown, particularly on their function. In this study, we sequenced three cotton seedling small RNA libraries, which were treated by control, salt and drought. 267 conserved miRNAs and 75 novel miRNAs were identified from the three libraries, in which at least 18 and 27 miRNAs were salt-specific and drought-specific, respectively. Evidences from miRNA target prediction, GO term classification, KEGG-based pathway analysis, and literature-based text mining, also support these identified miRNAs play a critical role in response to stress of salt and drought in cotton.

Method and material

Small RNA libraries preparation and sequencing

The seeds of *G. hirsutum L.* cultivar TM-1 were sterilized with 70% (v/v) ethanol for 60 s, 6% (v/v) bleach for 6-8 min, and then were washed with sterilized water for at least 4 times. The sterilized seeds were germinated on 1/2 Murashige and Skoog (MS) medium (pH 5.8) containing 0.8% agar under a 16 h light/8 h dark cycle at room temperature for 10 d. The MS mediums were supplemented with 0.5% NaCl as salinity treatment and with 5% PEG as drought treatment. Each treatment was replicated for 5 times as in five individual culture chamber and each chamber contained 5 seeds. Ten-day-old cotton seedlings (controls, 0.5% NaCl and 5% PEG treatment) were harvested and immediately frozen in liquid nitrogen. Total RNAs was extracted from each tissue sample using the mirVana miRNA isolation kit (Ambion, Austin, TX) according to the manufacturer's protocol. RNAs were quantified and qualified by Nanodrop ND-1000 (Nanodrop

technologies, Wilmington, DE, USA). All RNA samples were submitted to BGI (Shenzhen, China) for high-throughput sequencing using Illumina HiSeq high-throughput sequencing platform.

Pipeline of bioinformatics analysis

All the raw sequences generated from the three small RNA libraries were cleaned first, including removing 5' and 3' adaptors and filtering low-quality reads. Then, the raw sequences were categorized to unique reads and read counts were also calculated. To evaluate the similarity coefficient of the three sequencing libraries, top 5000 abundant small RNAs were chosen to compute the Jaccard index (Mohorianu et al., 2011). Clean reads fully matching to other RNAs, including repeat RNA, rRNA, snRNA, snoRNA, and tRNA, which were excluded by blastn-short alignment (blast2.2.26+, ftp://ftp.ncbi.nih.gov/blast/executables/blast+/2.2.26/) against Sanger RNA family database (Rfam 10.1, ftp://ftp.sanger.ac.uk/pub/databases/Rfam) (Gardner et al., 2011). The remaining sequences were further aligned against miRBase (Release 20, http://www.mirbase.org/) (Kozomara and Griffiths-Jones, 2011) to discriminate conserved reads and non-conserved reads. A conserved read is defined to have no more than 3 mismatches with known miRNA sequences, otherwise a non-conserved read. Considering some well-known miRNAs cannot be identified to have miRNA* in a small RNA sequencing dataset in some model plant species (Li et al., 2011; Xie et al., 2012), only novel miRNAs and their miRNA*s were required to co-exist in at least one sequencing library.

The newest EST (300,288) and GSS (62,820) databases were assembled with CAP3-based TGICL (ftp://ftp.tigr.org/pub/software/tgi/tgicl/), respectively. Given only the draft D genome of *G. ramondii* in *Gossypium sp.* is available (Lin et al., 2010; Paterson et al., 2012; Wang et al.,

2012), the assembled databases of EST and GSS and the *G. ramondii* genome were used as data source to identify miRNA precursors. Our tool, miRDeepFinder (http://www.leonxie.com/deepfinder.php) was used to identify miRNAs and their targets with the default parameters setting in the software (Xie et al., 2012). Identified miRNA precursor candidates were performed all-to-all alignment to remove repeated. miRNA targets were predicted from cotton annotated mRNA database in NCBI and the assembled protein-coding EST contigs from our cotton EST database (www.leonxie.com) (Xie et al., 2011).

Three released cotton degradome sequencing dataset (GSM1008997: seedlings; GSM1008999: hypocotyl; and GSM1061853: anthers) of upland cotton from NCBI were used to validate the predicted miRNA targets using CleveLand4 with default parameters (Brousse et al., 2014). Considering degradome sequencing is for detecting the sliced site on miRNA targets that generally occurs on the 10^{th} and 11^{th} nucleotides of mature miRNAs. Thus, no mismatches is allowed on the two nucleotides in degradome analysis (Schwab et al., 2005). Finally, miRNA targets in a *p-value* of ≤ 0.05 were kept.

Comparison of miRNA expression profiles in control, salt, and drought-treated cotton seedlings

All miRNAs abundance was standardized to transcript expression level per million reads (RPM). If the original miRNA expression in a library was zero, the normalized expression was adjusted to 0.01 according to a previous report (Murakami et al., 2006). miRNA expression fold change in any two libraries was calculated with the formula, Fold change=$\log_2$(treatment 1 / treatment 2) (Marsit et al., 2006). Pearson's chi-squared test was performed for the significance of miRNA expression from two samples. Fold change and *p*-value were combined to determine the final miRNA

expression significance. We defined expression difference level as following rules: extremely significant (**) if (fold change ≥ 1 or fold change ≤ -1) and *p-value* ≤ 0.01; significant (*) if (fold change ≥ 1 or fold change ≤ -1) and 0.05 ≥ *p*-value > 0.01; otherwise insignificant.

Validation of miRNA expression profiles in control, salt, and drought-treated cotton seedlings

Using the stem-loop qRT-PCR method for miRNA expression assay (Chen et al., 2005), we randomly selected 13 miRNAs (ghr-miR156a, ghr-miR157a, ghr-miR160a, ghr-miR166a, ghr-miR167b, ghr-miR171a, ghr-miR2911, ghr-miR394a, ghr-miR3954a, ghr-miR395a, ghr-n65, ghr-n68, and ghr-n8) to validate miRNA expression of deep sequencing. Ten-day-old cotton seedlings in three biological replicates for qRT-PCR were treated in the same way with those for deep sequencing. qRT-PCRs were performed with Applied Biosystems ViiA 7 Real Time PCR System for each reaction with three technical replicates. Pearson correlation test was performed to test correlation significance of relative miRNA expression in the treatments of drought and salinity to the control between qRT-PCR and deep sequencing under a *p-value* of 0.05.

CitationRank-based literature mining

A great number of studies have performed on the stress of salinity and drought in plant, in which many genes have been proposed or validated to be crucial in response to stress of salinity and drought. To sort out how our identified miRNAs are correlated with these genes, we first used CitationRank-based text-mining (Yang et al., 2009) to sort out their importance to stress of salinity and drought, respectively. In this step, briefly, the key words, "Salt AND Plant AND stress" and "salinity AND Plant AND stress", were used to retrieve salinity-related PubMed entries. Similarly, drought-related PubMed entries were obtained through the key words, "Drought AND Plant AND

stress" and "Water deficit AND Plant AND stress". Retrieved PubMed entries that are from non-plant species were excluded according to the annotation in the Gene-to-PubMed dataset, gene2PubMed, which was downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/). Considering some genes in PubMed entries are homologous and also might be from different species, we categorized these genes as one functional gene. To this end, we used Orthomcl (Version 2.0, http://orthomcl.org/common/downloads/) (Li et al., 2003) to search ortholog among the genes in all retrieved PubMed entries with a threshold of 1e-20. We hypothesized a group of homolog genes should carry out a similar function. We followed the CitationRank algorithm to build up co-existence matrix of gene and compute CitationRank value with an iteration of 1000. We incorporated the whole process of CitationRank calculation into a software package named RNAKER, which is freely available at http://leonxie.com/citationRank.php. Cytoscape was used to visualize regulation network between cotton miRNAs and coding homolog cluster.

Results

High-throughput sequencing of control, salt-, and drought-treated small RNA libraries

A total of 51,857,063 reads were generated from the three cotton small-RNA libraries generated from salinity (18,808,997) and drought (16,938,676) treatment as well control (16,109,390), representing a total of 16,126,755 unique sequences (Table 4-1). Considering upland cotton genome is still not available, we first aligned these sequenced read against the GSS and EST database of upland cotton. It turned out on average 15.99% reads and 51.63% reads were fully (100%) matched back to upland cotton datasets (EST and GSS) and the D genome of *G. ramondii*,

respectively, resulting in a mean of 54.59% successful match in upland cotton and *G. ramondii* (Table 4-1). Overall, the yielded reads and matched reads are similar in the three libraries. We calculated the Jaccard index for the top 5,000 abundant small RNA reads in each library in order to evaluate overall sequencing similarity among the three libraries (Mohorianu et al., 2011). The similarity between salt- and drought-treated libraries was 97.39% (Table 4-2). Furthermore, the two libraries showed a similar similarity with the control library (Control vs Drought: 42.27% and Control vs Salt: 46.94%), respectively. This indicates that some common small RNAs with relatively rich abundance might be readily induced to cope with abiotic stress in cotton, like drought and salinity stress. All three libraries displayed similar distributions to other RNA families including rRNA (~1.34% for the unique- and ~6.57% for the redundant reads), snRNA (~0.02% for the unique and ~0.01% for the redundant), snoRNA (~0.01% for the unique- and ~0.00% for the redundant reads), and tRNA (~0.13% for the unique- and ~1. 03% for the redundant reads) (Table 4-1). A similar size distribution for redundant reads, unique reads and matches unique reads, was observed in the three libraries, in which the 24-nt reads account for the largest part (Figure 4-1). However, the matched redundant reads have the most reads in 21-nt following by 24-nt. Largely, the distribution of small RNA abundance and size in cotton were consistent with the results reported in *Arabidopsis* (Rajagopalan et al., 2006) and rice (Wei et al., 2011).

Identification of conserved miRNA families in cotton

miRNAs are well-known to be highly conserved across species. According to aligning all clean reads from three libraries against all known plant miRNAs in miRBase (Release 20) (Kozomara and Griffiths-Jones, 2011) with no more three than mismatches, we identified a total of 709 known

plant miRNA families in cotton; out of these, 515, 546 and 538 families are from control, drought-treated and salted-treated samples, respectively (Supplementary 4-1). These miRNA families accounted for around 0.47% of the total unique read sequences and 23.59% of the total redundant read sequences on average (Table 4-1). Among these miRNA families, 71 and 58 miRNA families were specific to drought and salt treatment, respectively, whereas 47 miRNA families were only found in control treatment (Figure 4-2A). For example, miR1868 and miR2099 expressed only in drought- and salt-treated samples, respectively (Supplementary 4-1). In addition, drought- and salt-treated libraries shared 65 miRNA families that didn't occur in the control library. 357 out of 709 miRNA families were identified in the three libraries, suggesting their key roles in maintaining normal biological activities, such as miR156/157, miR159, miR168, and miR172 (Table 4-3 and Supplementary 4-1). Interestingly, all three libraries share similar most frequent miRNA families, including miR156, miR157, miR166, miR167, and miR3954. Pearson's chi-squared test showed that 565 out of 709 (79.69%) miRNA families expressed differentially in the three libraries (*p-value* ≤ 0.05). We used both fold change and p-value to define expression significance (significant *: absolute fold change ≥ 1 and *p-value* ≤ 0.05; extremely significant **: absolute fold change ≥ 1 and *p-value* ≤ 0.01). It turned out a total of 443 (62.48%) miRNA families showed significant expression difference in the pairwise comparison of three treatments, including miR157, miR159, miR2948, and miR3694 (Table 4-3 and Supplementary 4-1). miR1854 and miR1148 had the largest fold change in both control-vs-drought and salt-vs-drought up to ≥ 10 folds, implicating drought stress strongly inhibited its expression. Similarly, miR1097 and miR5170 expressed 5-7 folds less in the salt-treated sample than those in the control and drought treatment.

Identification of miRNA precursors and novel miRNAs

To avoid possible sequencing errors, only a total of 1,284,088 unique small RNAs with at least 3 reads in one of the three libraries were used to search for miRNA precursors. Considering upland cotton genome is not available and it is consisted of A and D genome, miRNA precursor search were performed on EST/ GSS databases of upland cotton and D genome of *G. ramondii*, respectively. On average, it yielded 3,602,105 and 2,798,118 100%-nucleotide-match hits per library against the assembled EST/GSS databases and *G. ramondii* genome, respectively. Finally, after removing repeated precursors, a total of 337 miRNAs with precursors were obtained, including 289 known miRNAs and 48 novel miRNAs (Supplementary 4-2). Of these miRNAs, there are 31 from EST, 6 from EST contig, 13 from GSS and 10 from GSS contig. 277 out of 337 (82%) were identified from *G. ramondii* genome. The 277 *G. ramondii*-genome-derived miRNA precursors were aligned against the datasets of EST and GSS of upland cotton to see if these miRNAs could have homologue sequences with at least 95% identity in upland cotton. It turned out that 21, 9, 29, and 8 miRNA precursors are homologous to those from EST, EST contig, GSS, and GSS contig, respectively (Supplementary 4-2). Among the 337 identified miRNAs, there are at least 121 conserved miRNAs and 4 novel miRNAs that have been recently reported in upland cotton by Xue (Xue et al., 2013) and Li (Li et al., 2012). Moreover, the mature miRNAs of ghr-miR477*, ghr-miR2608*, ghr-miR827*, and ghr-miR166j* are the same with mature miRNA of novel_mir_986, novel_mir_1398, novel_mir_50, and novel_mir_848 from Xue and its co-workers, respectively (Xue et al., 2013). For all newly identified 48 novel miRNAs, no homologues was found in other plant species, implicating these novel miRNAs might be cotton-specific.

The 337 miRNAs consisted of 154 miRNA families, including 84 known miRNA families and 44 novel miRNA families. 106 (30.99%) miRNAs have only one member, whereas the other miRNA families have 2 to 16 members. The largest miRNA family is miR5528 with up to 16 members, following by miR166 (13 members), miR169 (13 members), miR171 (11 members), and miR172 (10 members). Interestingly, there are 5 members for the novelly identified miRNA ghr-n36 family.

A total of 8 miRNA clusters were also found in our identified miRNAs, in which 2 and 6 clusters were from EST/GSS of upland cotton and *G. ramondii* genome, respectively (Supplementary 4-3). Normally, miRNA stars are thought to degrade shortly after miRNA maturation and are in extremely low abundance level (Ambros, 2004). Recently, miRNA stars were also found to be bona-fide miRNAs, since they also participate in negative gene regulation with the same mechanism of miRNA (Kozomara and Griffiths-Jones, 2011). Here, we also identified 21 miRNA stars that were generally expressed more than the corresponding mature miRNAs at least in one sequenced library, such as ghr-miR156*, ghr-miR7495*, ghr-miR166j*, and ghr-miR169c*. In addition, some miRNA start have similar expression abundance with their mature miRNAs, we believed these miRNAs are likely to be functional in cotton species or in response to drought and salinity stress.

miRNA expression in response to drought and salt treatment

The overall expression of identified miRNAs was similar between control and salinity treatment but not between control and drought treatment (data not shown). According to the heatmap for the top 50 abundant known miRNAs and novel miRNAs that represent 99.80% and 99.33% of total

expression abundance of known miRNAs and novel miRNAs, respectively, control treatment is closer to drought treatment for known miRNAs, (Figure 4-3A), whereas overall novel miRNAs in drought treatment were expressed more similarly to salt treatment relative to control treatment (Figure 4-3B).

292 out of 337 (86.6%) miRNAs were expressed in all of the three treatments, including 246 known miRNAs and 46 novel miRNAs (Figure 4-2B and 4-2C). 2 and 9 miRNAs were found to express specifically to drought and salt treatments, respectively. 11 known miRNAs and 1 novel miRNAs coexist in both drought and salt treatments, but not in control treatment. Similarly, 15 known miRNAs and 1 novel miRNAs were merely specific to control and drought treatments, whereas 9 known miRNAs and 1 novel miRNAs were only specific to control and salt treatments (Figure 4-2B and 4-2C). These miRNAs expression specificity in the three treatments indicates differential roles in response to stress of salt and drought.

According to Pearson Chi-square test, 155 (50.0%) miRNAs were expressed differentially amongst control, drought and salt treatments ($p\text{-}value \leq 0.05$), including 140 known miRNAs and 15 novel miRNAs, like ghr-miR156a, ghr-miR166a, ghr-miR168, and ghr-n3 (Supplementary 4-2). Based on fold change ($|\text{fold change}| \geq 1$) and *p-value* ($p\text{-}value \leq 0.05$), a total of 77 miRNAs were found to have significant expression difference in either two samples, including ghr-miR160b, ghr-miR399c, ghr-miR172i, ghr-n3, and ghr-n50.

Besides drought/salinity-specific miRNAs, some miRNAs were significantly up-regulated or down-regulated by drought or salt treatment (Supplementary 4-2). For instance, ghr-miR157a/b, ghr-miR166a-j, ghr-miR167a, ghr-miR172a/b/c/f, and ghr-miR396g were down-regulated in both

drought and salinity treatments when compared to those in control treatment, whereas ghr-miR394a-d, ghr-miR160b/c, ghr-miR393c, ghr-miR5340 were up-regulated in drought and salinity treatments. Likewise, some miRNAs were up-regulated or down-regulated under stress of drought or salt and contrarily down-regulated or up-regulated in the other non-control treatment, such as ghr-miR156a/c/d, ghr-miR408a, ghr-miR2911, and ghr-miR3954a/b.

Validation of miRNA expression by qRT-PCR

To test the reliability of deep sequencing, 13 miRNAs (10 conserved miRNAs and 3 novel miRNAs) were randomly selected to perform stem-loop qRT-PCR. Compared with the control miRNA expression, most of the 13 miRNAs' expression in the treatments of drought and salinity appeared similar tendency between deep sequencing and qRT-PCR, respectively (Figure 4-4). Pearson-correlation test showed that the 13 miRNAs' expression relative to the control exhibited significantly positive correlations between deep sequencing and qRT-PCR (R^2 = 0.4398 & *p-value* = 0.0135 and R^2=0.4592 & *p-value* = 0.0111) (Figure 4-4A and 4-4B). Therefore, qRT-PCR-based validation indicated our deep sequencing is reliable in quantifying miRNA expression abundance in cotton. Overall, qRT-PCR showed higher miRNA expression fold change in drought-vs-control and salinity-vs-control than those in small RNA sequencing (Figure 4-4C and 4-4D). We inferred it might be due to exponential amplification in PCR that might elevate real miRNA expression.

miRNA target identification and validation

To predict cotton miRNAs targets, we utilized two upland cotton mRNA datasets, including cotton mRNA databases (2,493 sequences) in NCBI and our annotated cotton EST database (20,307 coding sequences) (Xie et al., 2011). After strict filtration with a series of miRNA target features,

a total of 1,895 unique coding gene were predicted to be targets of 271 conserved miRNAs and 20 novel miRNAs, consisting of a total of 5,430 miRNA-target pairs. Of these targets, 748 and 1,147 are from mRNA database and assembled EST database, respectively (Supplementary 4-4). Degradome sequencing, also known as PARE (parallel analysis of RNA ends), is extensively used to discover in vivo miRNA targets through detecting cleaved miRNA targets from degradome data (Addo-Quaye et al., 2009). Based on three cotton degradome sequencing data from seedlings, hypocotyl, and anthers, 114 miRNA-target pairs were further verified in our miRNA prediction result, including 60 miRNAs and 55 coding genes (Supplementary 4-4). Many of these degradome-validated miRNA-target pairs are well-known in other species, such as ghr-miR156&SBP (squamosa promoter-binding protein), ghr-miR160&auxin response factor, ghr-miR168&argonaute protein, and ghr-miR172&AP2 (Backman et al., 2008).

miRNA target prediction revealed at least 1,019 important genes, which are involved in a variety of biological processes, including fiber development and stress response. These genes were manually classified to 11 major groups based on previous reports, including apoptosis, cell cycle, fiber development, gossypol biosynthesis, stress response, signal transduction, and transcription factors (Table 4-4 and Supplementary 4-4). A majority of miRNAs were predicted to target genes that are associated with stress response, transcription factors, metabolism and fiber development (Table 4-4). For example, based on our target prediction, 151 miRNAs might regulate 229 stress-response-related genes, whereas 217 transcription factors might be targets of 183 miRNAs.

Generally, miRNAs are well-known to be highly conserved across different species, even for miRNA regulatory function. In our target prediction result, some established miRNA-targets were

also similarly identified in cotton. For example, ghr-miR156b/c/d was predicted to target SBP transcription factors (contig7220, contig7221, and contig20213), which have been validated to be miR156 targets and are involved in flower development in *Arabidopsis* (Brousse et al., 2014) and rice (Schwab et al., 2005) (Supplementary 4-4). Through negatively regulating expression of auxin response factor ARF-10, -16, and -17 in *Arabidopsis*, miR160 is important for various plant development processes including seedlings, embryo development and inflorescences (Xue et al., 2013). We also found and validated ghr-miR160b/c might target ARFs in cotton, implicating ghr-miR160b/c might participate in cotton seedling development and seedling resistance to salinity and drought stress. Additionally, some other traditional conserved miRNA-target pairs were also identified in cotton, such as ghr-miR168-AGO1, ghr-miR164a-NAC and ghr-miR172c-AP2 (Supplementary 4-4).

Abiotic stress including drought and salinity stress always induces metabolic rearrangements and regulatory networks in terms of osmotic stress, disorganized membrane, low-activity or denatured protein, and excessive reactive oxygen species (ROS) accumulation (Krasensky and Jonak, 2012). Of these, ROS overproduction in plants is highly reactive and toxic, causing damage to protein, lipids, carbohydrates, and DNA and finally resulting in oxidative stress. Fortunately, plant possesses various enzymes against oxidative stress, including superoxide dismutase (SOD), catalase (CAT), ascorbate peroxidase (APX), glutathione reductase (GR), monodehydroascorbate reductase (MDHAR), dehydroascorbate reductase (DHAR); glutathione peroxidase (GPX) and guaicol peroxidase (GOPX) (Li et al., 2012). There are a lot of ROS-related genes that were identified to be miRNA targets in cotton. Copper/zinc superoxide dismutase (FJ415203.1) might

be targeted by ghr-miR398b. In fact, miR398 has a conserved function that was validated to act on two closely related Cu/Zn superoxide dismutase (cytosolic CSD1 and chloroplastic CSD2) and detoxify intracellular superoxide radicals in *Arabidopsis* (Sunkar et al., 2006). This suggests that ghr-miR398b might be involved in response to stress in cotton seedlings resulted by salinity and drought through modulating its target, copper/zinc superoxide dismutase. Similarly, cytosolic ascorbate peroxidase (EU244476.1, FJ793812.1, and EF432582.1) might be targeted by ghr-miR447a and ghr-miR6190 in cotton (Supplementary 4-4). Furthermore, we also found some specific drought/salt-responsive proteins might be miRNA target candidates in cotton, such as salt overly sensitive protein 2a (ghr-miR6190) (Addo-Quaye et al., 2009).

Interestingly, our degradome sequencing analysis also confirmed some stress-related miRNA targets, like ghr-miR171a-g and contig7077 (scarecrow-like protein 6-like) (Figure 4-5A), ghr-miR395a/b and contig4429 (Sulfate adenylyltransferase) (Figure 4-5B), ghr-miR390a-d and contig13815 (DEAD-box ATP-dependent RNA helicase 21-like) (Figure 4-5C), and ghr-miR172a/b/c/f and contig14537 (Avr9/Cf-9 rapidly elicited protein) (Figure 4-5D). For instance, low expression of osmotically responsive genes 4 (LOS4), a DEAD box RNA helicase gene, was found to be essential for mRNA export and important for development and stress responses in Arabidopsis, whose mutation could enhance cold stress-induction of the master regulator of cold tolerance, C-repeat binding factor 2 (CBF2) and its downstream target genes (Gong et al., 2005). Sulfur is a macronutrient that is necessary for plant growth and development, referring to assimilation of cysteine, methionine, glutathione and other sulfur-containing metabolites (Liang et al., 2010). It has been reported that sulfur metabolism play significant roles in drought stress

signaling transduction, since primary and secondary sulfur metabolism should be coordinated until a certain complex balance in plant (Chan et al., 2013). To date, miRNA395 were validated to participate in sulfur metabolism by targeting ATP-sulfurylase (Liang et al., 2010) and sulfate transporter (Allen et al., 2005). Both our miRNA target prediction and degradome sequencing analysis showed ghr-miR395a/b is also likely to be involved in sulfur metabolism by regulating sulfate adenylyltransferase in cotton. Thus, these stress-related miRNAs and their targets might also play roles in response to stresses of drought and salinity.

miRNAs not only target genes associate to drought and salinity stress but also genes involved in cotton fiber development. The fiber-related function distributes in fiber cell initiation, fiber-related carbohydrate metabolism, cellulose biosynthesis, fiber cell elongation, and some other important transcription factors (Supplementary 4-4). 163 miRNAs were predicted to take part in fiber development by targeting 210 unique genes in cotton (Table 4-4 and Supplementary 4-4). For instance, at least 15 cellulose synthase or cellulose synthase-like proteins might be targets of 34 miRNAs including ghr-miR166m, ghr-miR167b, ghr-miR169k, and ghr-miR172j. An MYB-transcription factor, CAPRICE (*CPC*) was initially identified to be a negative regulator of non-root hair cells and later also found to play a role in inhibiting leaf trichome development in *Arabidopsis* (Backman et al., 2008). Here, we found ghr-miR447a and ghr-miR5255a/b/c/e/f/g/h might target *CPC*, indicating ghr-miR447a and ghr-miR5255a/b/c/e/f/g/h may play a role in root development to respond stress of drought and salinity or fiber development by regulating *CPC* in cotton (Supplementary 4-4).

GO and KEGG pathway analysis

GO-based analysis allows us to know what GO terms (biological process, molecular function, and cellular component) a gene belongs to (Ashburner et al., 2000). Therefore, GO-based analysis could give us more ideas on understanding miRNA function. A total of 274 miRNAs (256 conserved miRNA and 18 novel miRNAs) and their 1,252 targets were classified to 557 molecular functions, 729 biological processes, and 188 cellular components (Supplementary 4-5). At least 151 miRNAs and their 229 targets that are associated with stress response were able to be categorized to 104 molecular functions, 136 biological processes, and 40 cellular components (Table 4-4). 54 pairs of miRNA-target belong to the biological process of response to salt stress (GO:0009651), involving ghr-miR156e, ghr-miR162b, ghr-miR169h, ghr-miR172e, ghr-miR396h, and ghr-miR399i. Similarly, at least 35 pairs of miRNA-target belong to the biological process of response to desiccation (GO:0009269) and response to water deprivation (GO:0009414), like ghr-miR159b, ghr-miR166h, ghr-miR399i, ghr-n26, and ghr-miR399f. Many of classified biological processes were associated with signal transduction, such as auxin metabolism (GO:0009850) and biosynthesis (GO:0009851), ethylene mediated signaling pathway (GO:0009873), response to biotic stimulus (GO:0009607), cytokinin metabolism (GO:0009690) and biosynthesis (GO:0009691), and jasmonic acid metabolism (GO:0009694) and biosynthesis (GO:0009695) (Supplementary 4-5).

KEGG-based analysis allows us to enrich 159 miRNAs and 235 targets to 93 pathways, including photosynthesis (ath00195), glycolysis/gluconeogenesis (ath00010), oxidative phosphorylation (ath00190), biosynthesis of plant hormones (ath01070), and starch and sucrose metabolism (ath00500) (Supplementary 4-6). 22 out of 93 pathways might interact with stress

response through 151 miRNAs and their 229 targets. Interestingly, many of these miRNAs and targets are involved in the pathways that metabolize and biosynthesize some intermediate products important for fiber development, such as glucose, sucrose, starch, and fatty acid.

Text-mining drought/salinity responsive genes

A great number of genes have been identified in response to drought and salinity stress, which widely exist in huge literatures. Currently, there is no any database that summarizes genes associated with response to stress of drought and salinity. Meanwhile, we also wanted to investigate how our identified miRNAs act on these reported stress-responsive genes. To this end, we hypothesized that a genes frequently mentioned or cited in a certain topic or context should be an important gene for the topic or context. We named the gene as a frequent gene. In addition, under the same topic or context, a gene coexisting with a frequent gene in literature should be more important than other genes that singly exists in a paper or coexists with an infrequent gene. The idea was successfully applied to sort the importance of genes to serious adverse drug reaction (SADR), known as CitationRank algorithm (Yang et al., 2009). Based on the idea, we first implemented the algorithm with PERL and then applied it to rank gene's relevance to stress of drought and salt. After searching with keywords of drought and salinity and discarding the genes and PubMed entries from non-plant species (see Methods), a total of 595 and 1,078 effective coding genes are retrieved from 228 and 419 effective PubMed entries under the context of salinity and drought, respectively (data not shown). These genes are from 12 (salinity) and 15 (drought) plant species, respectively. Considering these coding genes are likely homologous in different species, we used Orthomcl to cluster them and then calculated CitationRank value for each cluster.

The salinity-context-based and drought-context-based genes were clustered to 351 and 918 groups with a cutoff E-value of 1e-20, respectively. To build the connection between context-based genes and identified miRNAs&targets, only the protein sequence of the longest gene in each cluster was compiled together as subject datasets, which was then performed a BLASTX alignment with miRNA targets under the cutoff E-value of 1e-20.

In our CitationRank result, the top 5 ranked genes in drought context are MYB3R transcription factor, serine/threonine-protein kinase EDR1, SKP1-like protein 18, aquaporin PIP1-5, and abscisic acid receptor PYL7, corresponding to 27 miRNAs and 14 targets, such as ghr-miR156e, ghr-miR172g, ghr-miR447a, ghr-miR1876a, ghr-n6, contig11235 (MYB73 transcription factor), GU207868.1 (serine/threonine protein kinase 1) (Supplementary 4-7). Meanwhile, the genes including Granulin repeat cysteine protease, alcohol dehydrogenase 1, glycine-rich RNA-binding protein 3, gibberellin 2-beta-dioxygenase 7, and allene oxide cyclase 3, were ranked top 5 in salinity context, involving 11 miRNAs and 8 targets (Supplementary 4-7). Interestingly, 5 miRNAs (ghr-miR5284, ghr-miR6158a, ghr-miR6158b, ghr-miR6190, and ghr-miR6424e) were associated with top-ranked genes in both drought and salt context, implicating they are drought/salinity responsive miRNAs.

Discussion

Differentially expressed miRNAs involved in abiotic stress response

Under drought and salt stress, many conserved and novel miRNAs were expressed differentially; some miRNAs were even specifically expressed in drought and/or salt treatment. miR156/157

family is the most abundant miRNAs in all of three treatments, accounting for 60% of total miRNA reads (Table 4-3 and Supplementary 4-1). miR156 and miR157 were down-regulated in drought treatment by 0.44 and 1.23 folds, respectively, when compared with the control. However, miR156 was up-regulated (0.43 fold) and miR157 was down-regulated (0.41 fold) by salt treatment (Table 4-3 and Supplementary 4-1). Both of miR156 and miR157 in cotton root and leaf were reported to be down-regulated in high concentration of salt (>2.5%) and with the increase of PEG (drought) in a dose-dependent manner (Wang et al., 2013). Our result was also largely consistent with the results. However, miR156 was up-regulated in salt treatment relative to the control. It might be caused by that miR156 expression in our result represents a miR156 family, but not a certain miR156 member. miR156/157 is well-known as negatively targeting SPL transcription factor in plant, and miR156/157 overexpression results in a delayed onset of adult traits and flowering in *Arabidopsis* (Park et al., 2010). Overproduction of miR156 (Corngrass1, *Cg1*) caused an extension of the juvenile vegetative phase in maize (Aharon et al., 2003). Current research on miR156/157 is mainly focusing on its role in morphology change and blooming regulation. Here we offered evidence from small sequencing that drought and salinity stress disturb miR156/157 expression, implicating miR156/157's novel role in response to drought and salinity stress.

NF-YA (GmNFYA3) of the NF-Y complex in soybeans was able to be induced by abscisic acid (ABA) and abiotic stresses including drought, NaCl and cold. Overexpression of GmNFYA3 in *Arabidopsis* leads to reduced leaf water loss and enhanced drought tolerance, and elevates its sensitivity to high salinity and exogenous ABA. *In vivo* experiment in tobacco showed miR169 directs GmNFYA3 mRNA cleavage (Xue et al., 2009). In cotton, we predicted NF-YA3

(contig16841 and contig22907) is the target of ghr-miR169. Additionally, compared with the control, ghr-miR169a/b/c expression in drought and salinity treatment was significantly down-regulated by 0.04-0.93 fold. In contrast, ghr-miR169d/e/f/g were significantly up-regulated in drought and salinity treatment by 0.04-0.85 fold. Ghr-miR169i/j expression was inhibited in drought treatment and up-regulated in salinity treatment. Therefore, we infer that at least ghr-miR169a/b/c might play positive role in combating drought and salinity stress by acting on NF-YA3 in cotton.

Plant aquaporin proteins are a class of large major intrinsic protein family, which are well-known to play a role in transport of diverse small molecules including water and other small nutrients through biological membranes (Park et al., 2010). Besides mediated in absorption and transportation of water and nutrient, aquaporin is also found to get involved in a series of abiotic stress, like stress of drought and cold. For example, overexpression of a plasma membrane aquaporin in transgenic tobacco improves plant vigor under favorable growth conditions but not under drought or salt stress, since symplastic water transport via plasma membrane aquaporins has a deleterious effect during drought or salt stress (Aharon et al., 2003). Recently, Rh-TIP1, a TIP type aquaporin gene isolated from rose, was found to be repressed by both of treatments of ethylene and water deficit (Xue et al., 2009). We identified that 20 miRNAs might target 22 aquaporins proteins in cotton, in which two miRNA-target pairs (ghr-miR4371&contig780 and ghr-miR4371&BK007054.1) were also detected by degradome sequencing analysis (Supplementary 4-4). Thus, aquaporins might be involved in response to stress of drought and salinity in cotton by being a miRNA target, such as ghr-miR4371.

Transgenic rice plants with overexpression of miR393 are more sensitive to salt and alkali treatment, suggesting miR393 is a negative regulator in response to salt and alkali stress by targeting abiotic related genes (Gao et al., 2011). In this study, we obtained an opposite result that ghr-miR393a/b/c/d/e expression was up-regulated in drought and salinity treatment. However, the predicted ghr-miR393 targets in cotton are also stress-related genes, as well several hormone-responsive genes, including NADPH:cytochrome P450 reductase (CPR1), class III peroxidase (POX4), protein AUXIN SIGNALING F-BOX 3, TIR1 (TRANSPORT INHIBITOR RESPONSE 1). Therefore, if miR393 in cotton is also an negative stress-combating regulator as that in rice, one possible explanation is that cotton seedlings' miR393's up-regulation in drought and salinity stress contributes to enlarging stress signal in cotton and then triggers a more efficient or powerful pathway to counter stress of drought and salinity by cross-talking of signal transduction, like by auxin-related pathway.

miRNAs and targets important for fiber development

Cotton fiber development and maturation determine fiber yields and quality. Therefore, we also paid extra more attention on miRNAs and targets that are related to cotton fiber development. Interestingly, there are many identified miRNAs and their targets that are likely to play crucial roles in fiber development and maturation. MYB transcription factors have been known to play a role in promoting ovule epidermal cells into the elongated cotton fiber. Recently, GhMYB25-like, an R2M3 MYB, was newly identified as a key factor in early cotton fiber development, since GhMYB25-like silence resulted in cotton plants with fibreless seeds (Walford et al., 2011). Three GhMYB25 or GhMYB25-like (AF336283.1, AY464054.1, and HM134084.1) were detected to be

targeted by ghr-miR4370, ghr-miR5565a/c, and ghr-miR6158a/b in cotton, implicating the five miRNAs might get involved in early fiber development. In addition, of the large MYB family, there are many other MYB members that were also reported to be important during fiber development, such as MYB2 (Wang et al., 2004) and MYB109 (Pu et al., 2008). This suggests that MYB transcription factors are so important that their regulation function should receive more concerns in understanding fiber development. We uncovered a total of 40 miRNAs targeted at least 25 MYB transcription factors, such as ghr-miR156e, ghr-miR159a, ghr-miR162b, ghr-miR167b, ghr-miR169b, and ghr-miR172a. A recent study shows that two miRNAs (miR828 and miR858) targeting MYB2, which may play a role in fiber development (Guan et al., 2014). In addition to the genes involving in fiber cell initiation and fiber early development, we also found many identified miRNAs target a batch of genes that are related to fiber elongation and maturation. For example, after fiber elongation, fiber cell enters into secondary cell-wall formation that is characterized by massive synthesis of cellulose comprising multiple 1,4-β-glucan chains. Cellulose synthase is responsible for the glucan-chain elongation (Somerville et al., 2004). Therefore, cellulose synthase is an indispensable part of fiber maturation. In our study, at least 17 cellulose synthase were predicted to be 35 miRNAs' targets in cotton, suggesting these miRNAs and targets can be utilized to study their roles in fiber posterior development. Overall, our result provided a good data resource for better understanding fiber development. Several other studies also show many miRNAs were differentially expressed during cotton fiber initiation and development (Chen et al., 2013; Gong et al., 2013; Pang et al., 2009; Zhang et al., 2013).

Top-ranked genes and miRNAs for drought and salinity response

Using CitationRank-based algorithm, we first sorted out coding genes relevance in the context of drought and salinity stress in plant, respectively. MYB3R was ranked to be the most important gene in drought context, which is targeted by 10 miRNAs in cotton (Figure 4-6A). Recently, increasing research reported MYB is associated with drought tolerance in plant. For instance, NbPHAN, a R2R3-type MYB gene in tobacco whose silence led to severe wilting and increased rate of water loss in tobacco leaf, was considered to play a positive role in addressing drought stress (Huang et al., 2013). As the discussion above, MYB is also crucial for fiber development. Therefore, related regulation mechanism between miRNA and MYB transcription factor in cotton is promising to serve for dual purposes, fiber development and drought tolerance. For salinity response, genes including granulin repeat cysteine protease, alcohol dehydrogenase 1, glycine-rich RNA-binding protein 3, and gibberellin 2-beta-dioxygenase 7, were sorted out to be high-ranked important for salinity response (Figure 4-6B). Interestingly, there is many reports that in fact didn't associate granulin repeat cysteine protease with salinity response but with programmed cell death in plant (Andeme Ondzighi et al., 2008; Chen et al., 2006; Jiang et al., 2007; Yamada et al., 2001). Salinity stress might be more readily to trigger programmed cell death in plant through cysteine protease. However, the ranking result is merely based on literature text-mining and more experimental evidence is needed in future. Overall, our relevance ranking result for the context of drought and salinity stress provides an idea of gene/miRNA importance to drought and salinity stress in cotton, which could contribute to future cotton drought/salt tolerance research.

Supporting Information

Supplementary 4-1. Summary of miRNA family comparison among control, salt, and drought libraries in cotton. Many miRNAs are significantly differentially expressed after drought and salinity treatment.

Supplementary 4-2. Characteristics of all conserved miRNAs, including miRNA chromosome location, read number, and their precursor sequence. Individual miRNA read number and the comparison among different treatments.

Supplementary 4-3. miRNA clusters in cotton and their expression among different treatments.

Supplementary 4-4. miRNA targets for conserved and species-specific miRNAs in cotton.

Supplementary 4-5. GO ontology classification of identified miRNA families in cotton.

Supplementary 4-6. Gene pathway analysis for cotton miRNA targets based on GO and KEGG analysis

Supplementary 4-7. CitationRank results on gene clusters involving in drought and salinity treatment.

Reference

Addo-Quaye, C., W. Miller, and M.J. Axtell. 2009. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 25:130-131.

Aharon, R., Y. Shahak, S. Wininger, R. Bendov, Y. Kapulnik, and G. Galili. 2003. Overexpression of a plasma membrane aquaporin in transgenic tobacco improves plant vigor under favorable growth conditions but not under drought or salt stress. *The Plant cell* 15:439-

447.

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

Ambros, V. 2004. The functions of animal microRNAs. *Nature* 431:350-355.

Andeme Ondzighi, C., D.A. Christopher, E.J. Cho, S.C. Chang, and L.A. Staehelin. 2008. Arabidopsis protein disulfide isomerase-5 inhibits cysteine proteases during trafficking to vacuoles before programmed cell death of the endothelium in developing seeds. *The Plant cell* 20:2205-2220.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25:25-29.

Backman, T.W., C.M. Sullivan, J.S. Cumbie, Z.A. Miller, E.J. Chapman, N. Fahlgren, S.A. Givan, J.C. Carrington, and K.D. Kasschau. 2008. Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic acids research* 36:D982-985.

Brousse, C., Q. Liu, L. Beauclair, A. Deremetz, M.J. Axtell, and N. Bouche. 2014. A non-canonical plant microRNA target site. *Nucleic acids research*

Chan, K.X., M. Wirtz, S.Y. Phua, G.M. Estavillo, and B.J. Pogson. 2013. Balancing metabolites in drought: the sulfur assimilation conundrum. *Trends in plant science* 18:18-29.

Chen, C., D.A. Ridzon, A.J. Broomer, Z. Zhou, D.H. Lee, J.T. Nguyen, M. Barbisin, N.L. Xu, V.R.

Mahuvakar, M.R. Andersen, K.Q. Lao, K.J. Livak, and K.J. Guegler. 2005. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic acids research* 33:e179.

Chen, H.J., D.J. Huang, W.C. Hou, J.S. Liu, and Y.H. Lin. 2006. Molecular cloning and characterization of a granulin-containing cysteine protease SPCP3 from sweet potato (Ipomoea batatas) senescent leaves. *J Plant Physiol* 163:863-876.

Chen, X., W. Gao, J. Zhang, X. Zhang, and Z. Lin. 2013. Linkage mapping and expression analysis of miRNAs and their target genes during fiber development in cotton. *BMC Genomics* 14:706.

Covarrubias, A.A., and J.L. Reyes. 2010. Post-transcriptional gene regulation of salinity and drought responses by plant microRNAs. *Plant, cell & environment* 33:481-489.

Dong, H. 2012. Combating salinity stress effects on cotton with agronomic practices. *African Journal of Agricultural Research* 7:8.

Dugas, D.V., and B. Bartel. 2004. MicroRNA regulation of gene expression in plants. *Current opinion in plant biology* 7:512-520.

Gao, P., X. Bai, L. Yang, D. Lv, X. Pan, Y. Li, H. Cai, W. Ji, Q. Chen, and Y. Zhu. 2011. osa-MIR393: a salinity- and alkaline stress-related microRNA gene. *Molecular biology reports* 38:237-242.

Gardner, P.P., J. Daub, J. Tate, B.L. Moore, I.H. Osuch, S. Griffiths-Jones, R.D. Finn, E.P. Nawrocki, D.L. Kolbe, S.R. Eddy, and A. Bateman. 2011. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic acids research* 39:D141-145.

Gong, L., A. Kakrana, S. Arikit, B.C. Meyers, and J.F. Wendel. 2013. Composition and Expression

of Conserved MicroRNA Genes in Diploid Cotton (Gossypium) Species. *Genome Biology and Evolution* 5:2449-2459.

Gong, Z., C.H. Dong, H. Lee, J. Zhu, L. Xiong, D. Gong, B. Stevenson, and J.K. Zhu. 2005. A DEAD box RNA helicase is essential for mRNA export and important for development and stress responses in Arabidopsis. *The Plant cell* 17:256-267.

Guan, X., M. Pang, G. Nah, X. Shi, W. Ye, D.M. Stelly, and Z.J. Chen. 2014. miR828 and miR858 regulate homoeologous MYB2 gene functions in Arabidopsis trichome and cotton fibre development. *Nat Commun* 5:

He, L., X. Yang, L. Wang, L. Zhu, T. Zhou, J. Deng, and X. Zhang. 2013. Molecular cloning and functional characterization of a novel cotton CBL-interacting protein kinase gene (GhCIPK6) reveals its involvement in multiple abiotic stress tolerance in transgenic plants. *Biochemical and biophysical research communications* 435:209-215.

Huang, C., G. Hu, F. Li, Y. Li, J. Wu, and X. Zhou. 2013. NbPHAN, a MYB transcriptional factor, regulates leaf development and affects drought tolerance in Nicotiana benthamiana. *Physiol Plant* 149:297-309.

Huang, G.T., S.L. Ma, L.P. Bai, L. Zhang, H. Ma, P. Jia, J. Liu, M. Zhong, and Z.F. Guo. 2012. Signal transduction during cold, salt, and drought stresses in plants. *Molecular biology reports* 39:969-987.

Huang, S.Q., A.L. Xiang, L.L. Che, S. Chen, H. Li, J.B. Song, and Z.M. Yang. 2010. A set of miRNAs from Brassica napus in response to sulphate deficiency and cadmium stress. *Plant biotechnology journal* 8:887-899.

Jiang, Y., B. Yang, N.S. Harris, and M.K. Deyholos. 2007. Comparative proteomic analysis of NaCl stress-responsive proteins in Arabidopsis roots. *Journal of experimental botany* 58:3591-3607.

Jones-Rhoades, M.W., and D.P. Bartel. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular cell* 14:787-799.

Kozomara, A., and S. Griffiths-Jones. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* 39:D152-157.

Krasensky, J., and C. Jonak. 2012. Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *Journal of experimental botany* 63:1593-1608.

Li, L., C.J. Stoeckert, Jr., and D.S. Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13:2178-2189.

Li, Q., X. Jin, and Y.X. Zhu. 2012. Identification and analyses of miRNA genes in allotetraploid Gossypium hirsutum fiber cells based on the sequenced diploid G. raimondii genome. *J Genet Genomics* 39:351-360.

Li, T., H. Li, Y.X. Zhang, and J.Y. Liu. 2011. Identification and analysis of seven H(2)O(2)-responsive miRNAs and 32 new miRNAs in the seedlings of rice (Oryza sativa L. ssp. indica). *Nucleic acids research* 39:2821-2833.

Liang, G., F. Yang, and D. Yu. 2010. MicroRNA395 mediates regulation of sulfate accumulation and allocation in Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* 62:1046-1057.

Lin, L., G.J. Pierce, J.E. Bowers, J.C. Estill, R.O. Compton, L.K. Rainville, C. Kim, C. Lemke, J.

Rong, H. Tang, X. Wang, M. Braidotti, A.H. Chen, K. Chicola, K. Collura, E. Epps, W. Golser, C. Grover, J. Ingles, S. Karunakaran, D. Kudrna, J. Olive, N. Tabassum, E. Um, M. Wissotski, Y. Yu, A. Zuccolo, M. ur Rahman, D.G. Peterson, R.A. Wing, J.F. Wendel, and A.H. Paterson. 2010. A draft physical map of a D-genome cotton species (Gossypium raimondii). *BMC genomics* 11:395.

Lu, W., X. Chu, Y. Li, C. Wang, and X. Guo. 2013. Cotton GhMKK1 induces the tolerance of salt and drought stress, and mediates defence responses to pathogen infection in transgenic Nicotiana benthamiana. *PloS one* 8:e68503.

Mahajan, S., and N. Tuteja. 2005. Cold, salinity and drought stresses: an overview. *Archives of biochemistry and biophysics* 444:139-158.

Marsit, C.J., K. Eddy, and K.T. Kelsey. 2006. MicroRNA responses to cellular stress. *Cancer research* 66:10843-10848.

Mohorianu, I., F. Schwach, R. Jing, S. Lopez-Gomollon, S. Moxon, G. Szittya, K. Sorefan, V. Moulton, and T. Dalmay. 2011. Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *The Plant journal : for cell and molecular biology* 67:232-246.

Murakami, Y., T. Yasuda, K. Saigo, T. Urashima, H. Toyoda, T. Okanoue, and K. Shimotohno. 2006. Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. *Oncogene* 25:2537-2545.

Pang, M., A.W. Woodward, V. Agarwal, X. Guan, M. Ha, V. Ramachandran, X. Chen, B.A. Triplett, D.M. Stelly, and Z.J. Chen. 2009. Genome-wide analysis reveals rapid and

dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). *Genome Biology* 10:

Park, W., B.E. Scheffler, P.J. Bauer, and B.T. Campbell. 2010. Identification of the family of aquaporin genes and their expression in upland cotton (Gossypium hirsutum L.). *BMC plant biology* 10:142.

Paterson, A.H., J.F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, D. Llewellyn, K.C. Showmaker, S. Shu, J. Udall, M.J. Yoo, R. Byers, W. Chen, A. Doron-Faigenboim, M.V. Duke, L. Gong, J. Grimwood, C. Grover, K. Grupp, G. Hu, T.H. Lee, J. Li, L. Lin, T. Liu, B.S. Marler, J.T. Page, A.W. Roberts, E. Romanel, W.S. Sanders, E. Szadkowski, X. Tan, H. Tang, C. Xu, J. Wang, Z. Wang, D. Zhang, L. Zhang, H. Ashrafi, F. Bedon, J.E. Bowers, C.L. Brubaker, P.W. Chee, S. Das, A.R. Gingle, C.H. Haigler, D. Harker, L.V. Hoffmann, R. Hovav, D.C. Jones, C. Lemke, S. Mansoor, M. ur Rahman, L.N. Rainville, A. Rambani, U.K. Reddy, J.K. Rong, Y. Saranga, B.E. Scheffler, J.A. Scheffler, D.M. Stelly, B.A. Triplett, A. Van Deynze, M.F. Vaslin, V.N. Waghmare, S.A. Walford, R.J. Wright, E.A. Zaki, T. Zhang, E.S. Dennis, K.F. Mayer, D.G. Peterson, D.S. Rokhsar, X. Wang, and J. Schmutz. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492:423-427.

Pu, L., Q. Li, X. Fan, W. Yang, and Y. Xue. 2008. The R2R3 MYB transcription factor GhMYB109 is required for cotton fiber development. *Genetics* 180:811-820.

Qiu, C.X., F.L. Xie, Y.Y. Zhu, K. Guo, S.Q. Huang, L. Nie, and Z.M. Yang. 2007. Computational identification of microRNAs and their targets in Gossypium hirsutum expressed sequence

tags. *Gene* 395:49-61.

Rajagopalan, R., H. Vaucheret, J. Trejo, and D.P. Bartel. 2006. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes & development* 20:3407-3425.

Ranjan, A., N. Pandey, D. Lakhwani, N.K. Dubey, U.V. Pathre, and S.V. Sawant. 2012. Comparative transcriptomic analysis of roots of contrasting Gossypium herbaceum genotypes revealing adaptation to drought. *BMC genomics* 13:680.

Schwab, R., J.F. Palatnik, M. Riester, C. Schommer, M. Schmid, and D. Weigel. 2005. Specific effects of microRNAs on the plant transcriptome. *Developmental cell* 8:517-527.

Somerville, C., S. Bauer, G. Brininstool, M. Facette, T. Hamann, J. Milne, E. Osborne, A. Paredez, S. Persson, T. Raab, S. Vorwerk, and H. Youngs. 2004. Toward a systems approach to understanding plant cell walls. *Science* 306:2206-2211.

Sunkar, R. 2010. MicroRNAs with macro-effects on plant stress responses. *Seminars in cell & developmental biology* 21:805-811.

Sunkar, R., A. Kapoor, and J.K. Zhu. 2006. Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell* 18:2051-2065.

Trivedi, I., A. Ranjan, Y.K. Sharma, and S. Sawant. 2012. The histone H1 variant accumulates in response to water stress in the drought tolerant genotype of Gossypium herbaceum L. *The protein journal* 31:477-486.

Walford, S.A., Y. Wu, D.J. Llewellyn, and E.S. Dennis. 2011. GhMYB25-like: a key factor in early cotton fibre development. *The Plant journal : for cell and molecular biology* 65:785-797.

Wang, K., Z. Wang, F. Li, W. Ye, J. Wang, G. Song, Z. Yue, L. Cong, H. Shang, S. Zhu, C. Zou, Q. Li, Y. Yuan, C. Lu, H. Wei, C. Gou, Z. Zheng, Y. Yin, X. Zhang, K. Liu, B. Wang, C. Song, N. Shi, R.J. Kohel, R.G. Percy, J.Z. Yu, Y.X. Zhu, and S. Yu. 2012. The draft genome of a diploid cotton Gossypium raimondii. *Nature genetics* 44:1098-1103.

Wang, M., Q. Wang, and B. Zhang. 2013. Response of miRNAs and their targets to salt and drought stresses in cotton (Gossypium hirsutum L.). *Gene* 530:26-32.

Wang, S., J.W. Wang, N. Yu, C.H. Li, B. Luo, J.Y. Gou, L.J. Wang, and X.Y. Chen. 2004. Control of plant trichome development by a cotton fiber MYB gene. *Plant Cell* 16:2323-2334.

Wei, L.Q., L.F. Yan, and T. Wang. 2011. Deep sequencing on genome-wide scale reveals the unique composition and expression patterns of microRNAs in developing pollen of Oryza sativa. *Genome biology* 12:R53.

Xie, F., G. Sun, J.W. Stiller, and B. Zhang. 2011. Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database. *PloS one* 6:e26980.

Xie, F., P. Xiao, D. Chen, L. Xu, and B. Zhang. 2012. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant molecular biology*

Xue, J., F. Yang, and J. Gao. 2009. Isolation of Rh-TIP1;1, an aquaporin gene and its expression in rose flowers in response to ethylene and water deficit. *Postharvest Biology and Technology* 51:407-413.

Xue, W., Z. Wang, M. Du, Y. Liu, and J.Y. Liu. 2013. Genome-wide analysis of small RNAs reveals eight fiber elongation-related and 257 novel microRNAs in elongating cotton fiber cells. *BMC genomics* 14:629.

Yamada, K., R. Matsushima, M. Nishimura, and I. Hara-Nishimura. 2001. A slow maturation of a cysteine protease with a granulin domain in the vacuoles of senescing Arabidopsis leaves. *Plant physiology* 127:1626-1634.

Yang, L., L. Xu, and L. He. 2009. A CitationRank algorithm inheriting Google technology designed to highlight genes responsible for serious adverse drug reaction. *Bioinformatics* 25:2244-2250.

Yang, X., L. Wang, D. Yuan, K. Lindsey, and X. Zhang. 2013. Small RNA and degradome sequencing reveal complex miRNA regulation during cotton somatic embryogenesis. *Journal of Experimental Botany* 64:1521-1536.

Yao, D., X. Zhang, X. Zhao, C. Liu, C. Wang, Z. Zhang, C. Zhang, Q. Wei, Q. Wang, H. Yan, F. Li, and Z. Su. 2011. Transcriptome analysis reveals salt-stress-regulated biological processes and key pathways in roots of cotton (Gossypium hirsutum L.). *Genomics* 98:47-55.

Yin, Z., Y. Li, J. Yu, Y. Liu, C. Li, X. Han, and F. Shen. 2012. Difference in miRNA expression profiles between two cotton cultivars with distinct salt sensitivity. *Molecular biology reports* 39:4961-4970.

Zhang, B., Q. Wang, K. Wang, X. Pan, F. Liu, T. Guo, G.P. Cobb, and T.A. Anderson. 2007. Identification of cotton microRNAs and their targets. *Gene* 397:26-37.

Zhang, H., Q. Wan, W. Ye, Y. Lv, H. Wu, and T. Zhang. 2013. Genome-Wide Analysis of Small RNA and Novel MicroRNA Discovery during Fiber and Seed Initial Development in Gossypium hirsutum. L. *PLoS One* 8:e69743.

Zhang, L., D. Xi, S. Li, Z. Gao, S. Zhao, J. Shi, C. Wu, and X. Guo. 2011. A cotton group C MAP kinase gene, GhMPK2, positively regulates salt and drought tolerance in tobacco. *Plant molecular biology* 77:17-31.

Zhao, J., Y. Gao, Z. Zhang, T. Chen, W. Guo, and T. Zhang. 2013. A receptor-like kinase gene (GbRLK) from Gossypium barbadense enhances salinity and drought-stress tolerance in Arabidopsis. *BMC plant biology* 13:110.

Zhao, Y.T., M. Wang, S.X. Fu, W.C. Yang, C.K. Qi, and X.J. Wang. 2012. Small RNA profiling in two Brassica napus cultivars identifies microRNAs with oil production- and development-correlated expression and new small RNA classes. *Plant physiology* 158:813-823.

Table 4-1. Small RNA categorization in cotton *

	Unique (C)	Redundant (C)	Unique (D)	Redundant (D)	Unique (S)	Redundant (S)
Matched (E/G)	246,723 (4.31%)	2,740,263 (17.01%)	279,212 (3.65%)	2,714,775 (14.43%)	248,892 (3.90%)	2,800,217 (16.53%)
Matched (G)	1,775,856 (31.02%)	8,740,467 (54.26%)	2,314,399 (30.25%)	9,198,501 (48.90%)	1,965,288 (30.82%)	8,764,559 (51.74%)
Matched (A)	1,847,787 (32.27%)	9,229,827 (57.29%)	2,398,328 (31.35%)	9,719,543 (51.67%)	2,040,086 (31.99%)	9,281,862 (54.80%)
miRNA	29,145 (0.51%)	4,265,679 (26.48%)	31,229 (0.41%)	3,761,665 (20.00%)	31,343 (0.49%)	4,114,428 (24.29%)
rRNA	87,185 (1.52%)	1,230,970 (7.64%)	95,647 (1.25%)	1,117,486 (5.94%)	80,630 (1.26%)	1,038,386 (6.13%)
snRNA	1,250 (0.02%)	2,041 (0.01%)	1,532 (0.02%)	2,619 (0.01%)	1,353 (0.02%)	2,338 (0.01%)
snoRNA	486 (0.01%)	749 (0.00%)	611 (0.01%)	1,148 (0.01%)	530 (0.01%)	746 (0.00%)

tRNA	8,586 (0.15%)	182,011 (1.13%)	10,198 (0.13%)	201,042 (1.07%)	7,410 (0.12%)	151,590 (0.89%)
unann	5,598,655 (97.79%)	10,427,939 (64.73%)	7,511,572 (98.18%)	13,725,037 (72.97%)	6,256,412 (98.10%)	11,631,188 (68.67%)
Total	5,725,308	16,109,390	7,650,789	18,808,997	6,377,678	16,938,676

*: the number represented the raw data generated directly from deep sequencing; C: Control; S: Salt; D: Drought; Matched (E/G): Matched to EST and GSS of upland cotton; Matched (G): Matched to *ramondii*'s genome; Matched (A): Matched to EST and GSS of upland cotton and *ramondii*'s genome; unan: Unannotated.

Table 4-2. The similarity of the three cotton small RNA libraries treated by control drought, and salt.

Jaccard Index	Control	Drought	Salt
Control	-	42.27%	46.94%
Drought	42.27%	-	97.39%
Salt	46.94%	97.39%	-

Table 4-3. The expression of conserved miRNA families among control (C), drought (D), and salt (S) treatments

miRNA	miRNA normalization*			Fold change			Statistical significance			Expression significance		
	C	D	S	D/C	S/C	S/D	D vs C	S vs C	S vs D	D vs C	S vs C	S vs D
miR156	30118.5	25891.5	42784.9	-0.44	0.43	0.88	**	**	**			
miR157	129929.3	64844.7	103178.0	-1.23	-0.41	0.82	**	**	**	**		
miR158	4.7	5.4	3.8	-0.04	-0.39	-0.36			*			
miR159	149.7	151.6	484.3	-0.20	1.62	1.83		**	**		**	**
miR160	130.9	132.2	65.5	-0.21	-1.07	-0.86		**	**		**	
miR161	4.7	5.2	3.8	-0.06	-0.35	-0.29						
miR162	73.9	53.2	71.4	-0.70	-0.12	0.58	**		**			
miR163	3.5	0.5	0.2	-3.11	-4.39	-1.28	**	**		**	**	
miR164	75.7	61.4	105.7	-0.53	0.41	0.94	**	**	**			
miR165	72.8	53.0	49.8	-0.68	-0.62	0.06	**	**				
miR166	13680.8	9704.4	12182.0	-0.72	-0.24	0.48	**	**	**			
miR167	3984.1	2999.4	3004.2	-0.63	-0.48	0.15	**	**				
miR168	1116.1	1017.5	1068.7	-0.36	-0.14	0.22	**	**	**			
miR169	176.7	128.6	101.8	-0.68	-0.87	-0.19	**	**	**			
miR170	0.6	1.3	0.8	1.03	0.39	-0.64	*			*		
miR171	166.1	179.4	129.1	-0.11	-0.44	-0.32	**	**	**			
miR172	477.9	394.1	316.5	-0.50	-0.67	-0.17	**	**	**			

miR173	2.3	0.0	0.0	-3.83	-3.83	0.00	**	**		**	**	
miR319	1.9	1.1	4.3	-1.08	1.09	2.17	*	**	**	*	**	**
miR390	237.7	230.3	215.1	-0.27	-0.22	0.05		**	**			
miR391	0.0	0.3	0.0	0.76	0.00	-0.76	*		*			
miR393	5.7	4.4	6.7	-0.61	0.16	0.78			**			
miR394	4.2	9.0	17.0	0.87	1.94	1.07	**	**	**		**	**
miR395	11.9	5.4	9.6	-1.35	-0.38	0.97	**	*	**	**		
miR396	387.2	274.8	363.8	-0.72	-0.16	0.56	**	**	**			
miR397	860.2	662.7	964.1	-0.60	0.09	0.69	**	**	**			
miR398	45.3	40.0	63.0	-0.40	0.40	0.81	*	**	**			
miR399	11.4	12.5	9.5	-0.09	-0.34	-0.25			**			
miR400	0.0	0.3	0.0	0.50	0.00	-0.50	*		*			
miR401	1.2	9.7	0.0	2.74	-2.95	-5.68	**	**	**	**	**	**
miR403	23.3	22.9	18.9	-0.25	-0.37	-0.13		**	**			
miR407	0.0	0.3	0.0	0.50	0.00	-0.50	*		*			
miR408	226.3	150.5	308.5	-0.81	0.37	1.19	**	**	**			**
miR413	0.0	0.0	0.6	0.00	1.80	1.80		**	**		**	**
miR414	8.4	3.1	9.6	-1.64	0.12	1.76	**		**	**		**
miR415	38.9	34.1	39.2	-0.41	-0.06	0.35	*		*			
miR416	3.9	0.1	0.0	-6.42	-4.60	1.82	**	**		**	**	

miR417	0.0	3.9	0.3	4.37	0.80	-3.57	**	*	**	**		**
miR418	17.6	20.8	21.7	0.01	0.23	0.21	*	**				
miR419	1.1	2.4	62.6	0.91	5.74	4.83	**	**	**		**	**
miR420	0.3	5.4	14.6	3.89	5.48	1.59	**	**	**	**	**	**
miR426	0.3	0.3	0.5	-0.45	0.53	0.98						
miR437	0.6	1.0	0.4	0.40	-0.66	-1.06						
miR440	0.0	0.0	0.6	0.00	1.80	1.80		**	**		**	**
miR442	2.2	0.0	1.1	-3.75	-1.03	2.73	**	*	**	**	*	**
miR443	5.4	5.3	5.1	-0.25	-0.14	0.10						
miR444	5.0	28.6	0.1	2.28	-5.48	-7.77	**	**	**	**	**	**
miR445	0.0	0.2	0.0	0.18	0.00	-0.18						
miR446	1.0	0.0	0.1	-2.62	-3.14	-0.52	**	**		**	**	
miR447	10.6	8.9	13.3	-0.48	0.25	0.73		*	**			
miR472	0.7	6.2	0.7	2.95	-0.02	-2.97	**		**	**		**
miR473	9.8	3.1	3.7	-1.87	-1.47	0.40	**	**		**	**	
miR474	5.1	7.0	8.3	0.24	0.64	0.40	*	**				
miR475	0.0	0.3	0.0	0.50	0.00	-0.50	*		*			
miR476	2.4	2.1	1.9	-0.41	-0.39	0.02						
miR477	37.9	24.2	32.8	-0.87	-0.28	0.59	**	*	**			
miR478	0.0	7.4	0.5	5.31	1.48	-3.83	**	**	**	**	**	**

miR479	0.3	0.2	0.3	-1.18	-0.14	1.04					
miR480	0.0	0.6	0.0	1.76	0.00	-1.76	**		**	**	**
miR482	37.5	27.2	29.8	-0.69	-0.40	0.28	**	**			
miR528	0.8	0.4	0.6	-1.15	-0.52	0.62				*	*
miR529	9.1	7.9	7.4	-0.42	-0.37	0.05					
miR535	1486.8	1097.5	1382.5	-0.66	-0.18	0.48	**	**	**		
miR774	415.3	365.6	421.2	-0.41	-0.05	0.36	**		**		
miR894	569.7	337.0	384.4	-0.98	-0.64	0.34	**	**	**		
miR2089	1578.8	1149.4	1585.8	-0.68	-0.07	0.62	**		**		
miR2911	621.8	762.6	477.7	0.07	-0.45	-0.52	**	**	**		
miR3476	640.4	562.9	535.0	-0.41	-0.33	0.08	**	**	**		
miR3954	69714.5	82391.4	64928.4	0.02	-0.18	-0.19	**	**	**		
miR4377	2086.5	1551.5	2047.8	-0.65	-0.10	0.55	**	*	**		

*:All miRNA families with redundant abundance were normalized to transcript expression levels per million reads (RPM). If the original miRNA expression in a library was zero, the normalized expression was adjusted to 0.01 according to a previous report (Murakami et al., 2006). miRNA expression fold change in any two libraries was calculated with the formula, Fold change=$\log_2$(treatment 1 / treatment 2) (Marsit et al., 2006). A 2 × 2 contingency table was used to perform Pearson's chi-squared test for significance of miRNA expression from two samples. Fold change and *p*-value were combined to determine the final miRNA expression significance. We defined expression difference level as following rules: extremely significant (**) if (fold change ≥ 1 or fold change ≤ -1) and *p-value* ≤ 0.01; significant (*) if (fold change ≥ 1 or fold change ≤ -1) and 0.05 ≥ *p*-value > 0.01; otherwise insignificant.

Table 4-4. Stress-, resistance-, and fiber-related miRNAs, miRNA targets, GO terms, and KEGG pathways in cotton

Function type	miRNAs	Targets	Cellular component	Biological process	Molecular function	Pathway
Apoptosis	14	4	3	4	3	
Cell cycle	13	7	1	7	3	
Cell migration	23	15	2	2	3	
Circadian clock	2	3	1	12	4	1
Development	46	17	11	32	20	11
Fiber development	163	210	41	132	94	21
Gossypol biosynthesis	3	1	1	1	1	
Metabolism	181	239	48	198	187	57
Signal transduction	110	104	27	85	46	6
Stress response	151	229	40	136	104	22
Transcription factor	183	217	29	131	47	4

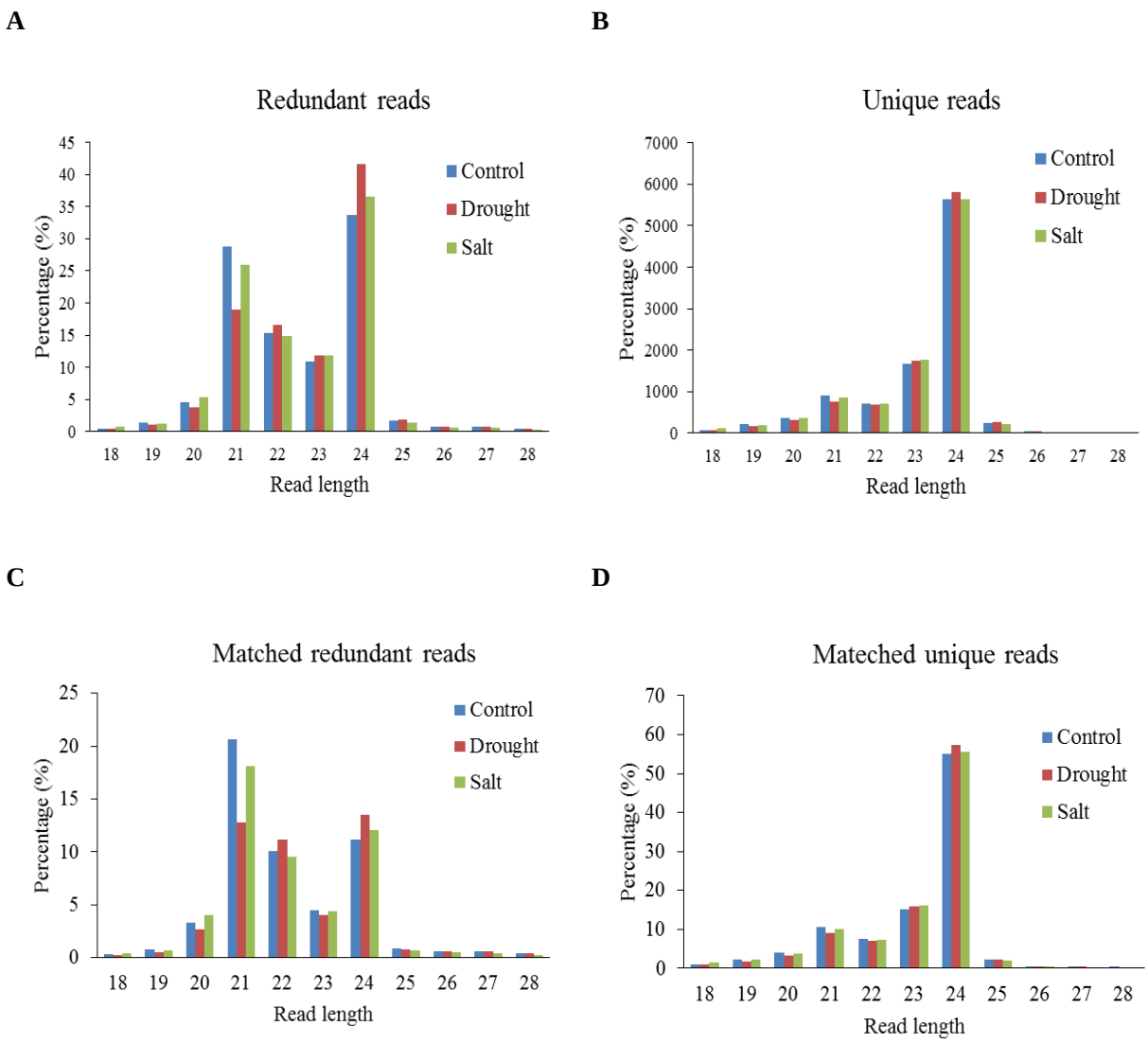


Figure 4-1. Size distribution of redundant and unique small RNA reads in cotton. **A and C:** Size distribution of redundant sRNA reads from control, drought, and salt libraries. **B and D:** Size distribution of unique sRNA reads from control, drought, and salt libraries. C and D: small RNA reads were fully mapped back to EST and GSS of upland cotton and *G. ramondii* genome.

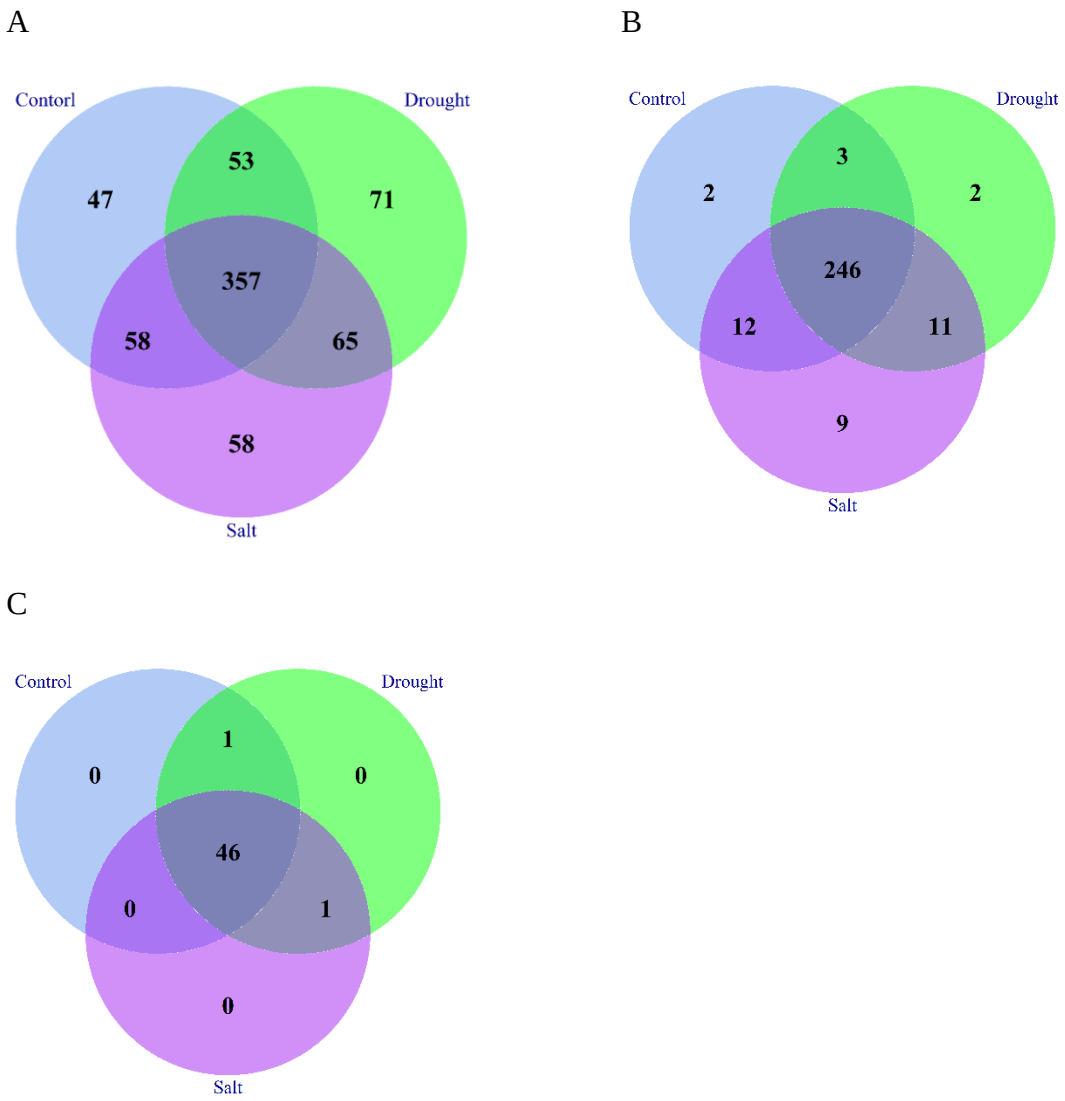


Figure 4-2. Distribution of miRNAs in control, drought and salinity treatment. A: conserved miRNA families; B: known miRNAs; C: novel miRNAs.

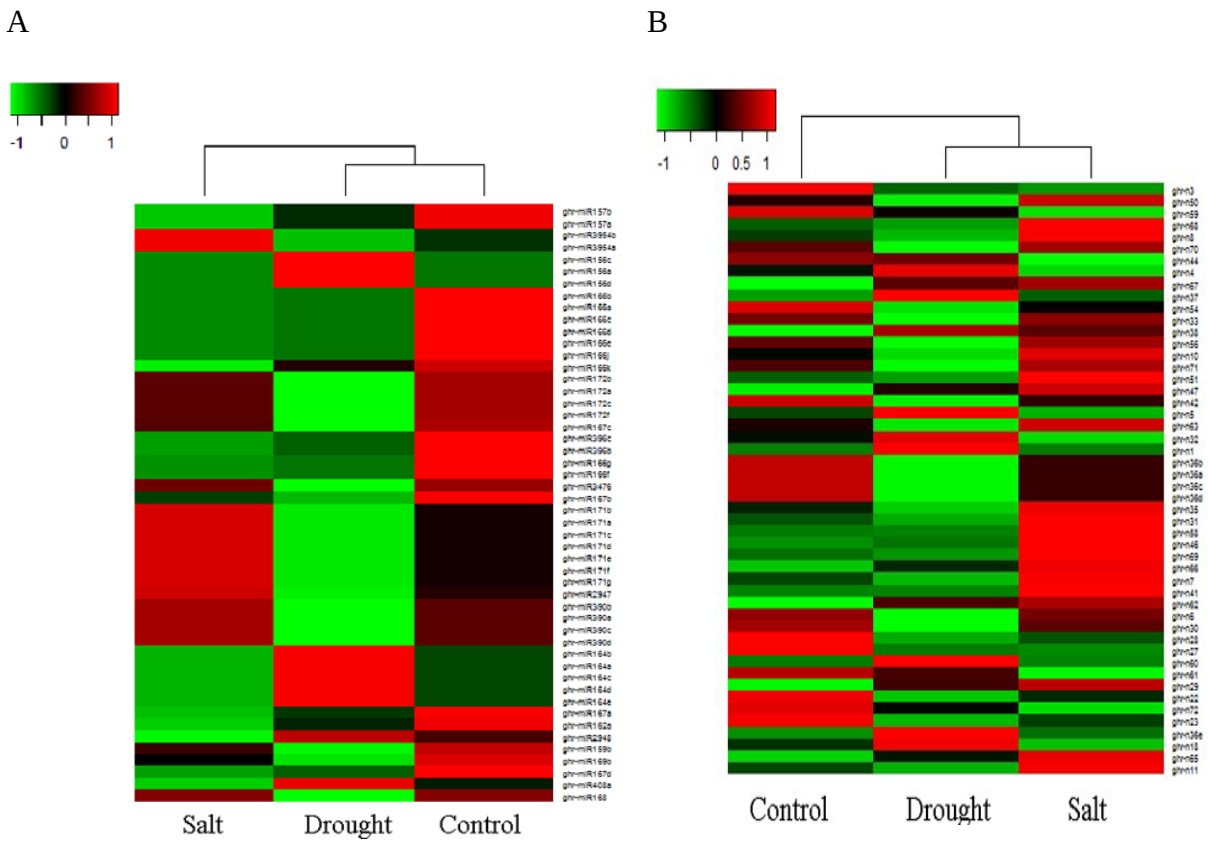


Figure 4-3. Heatmaps of A) top 50 abundant conserved miRNAs and B) top 50 abundant novel miRNAs in control, salt, and drought libraries in cotton. Red: up-regulated; green: down-regulated.

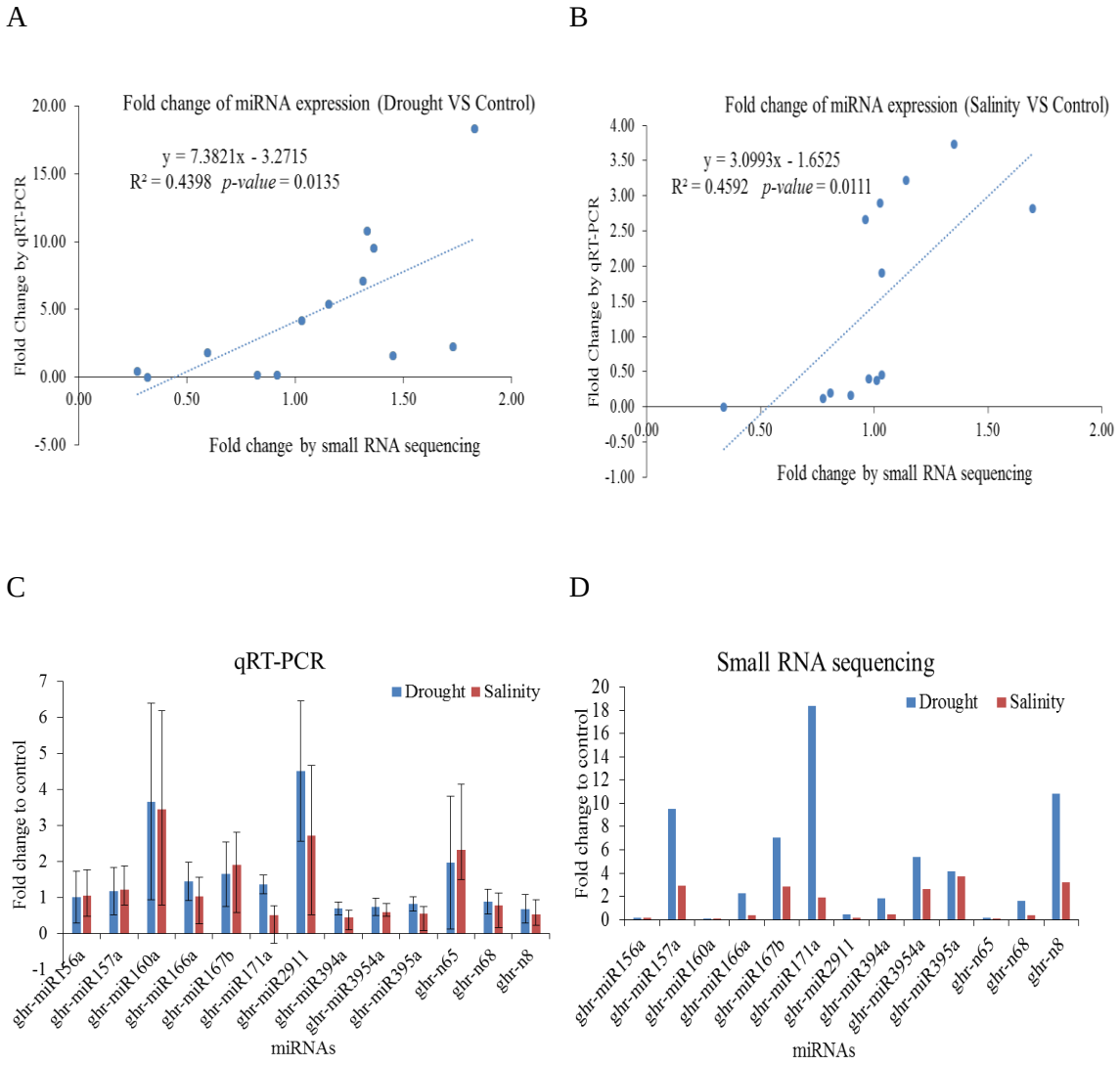
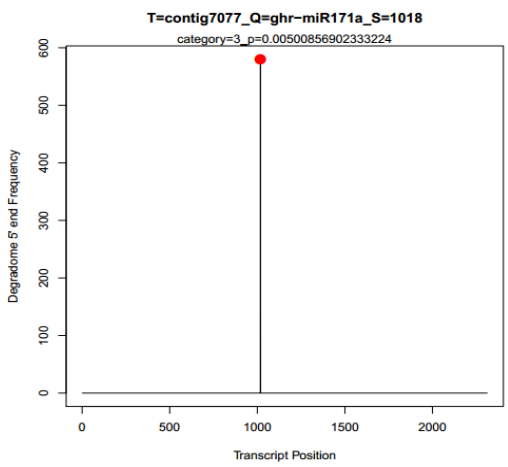


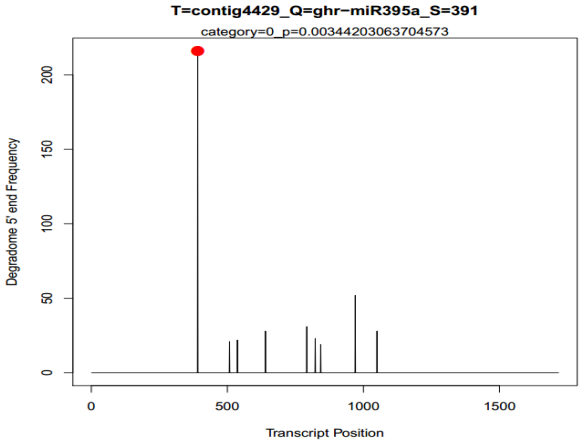
Figure 4-4. Validation and comparison of 13 cotton miRNAs' expression between qRT-PCR and small RNA sequencing. miRNA expression correlation between qRT-PCR and small RNA sequencing (A: drought VS control and B: salinity VS control). miRNA expression fold change to the control library (C: qRT-PCR and D: small RNA sequencing). qRT-PCR assay of each miRNA expression was performed on ten-day-old cotton seedlings in three biological replicates, which were treated with control, drought and salinity.

A



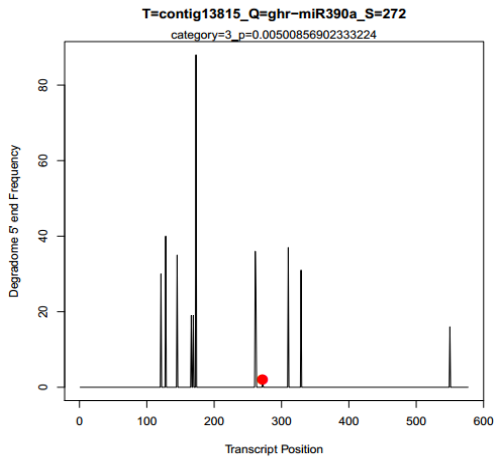

```
3' CUAUAACCGUGCCGAGUUAGU 5' ghr-miR171a-g
   |||||||||o||||||||||
5' GAUAUUGGCGCGGCUCAAUCA 3' contig7077
Scarecrow-like protein 6-like
```

B



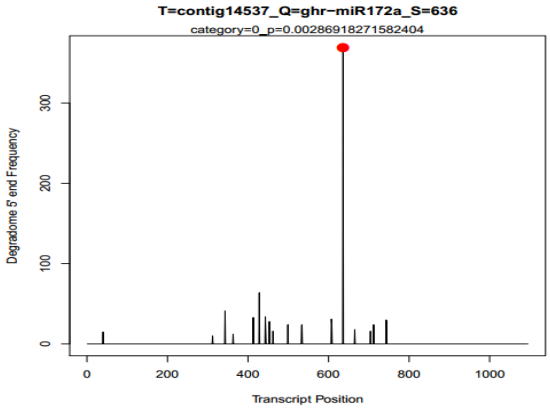

```
3' CUCAAGGGGGUUUGUGAAGUC 5' ghr-miR395a/b
   |||||||o|||||| |||||
5' GAGUUCCUCCAAACUCUUCAU 3' contig4429
Sulfate adenylyltransferase
```

C




```
3' CCGCGAUAGGGAGGACUCGAA 5' ghr-miR390a-d
   ||||||o|||||||||||||||
5' AGCCCUGUCCCUCCUGAGCUG 3' contig13815
DEAD-box ATP-dependent RNA helicase 21-like
```

D




```
3' UACGUCG-UAGUAGUUCUAAGA 5' ghr-miR172a/b/c/f
   | ||||| ||||||o|||||||
5' AAGCAGCAAUCAUCGAGAUUCU 3' contig14537
Avr9/Cf-9 rapidly elicited protein
```

Figure 4-5. Cotton miRNA target alignment and its T-plot validated by degradome sequencing. A:

ghr-miR171a-g and contig7077 (Scarecrow-like protein 6-like); B: ghr-miR390a-d and contig13815 (DEAD-box ATP-dependent RNA helicase 21-like); C: ghr-miR395a/b and contig4429 (Sulfate adenylyltransferase); and D: ghr-miR172a/b/c/f and contig14537 (Avr9/Cf-9 rapidly elicited protein). Both red arrow and red dot represent slice site on miRNA target.

A **B**

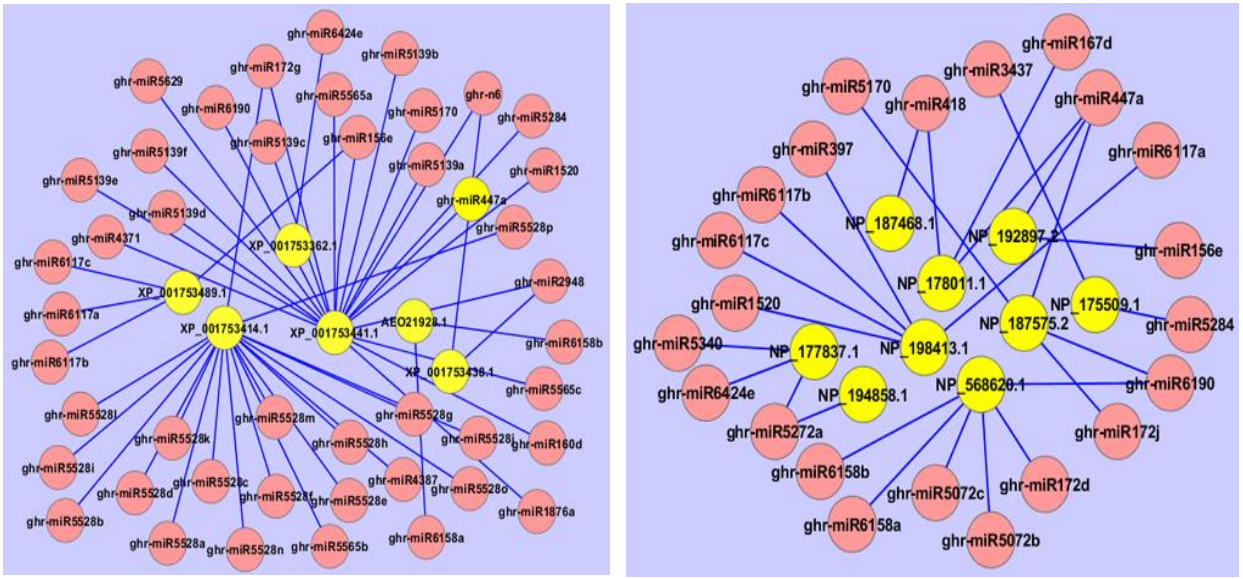


Figure 4-6. Top-ranked miRNA regulation networks involved in (A) drought response and (B) salinity response in cotton.

CHAPTER 5: Small RNA sequencing identifies miRNA roles in fiber development

Abstract

MicroRNAs (miRNAs) have been found to be differentially expressed during cotton fiber development. However, what miRNAs and how they are involved in fiber development is unclear. Here, using deep sequencing, 65 conserved miRNA families were identified; 32 families were differentially expressed between leaf and ovule. At least 40 miRNAs were either leaf or ovule-specific, whereas 62 miRNAs were shared in both leaf and ovule. qRT-PCR confirmed these miRNAs were differentially expressed during fiber early development. A total of 820 genes were potentially targeted by 99 identified miRNA families, whose functions are involved in a series of biological processes including fiber development, metabolism, and signal transduction. 22 predicted miRNA-target pairs were subsequently validated by degradome sequencing analysis. GO and KEGG analyses showed that the identified miRNAs and their targets were classified to 1,027 GO terms including 568 biological processes, 324 molecular functions, and 135 cellular components, and were enriched to 78 KEGG pathways. At least 7 unique miRNAs participate in trichome regulatory interaction network. Eleven trans-acting siRNA (tasiRNA)-derived candidate genes were also identified in cotton. One has never been found in other plant species and two of them are *MYB* and *ARF,* both play important role in cotton fiber development. Sixteen genes were predicted to be tasiRNA targets, including sucrose synthase and *MYB2*. Together, this study discovered new miRNAs in cotton and offered evidences that miRNAs play important roles in cotton ovule/fiber development. The identification of tasiRNA genes and their targets broadens

our understanding of the complicated regulatory mechanism of miRNAs in cotton.

Introduction

Cotton is one of most important economic crops because that it produces the majority of nature textile fiber material in the world. The most widely planted cotton is a allotetraploid (AADD) species, upland cotton (*Gossypium hirsutum*), derived from the union of A-genome *Gossypium* species and D-genome *Gossypium* species around 1-2 million years ago (Wendel, 1989). The genome expansion of upland cotton resulted in a great improvement of fiber quality and yields, such as much longer and stronger fibers than those in ancestor species (Adams et al., 2003). Cotton fibers are seed trichomes, whose development undergo four major overlapping stages (initiation, elongation, secondary cell wall biosynthesis, and maturation) (Qin and Zhu, 2011). Fiber development is initiated through the differentiation of partial epidermal cells on fertilized ovule into seed trichome cells. After initiation, single-celled seed trichomes subsequently perform endoduplication and rapid elongation (Larkin et al., 2003).

Cotton fiber is an excellent model for studying cell differentiation and elongation. Interestingly, cotton fiber development shares many similarities with *Arabidopsis* leaf trichome development. Decades of study has discovered at least two major types of regulators involved in regulating leaf trichome development, including both positive and negative regulators. Leaf trichome development in *Arabidopsis* is promoted by positive transcription regulators, including the *R2R3 MYB* protein, *WD40* domain–containing protein and a basic helix-loop-helix–related protein (*bHLH*), like *GLABROUS1* (*GL1*), *TRANSPARENT TESTA GLABRA1* (*TTG1*), *GLABRA3*

(*GL3*), and *ENHANCER of GL3* (*EGL3*). The negative regulators consist of a family of single-repeat *R3 MYB* transcription factors including *TRIPTYCHON* (*TRY*), *CAPRICE* (*CPC*), *ENHANCER of TRY and CPC1* (*ETC1*, 2, and 3), and *TRICHOMELESS1* (*TCL1*) (Pesch and Hulskamp, 2009). There are many gene homoeologues of positive and negative regulators in cotton, some of which are expressed differentially during fiber development and even are capable of rescuing trichome formation in trichomeless mutants in *Arabidopsis*. The most representative genes are *MYB*-domain transcription factors. For instance, *GhMYB109* encodes a *R2R3 MYB* transcription factor expressed specifically in fiber initials and elongating fibers of upland cotton (Suo et al., 2003). Transferring the *GL1* homolog (*GaMYB2*) of cotton into *Arabidopsis* rescued trichome formation of a *gl1* mutant, and also induced sporadic seed trichomes on *Arabidopsis*. This indicates that cotton might share similar regulators in fiber trichome development with those in *Arabidopsis* leaf trichome (Wang et al., 2004). In addition another *R2R3 MYB* transcription factor, *GhMYB25,* was identified to express specifically in cotton fiber initiation and elongation, whose silence shortens fiber and reduces fiber trichomes and other trichomes in cotton (Machado et al., 2009).

miRNAs are a class of small non-coding RNAs in length of ~22 nt, negatively regulating gene expression by either mRNA degradation or translation inhibition. miRNAs play important roles in a series of biological processes such as development, cell proliferation, stress response, and metabolism (Eldem et al., 2013; Zhang and Wang, 2014). Both computational prediction and experimental approaches show that many transcription factors in plants including *TCP* and *MYBs* are miRNA targets (Bartel, 2004; Pang et al., 2009; Qiu et al., 2007). miR319a was demonstrated

to be critical for petal growth and development through targeting *TCP4* transcription factor in *Arabidopsis* (Nag et al., 2009). Recently, homologous *MYB2* gene was reported to be targeted by miR828 and miR858 during fiber initiation, resulting in biogenesis of trans-acting siRNA (tasiRNA) in the *TAS4* family (Guan et al., 2014). Moreover, some miRNAs were even found to directly regulate *MYB* gene expression. A good examples is the *Arabidopsis GAMYB*-like genes, *MYB33* and *MYB65*, which redundantly facilitate anther development and is cross-regulated by miR319 and miR159 (Palatnik et al., 2007). To date, some miRNAs are identified in cotton, many of which were detected to be expressed differentially during fiber development as well as in other tissue development (Gong et al., 2013; Wang et al., 2012b; Xue et al., 2013). Taken collectively, cotton miRNAs may play important roles in modulating fiber initation and elongation, demonstrating great potential in improving fiber quality and yields in the future (Lu et al., 2013; Pang et al., 2009; Wang et al., 2012b).

Recently, the emergence of next-generation sequencing technologies (NGS) has unprecedently promoted identification of miRNAs and their targets (Kozomara and Griffiths-Jones, 2011; Mi et al., 2013). To date, there are 337 and 713 miRNAs identified from Arabidopsis and rice, respectively (miRBase: Release 20) (Kozomara and Griffiths-Jones, 2011). Upland cotton, as a tetraploid species (AADD, 2n = 4x = 52) (Chen et al., 2007), should inherit two sets of miRNA system from A- and D-genome diploid cotton species in theory, and likely owns more miRNAs than Arabidopsis (2x= 10) and rice (2x=24). However, using computational prediction and small RNA sequencing, about 80 conserved miRNAs and novel miRNAs have been identified in cotton, but many of which were found to have no precursor and are merely homologous with mature

miRNAs from other plant species (Chen et al., 2007; Kwak et al., 2009; Lu et al., 2013; Pang et al., 2009; Qiu et al., 2007; Wang et al., 2012b; Zhang et al., 2013). Therefore, the complicated tetraploid genome of cotton apparently restrict miRNAs identification in cotton. As a result, miRNAs identification and functional validation in cotton still lags behind those of model species, particularly for cotton fiber development.

In this study, small RNA libraries from cotton leaf and ovule were sequenced. Multiple cotton datasets including EST and GSS databases of upland cotton and D-genome of *Gossypium raimondi*, were used in order to systematically identify miRNAs in cotton. Compared with previous studies for cotton miRNA identification, significantly more miRNA precursors were detected out in our result. Many transcription factors and other genes important for development of fiber were predicted to be miRNA targets. In addition to miRNAs, we also identified 11 potential tasiRNAs-derived genes, many of which also might be involved in fiber development. Thus, our study could contribute to our understanding miRNA roles in fiber development.

Results

Deep sequencing of small RNA libraries of cotton ovule and fiber

After removing low quality reads [low-quality reads; reads in length ≤ 15; reads with < 3 copies; and junk reads (≥ 80% A, C, G or T; ≥ 3 N; only A, C, or only G/T)], a total of 484,896 and 333,771 reads were generated from leaf and ovule, respectively, including 21,568 and 41,350 unique reads (Table 5-1). 74.34% redundant reads and 85.10% unique reads were 100% mapped to EST and GSS databases of upland cotton leaves, whereas 96.36% redundant and 96.54% unique

reads were fully matched back to EST and GSS databases in ovules. However, we only detected 59.46% and 25.36% reads mappable to the D genome *G raimondii*. In both leaf and ovule libraries, miRNA-derived reads that have ≤ 3 mismatches with known plant miRNA sequences accounted for significantly more reads than from other RNAs including tRNA, siRNA, and rRNA (Table 5-1). For redundant reads, unique reads, mappable redundant reads and mappable unique reads, all showed similar size distribution between leaf and ovule libraries, most of which were 21 -nt (Figure 5-1). The distribution of small RNA abundance and size in cotton leave were similar with the results reported in *Medicago truncatula* (Lelandais-Briere et al., 2009) and rice (Sunkar et al., 2008). However, the size distribution in cotton ovule is little bit of difference from cotton leave and other plant species; in cotton, except 21 nt length of small RNAs, 24 nt length small RNAs also show a large percentage in the small RNA libraries.

Identifying conserved miRNA families in cotton leaf and ovule

About 42.15% and 26.82% reads were identified to be miRNA homologs in leaf and ovule, respectively (Table 5-1). These reads belonged to a total of 65 conserved miRNA families, such as miR156/157, miR159, miR164, miR168, and miR395 (Table 5-2). Of these 29 miRNA families co-existed in both leaf and ovule, indicating their fundamental function for maintaining life activity in normal cell or tissue. However, 50 miRNA families were only detected in the leaf, whereas 45 families were only found in the ovule, suggesting the specific role of miRNAs for cotton leaf and fiber development. According to the comparison of miRNA family expression abundance, 32 out of 65 miRNA families exhibited remarkable expression difference between leaf and ovule,

including miR164, miR168, miR399, and miR397 (Table 5-2). To our surprise, only miR156/157 in leaf was the most abundant miRNA family, up to 30.90% in all conserved miRNA reads. In contrast, the miR156/167 in ovule was expressed in relatively low level (0.028%) (Table 5-2). The most abundant miRNA families in leaf were miR157, miR3954, miR166, miR156, and miR397, whereas miR166, miR167, miR172, miR3954, and miR2949 were the most abundant in ovule.

Identifying miRNA precursors from EST, GSS, and D genome sequences

It has been shown that miRNA and miRNA star are located on the opposite arms of miRNA precursors, owning a typical 2-nt overhangs on their 3' end. However, many conserved miRNA star were reported not to co-exist with their miRNAs in the same sequencing dataset (Meyers et al., 2008). To systematically identify miRNA precursors, we required both novel miRNAs and their miRNA stars to co-exist in one sequencing dataset. In this study, we identified a total of 128 miRNA precursors, based on the sequences from upland cotton EST and GSS, and D genome *G. raimondii* genome sequences; this included 120 conserved miRNAs and 8 novel miRNAs (Supplementary 5-1). Forty nine out of 128 miRNAs (38.3%) were found to have miRNA star. Thirty four and 6 miRNA stars were only detected in leaf and ovule, respectively. Of the 128 miRNAs, 23 miRNAs derived from the database of EST, GSS, or assembled EST, such as ghr-miR156c, ghr-n2, ghr-miR156d, ghr-miR166k, ghr-miR159c, and ghr-miR160d. Sixteen out of the 105 (15.2%) D-genome-derived miRNAs were found to have homolog counterparts from EST and GSS of upland cotton (Supplementary 5-1). The lengths of miRNAs were between 17 and 22 nt, of which the 21-nt is the largest percentage (68.8%) (Figure 5-2A). The majority of pre-miRNA

sizes ranged from 81 to 100 nt (39.8%), followed by the precursors in length of 58-80 nt (30.4%) and 101-150 nt (24.2%) (Figure 5-2B). 31 out of the 128 identified miRNAs were also detected by recent studies of cotton small RNA sequencing (Lu et al., 2013; Wang et al., 2012b). It is likely caused by criteria for miRNA identification, such as mismatches between miRNA and miRNA star, least small RNA reads, and required miRNA star or no. Furthermore, many leaf-specific miRNAs in our studies might also contribute to the difference with previous reports.

We also found 6 miRNA clusters, including 12 miRNAs, miR169g-miR169d, miR2947-miRn5, miRn6-miR477b, miR396d-miR396b, miR482a-miRn4, and miRn1-miR482d. Only miRn1- miR482d originates from EST (DT527030.1), while the others are identified from cotton D genome sequences.

miRNA expression in cotton leaf and ovule

59 conserved miRNAs and 3 novel miRNAs were expressed in both leaf and ovule. At least 12 miRNAs (11 conserved miRNAs and 1 novel miRNAs) were ovule-specific, since they were only found in the ovule library, including miR-n2, miR828a, and miR393d. In contrast, 50 conserved miRNAs and 4 novel miRNAs were leaf-specific, including miR156a, miR164a, and miR169b (Figure 5-3A and 5-3B). Expression abundance of these identified miRNAs differed sharply, scaling from 0 to 29,840 reads per million in leaf and ovule libraries. The miRNA expression difference between leaf and ovule were in absolute fold change of 0.03 -14.77. After Chi-squared test, 115 out of 128 (89.8%) identified miRNAs were expressed differentially at a significant level (*p-value* < 0.05 and |Fold change| ≥1) (Supplementary 5-1). Amongst the 115

differential miRNAs, 26 and 88 miRNAs were up-regulated in ovule and leaf, respectively, while the remainder were down-regulated.

miRNA target identification

A total of 2,534 upland cotton protein-coding mRNA sequences deposited in NCBI and the 21,991 assembled EST-derived coding contigs were combined as the data source for miRNA target prediction. From this, a total of 820 genes were predicted to be targeted by 99 cotton miRNA families belonging to 1,498 miRNA-target pairs (Supplementary 5-2). Panther-based protein classification shows that these miRNA targets could be categorized to 23 classes (Mi et al., 2013). Most targets are affiliated to transferase (PC00220) (19.10%), nucleic acid binding (PC00171) (16.20%), kinase (PC00137) (11.80%), hydrolase (PC00121) (8.80%), transcription factor (PC00218) (8.80%), and oxidoreductase (PC00176) (8.30%) (Figure 5-4).

Not surprisingly a series of genes important for fiber initiation and elongation were predicted to be miRNA targets; these genes include *AP2* transcription factor , tubulin, *MYB2* transcription factor, and cellulose synthase (Lee et al., 2007). According to these gene functions, we categorized these genes into 6 groups including transcription factors, stress response, signal transduction, and fiber development (Table 5-3). A total of 87 miRNA families and 498 genes were involved in these 6 groups. Our results also show that metabolism isthe most predominant among all predicted miRNA targets (60 miRNA families and 147 targets), followed by fiber development (36 miRNA families and 116 targets) and stress response (47 miRNAs and 94 targets).

miRNA target validation using degradome sequencing analysis

Degradome sequencing analysis has been become a powerful approach for identifying and

validating miRNA targets in plants through identifying the cleavage sites. Here, a total of 22 predicted miRNA-target pairs were able to be validated by degradome sequencing data, involving 9 unique mature miRNAs and 18 unique coding genes (Figure 5-5; Supplementary 5-2). At least 3 known miRNA-target pairs that had been validated in other species were also detected in cotton, including miR828-*MYB2D* (Figure 5-5A), miR164-*NAC* (Figure 5-5B) and miR160-*ARF* (Figure 5-5C) (Guan et al., 2014; Prigge et al., 2005). This indicates these miRNAs and their targets are also functionally conserved in cotton and likely participate in similar regulation of corresponding biological processes. In addition, we also validated some novel miRNA-target pairs, such as miR171-GRAS transcription factor (Figure 5-5D) and miR7484-RING/U-box protein.

Distribution of identified miRNAs and their targets in cotton D genome

One hundred twenty six (98.4%) miRNAs and 860 (99.5%) miRNA target genes were successfully mapped back to the cotton D genome under a cutoff of 95% identity (Supplementary 5-3). In addition, the other four miRNA target genes were mapped back to three scaffolds in the D genome. On average, 9 miRNAs and 66 miRNA targets were found in one chromosome with a density of 16 miRNAs and 114 miRNA targets per 10^8 nt, respectively. Chromosome 8 was found to have the largest miRNAs (21), whereas the largest miRNA targets (133) were from the longest chromosome 9. Interestingly, many of the identified cotton miRNAs and their targets were mapped on two ends of chromosomes in the D genome using circos analysis (Latchman, 1997) (Figure 5-6). We also checked the miRNA distribution on the genomes of Arabidopsis and rice. No significant biased miRNA distribution was detected (data not shown). Thus, we inferred the distribution of cotton miRNAs and their targets might hide a special unclear evolutionary story.

However, further evidence is needed to support this point.

Identification of tasiRNAs and their targets

Previous studies uncovered that phased tasiRNAs are generated by miRNA-triggered cleavage and employ a miRNA-like mechanism to act on their targets (Williams et al., 2005). There are at least 8 tasiRNA-derived genes identified in *Arabidopsis*, some of which are also conserved to target auxin response factors (*ARFs*) across plant species (Williams et al., 2005; Zhang et al., 2014). Non-miRNA and non-other-RNA reads from leaf and ovule were mapped back to EST and coding mRNA databases of upland cotton to detect the genes having phased siRNA clusters. The non-redundant sequences were obtained by removing the repeated sequences and were further used as a data source for miRNA target prediction. In the end 11 genes were predicted by 17 miRNAs in cotton, each of which contains 14 phased siRNAs on average (Table 5-4). Leaf and ovule libraries shared two tasiRNA-derived genes (ES834802 and DW502659). Eight out of the 11 genes are coding genes, including *MYB*-like DNA-binding domain protein (*MYB2*), beta-galactosidase, and ATP synthase. Two *ARF6* (*ARF6*: ES819493 and ES834779) were also predicted to be tasiRNA-derived genes in ovule, which are targeted by miR167b and miR167d (Table 5-4). However, in the plant tasiRNA database, we didn't find the example that miR167 could target a tasiRNA-derived gene of *ARF*s. We chose the most abundant reads in phased siRNAs as the representative tasiRNA for target prediction based on coding mRNA database of upland cotton in NCBI. Six tasiRNAs were found to target 16 genes including gland development related protein 79-like, sucrose synthase, polyphenol oxidase, and glycosyl hydrolase (Supplementary 5-4).

GO and KEGG pathway analysis

To better understand miRNA roles in fiber and leaf development, we performed an analysis of GO-based term classification and KEGG-based pathway enrichment. In total, 97 miRNA families and their 708 targets were categorized to 568 biological processes, 324 molecular functions, and 135 cellular components (Supplementary 5-3). The top 3 biological processes involved with miRNAs and their targets are single-organism process (GO:0044699), cellular process (GO:0009987), and response to stimulus (GO:0050896) (Figure 5-7A). Most miRNAs and their targets in cotton has the molecular function of binding (GO:0005488), catalytic activity (GO:0003824), nucleic acid binding transcription factor activity (GO:0001071) (Figure 5-7B). Similarly, the most frequent cellular component where most miRNA targets are located in are cell part (GO:0044464), membrane part (GO:0044425), and cell junction (GO:0030054) (Figure 5-7C).

Forty eight miRNAs and their 101 targets in cotton were enriched to 78 KEGG pathways including amino sugar and nucleotide sugar metabolism, biosynthesis of phenylpropanoids, biosynthesis of plant hormones, fatty acid biosynthesis, glycolysis / gluconeogenesis, oxidative phosphorylation, photosynthesis, starch and sucrose metabolism, and tryptophan metabolism (Supplementary 5-6). Fifty seven miRNAs and their 112 targets that were related to fiber development, were enriched to 12 different pathways, whereas transcript factors involving 84 miRNAs and 113 targets belonged to 5 pathways (Table 5-3). For instance, 4 UDP-D-glucose pyrophosphorylase genes and 4 sucrose synthase genes that were enriched to starch and sucrose metabolism, were targeted by ghr-miR162a, ghr-miR172e, ghr-miR172h, ghr-miR396g, ghr-miR4370, ghr-miR447a, ghr-miR5170, ghr-miR6158a, and ghr-miR6158b.

Validation and comparison of fiber-development-related miRNAs by qRT-PCR

According to the annotation of the identified miRNA targets, 46 miRNAs (36 unique miRNAs) that are closely related to fiber initiation and elongation as well as secondary cell wall synthesis, were chosen for validation and comparison of expression amongst -2, 0, and 2 DPA ovules. Our qRT-PCR result showed that all of the 36 unique miRNAs were successfully detected in the three developmental stages (Supplementary 5-7). Independent student t test revealed 1 and 11 miRNAs in 0 and 2 DPA ovules were expressed differentially with the counterparts of -2 DPA ovule, respectively (Figure 5-8). Most of these differentially expressed miRNAs were unregulated in 2 DPA than those in -2 and 0 DPA ovules, ranging from 2.5 to 8.3 folds, like ghr-miR159a, ghr-miR160a, ghr-miR164a/b/c/d/e, and ghr-miR167b (Figure 5-8). However, ghr-miR156b and ghr-miR169b were remarkably down-regulated in 0 and +2 DPA ovules, respectively. The differentially expressed miRNAs implicated their potential roles in promoting fiber development. It is likely for the rest miRNAs to be fundamental regulators in fiber development that always remain a relatively stable expression level from -2 to +2 DPA. These results allowed us a broader view on understanding the role of miRNAs and their targets in fiber development and other regulatory machineries.

Discussion

Newly identified miRNAs in cotton

Recently, a series of conserved and novel miRNAs of cotton had been also identified by other research groups using high-throughput small RNA sequencing methods (Kwak et al., 2009; Pang

et al., 2009; Wang et al., 2012b; Zhang et al., 2013). We compared our identified miRNAs with those from recent reports and from miRBase database. It turned out that 31 conserved cotton miRNA overlapped with previously reported miRNAs in cotton, belonging to 18 miRNA families (Supplementary 5-1). Thus, 89 conserved and 8 novel cotton miRNAs were newly identified in this study, including ghr-miR156, ghr-miR160, ghr-miR166, and ghr-n1. 76 out of the 97 (78.35%) newly identified miRNA were expressed specifically or higher in leaf, implicating their importance to leaf development. In addition, 91 of the 97 (93.81%) newly identified miRNA precursors were from cotton D genome, implicating current nucleotide resource for upland cotton is still limited for miRNA identification and more conserved and novel miRNAs would be found with the emerging of upland cotton genome.

miRNA role in the development of leaf and ovule

miR156/157 is a highly conserved and expressed miRNA family in the plant kingdom which has been demonstrated to take part in regulation of leaf development and flower development through targeting SQUAMOSA-promoter binding-like (*SPL*) family in *Arabidopsis* (Xing et al., 2010) and rice (Xie et al., 2012b). In this cotton study, we found that miR157a/b showed huge expression difference between leaf and ovule, up to 8.25 folds (leaf vs ovule). Additionally, miR157a/b is the most abundant miRNAs in cotton leaf. The expression of miR156/157 was consistent with two previous results (Kwak et al., 2009; Lu et al., 2013) in which both studies showed that miR156/157 is not the richest miRNAs in ovule. In our study, we detected miR166a/b/c/d/e/j as the most abundant conserved miRNA in the ovule, 1.4 folds higher than in the leaf. Target prediction analysis showed that three miR166a/b/c/d/e/j target 3 homeobox leucine

zipper proteins, 2 ribonucleotide reductase 2A and 2 rubisco methyltransferase proteins. Homeobox leucine zipper (HD-ZIP) protein is critical for meristem development in which HD-ZIP III family genes are well characterized as developmental regulators required for shoot apical meristem (SAM) establishment (Prigge et al., 2005). In *Arabidopsis*, miR165/166 directs the repression of HD-ZIP III genes, resulting in ectopic accumulation of the transcripts in abaxial domains of leaf primordial (Zhang and Zhang, 2012). Qiu et al., reported that a novel HD-Zip III gene named *GbHB1* may play a regulatory role in interfascicular fiber development in cotton, since *GbHB1* was highly expressed in ovule and stem, followed by in root, and low in leaf and cotyledon (Qiu et al., 2006). This suggests that miR166 might play a crucial role in promoting the fate of epidermal cell on ovary to fiber trichome by regulating HD-ZIP genes.

Transcription factors are proteins involved in promoting or repressing gene transcription through binding to regulatory sequences, allowing for specific expression of each regulated gene in different cell types during development (Latchman, 1997). Growing evidence reveals that transcription factors are one of most important miRNA targets, directly or indirectly participating in development, stress response, and metabolism. Amongst these categorized miRNA targets in cotton, transcription factors are also a substantial part in our identified miRNAs targets (Table 5-3), implying these factors may play important roles in cotton fiber and leaf development. In our miRNA target result, at least 9 types of transcription factors (*AP2*, *bHLH*, growth-regulating factor, *MYB*, *NAC*, *RAP2*, *SPL*, *TCP*, and *WD40*) were predicted to be miRNA targets that are associated with fiber development, involving 12 miRNA families including miR156, miR157, miR482, miR7484, miR7510, miR7584, miR828, miR858, miR8633, miR8718, miR8722, and miR8728

(Supplementary 5-3). Amongst these transcription factors, several MYBs were found to be preferentially expressed in fiber and were inferred to play crucial roles in fiber development, such as *GhMYB6*, *GhMYB109*, *GaMYB2*, *GhMYB25*, and *R2R3*-type *MYB* (Zhang et al., 2011). For instance, *GhMYB25* was identified to participate in regulating specialized outgrowths of epidermal cells to cotton fiber. Silence of *GhMYB25* is capable of inducing a series of abnormal fiber phenotypes, including altering the alternation of timing of rapid fiber elongation, fiber shortening, and reducing seed trichome deduction (Machado et al., 2009). In addition, *GhMYB109* was found to be expressed specifically at fiber initials (Suo et al., 2003) and elongating fibers and its silence resulted in reduced fiber length in cotton, implicating its potential role in fiber elongation (Pu et al., 2008; Zhang et al., 2011). Here, 11 MYB genes (*MYB 2/6/7/9/10/19/23/26/36/38/55/61/74/85*) were found to be targeted by miR7484, miR7584, miR828, miR858, and miR8722, indicating their roles in cotton fiber development.

Once initiation, fiber cells start to undergo rapid elongation during primary cell well synthesis. Ethylene functions in a set of hormone-mediated regulation including development of the main root, lateral roots, and root hairs, and biosynthesis of fatty acids (Qin et al., 2007). In cotton, activating ethylene biosynthesis could promote fiber growth and subsequently synthesized ethylene could further stimulate pectin biosynthesis and scaffold establishment (Pang et al., 2010; Qin et al., 2007). *RAP2*, also known as *APETALA2* (*AP2*)/*ERF* transcription factor, can bind as activators or repressors to the GCC box elements in the promoters of ethylene-responsive genes (Hinz et al., 2010). We detected five *RAP2*s (AY779338.1, contig1152, contig1153, contig1156, and contig228) were targeted by miR8633 and miR8718, respectively (Supplementary 5-2). Thus,

it indicated miR8633 and miR8718 might also take part in ethylene response and then regulate fiber development through targeting *RAP2* in cotton.

Many studies have shown that cotton fiber initially develops in a similar pattern as the leaf trichome of Arabidopsis and then undergoes elongation, secondary cell wall biosynthesis, and maturation (Pu et al., 2008; Wang et al., 2004). Therefore, we searched 110 functionally well-known genes from Arabidopsis and cotton that play important roles in trichome development and fiber development (Supplementary 5-9). Using Cytoscape, we built a miRNA-mediated interaction network during fiber development stages including initiation, elongation and secondary wall synthesis (Figure 5-9). Fifty eight miRNA families and their 49 targets were involved in fiber initiation, elongation and secondary wall synthesis. Taking fiber initiation as an example, a series of genes that were validated to be important for fiber trichome development were targeted by miRNAs, including *GL1, GL2, GL3, ETC2, CPC*, *GL1*, *TRY, MYB2, GL1, RDL1, MYB25,* and sucrose synthase *(Susy)* (Lee et al., 2007; Machado et al., 2009). Furthermore, other fiber-important genes were also detected to be miRNA targets such as *AP2*, tubulin, and *MAPK* (Machado et al., 2009). Together, the miRNA-mediated interaction network model may contribute to our understanding of miRNA in fiber development.

miRNAs and tasiRNAs in cotton

Recently, Guan and its colleagues reported *GhMYB2D* is targeted by miR828 and miR858 to generate phased tasiRNAs of *TAS4* family, resulting in biased using of *GhMYB2* that promotes fiber development in cotton (Guan et al., 2014). In our small sequencing data, miR858 were obtained, but we didn't find its precursor sequence in EST, GSS, or the D genome. However, we

obtained another miR828 which has 1-nt difference with Guan's miR828 (Supplementary 5-1). Both of our miR858 and miR828a were predicted to target *GhMYB2D and GhMYB2A*. At the same time, both of our miR858 and miR828a were only expressed in ovule at a low level. Moreover, we only detected phased-siRNA cluster on *GhMYB2A* and *GhMYB2D* (Supplementary 5-4) in the leaf sequencing library. During fiber initiation, *GhMYB2D* is targeted by miR858, miR828a, or both. *GhMYB2D* expression level is the opposite of the expression of miR858 or miR828a, in order to ensure fiber initiation and elongation. At the same time, *GhMYB2D* is only regulated by miR858 or miR828a, not generating tasiRNAs in ovule. *GhMYB2D* in leaf might be targeted by miR828 obtained by Guan et al., (Guan et al., 2014), and generate taisRNAs. The complex regulation of *GhMYB2D* leads to differentiation of leaf and ovule. One of *GhMYB2D/ GhMYB2A*-derived tasiRNA in our leaf sequencing data is perfectly complementary to the *GhMYB2D* and *GhMYB2A* (Supplementary 5-4). Does the tasiRNA really regulate *GhMYB2D* or *GhMYB2A* in a feedback regulation manner? If so, does the tasiRNA follow the way of miRNA or siRNA to target *GhMYB2D* or *GhMYB2A*? This speculation needs further experiments to validate.

Currently there are 8 tasiRNA genes identified in *Arabidopsis*, including *TAS1a*, *TAS1b*, *TAS1c*, *TAS2*, *TAS3a*, *TAS3b*, *TAS3c*, and *TAS4* (German et al., 2008). tasiRNA produced from the *TAS2* transcript was reported to target two clusters of pentatricopeptide repeat gene transcripts (*PPRs*), whereas tasiRNA from *TAS3* targets auxin response factor family members *ARF1*, 2, 3, or 4 (Allen and Howell, 2010). We found a tasiRNA-derived gene, *ARF6*. Our identified tasiRNAs were predicted to act on some genes important to fiber development and cotton seed development, such as fasciclin-like arabinogalactan protein 2, sucrose synthase, glycosyl hydrolase, and polyphenol

oxidase. It implies these tasiRNAs might play roles in fiber/leaf-related development.

EXPERIMENTAL PROCEDURES

Plant material and small RNA sequencing

Cotton (*Gossypium Hirsutum L.*) cultivar TM-1 were grown in the greenhouse. Cotton ovules -1 to +1 day post anthesis (DPA) and young leaves were harvested for five biological replicates and immediately frozen in liquid nitrogen. They were then stored at -80°C following RNA extraction. Totals RNAs were extracted from each tissue sample using the mirVana miRNA isolation kit (Ambion, Austin, TX) according to the manufacturer's protocol. The small RNA samples extracted from the five biological replicates were pooled together for leaf and ovule, respectively. Finally, the construction of pooled small RNA libraries and sequenced were performed by LC Sciences (Houston, TX) using Illumina high-throughput sequencing platform.

Deep sequencing analysis

The raw sequences generated from libraries of ovule and leaf were cleaned first, including trimming 5' and 3' adaptors and filtering low-quality reads. The raw sequences were then categorized to unique reads and read counts were also recorded. Clean reads were aligned against Sanger RNA family database (Rfam 10.1, ftp://ftp.sanger.ac.uk/pub/databases/Rfam) (Gardner et al., 2011) to detect other RNA types by BLASTn with an E-value of 100, which include repeat RNA, rRNA, snRNA, snoRNA, and tRNA. Only the fully matched were discarded. The remaining sequences were further aligned against miRBase (Release 20: June 2013) (Kozomara and Griffiths-Jones, 2011) to discriminate conserved reads and non-conserved reads by WATER (Xie et al.,

2010). If a read with no more than 3 mismatches including mismatch and gap is viewed as a conserved read with known mature miRNAs, otherwise it was considered a non-conserved read. The newly released cotton D genome, *G. raimondii*, sequences (Paterson et al., 2012; Wang et al., 2012a) were also used for identifying cotton miRNAs. Additionally, we assembled EST and GSS databases of upland cotton by CAP3-based TGICL with a set of default parameters (ftp://ftp.tigr.org/pub/software/tgi/tgicl/), respectively. The assembled databases of EST and GSS and the D genome of *raimondii* were used as data source for identifying miRNA precursors. miRDeepFinder was used to identify miRNAs and their targets with the default parameters in the software (Xie et al., 2012a). In brief, miRNA precursor candidate sequences were excised from 700-bp downstream and upstream sequences of loci where short reads were able to be 100% mapped back to non-coding sequences. Then, RNAfold was used to fold these candidate sequences into secondary RNA structures. If candidate sequences that could stratify the following two criteria are considered to be a miRNA precursor: candidate sequences could be into a perfect or near-perfect stem-loop hairpin structure, where mature miRNAs stand one arm of the hairpin structure and have less than 7 mimatches with opposite arm (Pu et al., 2008). In addition, considering many of conserved miRNAs in model plant species were not found to have miRNA star in a small RNA sequencing, only novel miRNAs were required to have miRNA stars at least in one cotton sequencing library (Xie et al., 2012a).

Particularly, miRNA candidates identified from the D genome were aligned against those identified from EST and GSS to remove repeated precursors. Using psRNA tool with a set of default parameters (Dai and Zhao, 2011), miRNA targets were predicted from cotton annotated

mRNA database in NCBI and the protein-coding assembled EST contigs from our cotton EST database (Xie et al., 2011).

miRNA target validation based on degradome sequencing

Degradome sequencing analysis has been become a powerful approach for identifying and validating miRNA targets in plants through identifying the cleavage sites. Here, we also employed this approach to validate the miRNA targets. Three small RNA degradome sequencing dataset (GSM1008997: seedlings; GSM100899: hypocotyl; and GSM1061853: anthers) of upland cotton were downloaded from NCBI for validating the predicted miRNA targets using CleveLand4 with default parameters (Xing et al., 2010). Considering degradome sequencing is for detecting the sliced site on miRNA targets that generally occurs on the 10th and 11th nucleotides of mature miRNAs. Thus, no mismatches was allowed on the two nucleotides in degradome analysis (Zhang et al., 2013). Finally, miRNA targets in a p-value of ≤ 0.05 were kept.

miRNA expression profiles and comparison between ovule and leaf

All miRNA abundance was normalized to reads per million (RPM). If the original miRNA expression in a library was zero, the normalized expression was adjusted to 0.01 according to a previous report (Murakami et al., 2006). miRNA expression fold change between ovule and leaf was computed with the formula, Fold change=$\log_2$(ovule / leaf) (Marsit et al., 2006). Pearson's chi-squared test was performed for the significance of miRNA expression from the two samples. Fold change and p-value were combined to determine the final miRNA expression significance. We defined expression difference level using the following rules: extremely significant (**) if (fold change ≥ 1 or fold change ≤ -1) and p-value ≤ 0.01; significant (*) if (fold change ≥ 1 or fold

change ≤ -1) and 0.05 ≥ p-value > 0.01; otherwise insignificant. A miRNA family expression abundance was calculated as the sum of expression of conserved reads that have no more than 3 mismatches with known plant miRNAs in a same family. Pearson's chi-squared test and fold change mentioned above were also used to distinguish difference of miRNA family expression between leaf and ovule.

Identification of tasiRNAs and their targets

tasiRNAs are initially generated from non-coding and coding transcripts from the product of miRNA-guided cleavage, and then form into double-strand RNA through RNA-dependent RNA polymerase (*RDR6*) (Vazquez et al., 2004). tasiRNAs negatively regulate target mRNAs in the same manner with miRNAs (Guan et al., 2014; Yoshikawa et al., 2005). To identify potential tasiRNAs, we referenced the approaches from Quintero et al., (Quintero et al., 2013) and Fei et al., (Fei et al., 2013) and made minor modifications. Briefly, 1) all sequences reads with an original abundance of ≥ 10 except miRNAs and other types of known RNAs were 100% mapped back to the EST database and coding genes of upland cotton; 2) mapped reads were in length of 18-24 nt; 3) ≥ 3 different mapped positions were found for small RNA reads and were spanned by 10-25 nt; 4) miRNA target loci could be predicted on non-phased-siRNA region of tasiRNA candidate genes; 5) all tasiRNA candidate genes were aligned against itself to remove the repeated sequences with cutoff E-value of 1e-20. The siRNA in highest abundance in the phased-siRNA cluster were chosen to do tasiRNA target prediction using Target-align (Li et al., 2012).

miRNA function annotation

Identified miRNA targets were performed alignment against GO protein database for GO

term classification and KEGG pathway enrichment according to our previously reported approach (Xie et al., 2010). To visualize the biological process and molecular function that identified miRNAs and their targets are involved in, corresponding *Arabidopsis* genes homologous to cotton miRNA targets were used an input for Gorilla (process ontology) (Eden et al., 2009).

Validation of miRNA role in fiber development by qRT-PCR

To validate the newly identified miRNAs and their role in fiber development, 30 miRNAs associated with fiber development were chosen to design miRNA-specific stem loop primers for miRNA expression analysis by qRT-PCR (Xie et al., 2012a). Total RNAs of -2, 0, and +2 DPA ovules were reversely transcribed to cDNA, according to the manufacture protocol of TaqMan® MicroRNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA). qRT-PCRs were performed with Applied Biosystems ViiA 7 Real Time PCR System. Three biological replicates were run for each reaction with three technical replicates. ΔCt-based fold change was calculated as each miRNA expression abundance, while ΔΔCt-based fold change was used for pair-wise comparison of miRNA expression amongst -2, 0, and +2 DPA ovules (Xie et al., 2012b). Independent student t test was performed to test whether a miRNA expression is significantly different from each other under a p-value of 0.05.

Supplementary Information

Supplementary 5-1. Summary of miRNA family comparison among control, salt, and drought libraries in cotton.

Supplementary 5-2. miRNA targets for conserved cotton miRNAs

Supplementary 5-3. GO ontology classification of identified miRNA families in cotton.

Supplementary 5-4. Gene pathway analysis for cotton miRNA targets based on GO and KEGG analysis

Supplementary 5-5. TasiRNAs and their targets

Supplementary 5-6. Chromosome mapping of cotton miRNAs and their targets.

Supplementary 5-7. Fiber-development-related miRNA expression data of -2, 0, and +2 DPA ovules by qRT-PCR.

Supplementary 5-8. Genes involved in cotton fiber initiation and development and potential miRNA-mediated gene network.

Reference

Adams, K.L., Cronn, R., Percifield, R. and Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National*

Academy of Sciences of the United States of America **100**, 4649-4654.

Allen, E. and Howell, M.D. (2010) miRNAs in the biogenesis of trans-acting siRNAs in higher plants. *Semin Cell Dev Biol* **21**, 798-804.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297.

Chen, Z.J., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X., Stelly, D.M., Rabinowicz, P.D., Town, C.D., Arioli, T., Brubaker, C., Cantrell, R.G., Lacape, J.M., Ulloa, M., Chee, P., Gingle, A.R., Haigler, C.H., Percy, R., Saha, S., Wilkins, T., Wright, R.J., Van Deynze, A., Zhu, Y., Yu, S., Abdurakhmonov, I., Katageri, I., Kumar, P.A., Mehboob Ur, R., Zafar, Y., Yu, J.Z., Kohel, R.J., Wendel, J.F. and Paterson, A.H. (2007) Toward sequencing cotton (Gossypium) genomes. *Plant physiology* **145**, 1303-1310.

Dai, X. and Zhao, P.X. (2011) psRNATarget: a plant small RNA target analysis server. *Nucleic acids research* **39**, W155-159.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 48.

Eldem, V., Okay, S. and Unver, T. (2013) Plant microRNAs: new players in functional genomics. *Turkish Journal of Agriculture and Forestry* **37**, 1-21.

Fei, Q., Xia, R. and Meyers, B.C. (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant cell* **25**, 2400-2415.

Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R.D., Nawrocki, E.P., Kolbe, D.L., Eddy, S.R. and Bateman, A. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic acids research* **39**, D141-145.

German, M.A., Pillay, M., Jeong, D.H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L.A., Nobuta,

K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B.C. and Green, P.J. (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol* **26,** 941-946.

Gong, L., Kakrana, A., Arikit, S., Meyers, B.C. and Wendel, J.F. (2013) Composition and Expression of Conserved MicroRNA Genes in Diploid Cotton (Gossypium) Species. *Genome biology and evolution* **5,** 2449-2459.

Guan, X., Pang, M., Nah, G., Shi, X., Ye, W., Stelly, D.M. and Chen, Z.J. (2014) miR828 and miR858 regulate homoeologous MYB2 gene functions in Arabidopsis trichome and cotton fibre development. *Nature communications* **5,** 3050.

Hinz, M., Wilson, I.W., Yang, J., Buerstenbinder, K., Llewellyn, D., Dennis, E.S., Sauter, M. and Dolferus, R. (2010) Arabidopsis RAP2.2: an ethylene response transcription factor that is important for hypoxia survival. *Plant physiology* **153,** 757-772.

Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* **39**, D152-157.

Kwak, P.B., Wang, Q.Q., Chen, X.S., Qiu, C.X. and Yang, Z.M. (2009) Enrichment of a set of microRNAs during the cotton fiber development. *BMC genomics* **10,** 457.

Larkin, J.C., Brown, M.L. and Schiefelbein, J. (2003) How do cells know what they want to be when they grow up? Lessons from epidermal patterning in Arabidopsis. *Annual review of plant biology* **54,** 403-430.

Latchman, D.S. (1997) Transcription factors: an overview. *Int J Biochem Cell Biol* **29,** 1305-1312.

Lee, J.J., Woodward, A.W. and Chen, Z.J. (2007) Gene expression changes and early events in cotton fibre development. *Annals of botany* **100,** 1391-1401.

Lelandais-Briere, C., Naya, L., Sallet, E., Calenge, F., Frugier, F., Hartmann, C., Gouzy, J. and Crespi, M. (2009) Genome-wide Medicago truncatula small RNA analysis revealed novel microRNAs and isoforms

differentially regulated in roots and nodules. *The Plant cell* **21,** 2780-2796.

Li, F., Orban, R. and Baker, B. (2012) SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. *The Plant journal : for cell and molecular biology* **70,** 891-901.

Lu, S., Li, Q., Wei, H., Chang, M.J., Tunlaya-Anukit, S., Kim, H., Liu, J., Song, J., Sun, Y.H., Yuan, L., Yeh, T.F., Peszlen, I., Ralph, J., Sederoff, R.R. and Chiang, V.L. (2013) Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in Populus trichocarpa. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10848-10853.

Machado, A., Wu, Y., Yang, Y., Llewellyn, D.J. and Dennis, E.S. (2009) The MYB transcription factor GhMYB25 regulates early fibre and trichome development. *The Plant journal : for cell and molecular biology* **59,** 52-62.

Marsit, C.J., Eddy, K. and Kelsey, K.T. (2006) MicroRNA responses to cellular stress. *Cancer research* **66,** 10843-10848.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J., Griffiths-Jones, S., Jacobsen, S.E., Mallory, A.C., Martienssen, R.A., Poethig, R.S., Qi, Y., Vaucheret, H., Voinnet, O., Watanabe, Y., Weigel, D. and Zhu, J.K. (2008) Criteria for annotation of plant MicroRNAs. *The Plant cell* **20,** 3186-3190.

Mi, H., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8,** 1551-1566.

Murakami, Y., Yasuda, T., Saigo, K., Urashima, T., Toyoda, H., Okanoue, T. and Shimotohno, K. (2006) Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. *Oncogene* **25**, 2537-2545.

Nag, A., King, S. and Jack, T. (2009) miR319a targeting of TCP4 is critical for petal growth and development in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 22534-22539.

Palatnik, J.F., Wollmann, H., Schommer, C., Schwab, R., Boisbouvier, J., Rodriguez, R., Warthmann, N., Allen, E., Dezulian, T., Huson, D., Carrington, J.C. and Weigel, D. (2007) Sequence and expression differences underlie functional specialization of Arabidopsis microRNAs miR159 and miR319. *Developmental cell* **13**, 115-125.

Pang, C.Y., Wang, H., Pang, Y., Xu, C., Jiao, Y., Qin, Y.M., Western, T.L., Yu, S.X. and Zhu, Y.X. (2010) Comparative proteomics indicates that biosynthesis of pectic precursors is important for cotton fiber and Arabidopsis root hair elongation. *Molecular & cellular proteomics : MCP* **9**, 2019-2033.

Pang, M., Woodward, A.W., Agarwal, V., Guan, X., Ha, M., Ramachandran, V., Chen, X., Triplett, B.A., Stelly, D.M. and Chen, Z.J. (2009) Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). *Genome biology* **10**, R122.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., Yoo, M.J., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M.V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T.H., Li, J., Lin, L., Liu, T., Marler, B.S., Page, J.T., Roberts, A.W., Romanel, E., Sanders, W.S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z., Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J.E., Brubaker, C.L., Chee, P.W., Das, S., Gingle, A.R., Haigler, C.H., Harker, D., Hoffmann, L.V., Hovav, R., Jones, D.C., Lemke, C., Mansoor, S., ur Rahman, M., Rainville, L.N., Rambani, A., Reddy, U.K., Rong, J.K., Saranga, Y., Scheffler, B.E., Scheffler, J.A., Stelly, D.M., Triplett, B.A., Van Deynze, A., Vaslin, M.F., Waghmare, V.N., Walford, S.A., Wright, R.J., Zaki, E.A., Zhang, T.,

Dennis, E.S., Mayer, K.F., Peterson, D.G., Rokhsar, D.S., Wang, X. and Schmutz, J. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* **492,** 423-427.

Pesch, M. and Hulskamp, M. (2009) One, two, three...models for trichome patterning in Arabidopsis? *Current opinion in plant biology* **12,** 587-592.

Prigge, M.J., Otsuga, D., Alonso, J.M., Ecker, J.R., Drews, G.N. and Clark, S.E. (2005) Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in Arabidopsis development. *The Plant cell* **17,** 61-76.

Pu, L., Li, Q., Fan, X., Yang, W. and Xue, Y. (2008) The R2R3 MYB transcription factor GhMYB109 is required for cotton fiber development. *Genetics* **180**, 811-820.

Qin, Y.M., Hu, C.Y., Pang, Y., Kastaniotis, A.J., Hiltunen, J.K. and Zhu, Y.X. (2007) Saturated very-long-chain fatty acids promote cotton fiber and Arabidopsis cell elongation by activating ethylene biosynthesis. *The Plant cell* **19**, 3692-3704.

Qin, Y.M. and Zhu, Y.X. (2011) How cotton fibers elongate: a tale of linear cell-growth mode. *Current opinion in plant biology* **14,** 106-111.

Qiu, C., Zuo, K., Qin, J., Zhao, J., Ling, H. and Tang, K. (2006) Isolation and characterization of a class III homeodomain-leucine zipper-like gene from Gossypium barbadense. *DNA sequence : the journal of DNA sequencing and mapping* **17,** 334-341.

Qiu, C.X., Xie, F.L., Zhu, Y.Y., Guo, K., Huang, S.Q., Nie, L. and Yang, Z.M. (2007) Computational identification of microRNAs and their targets in Gossypium hirsutum expressed sequence tags. *Gene* **395,** 49-61.

Quintero, A., Perez-Quintero, A.L. and Lopez, C. (2013) Identification of ta-siRNAs and cis-nat-siRNAs in cassava and their roles in response to cassava bacterial blight. *Genomics, proteomics & bioinformatics* **11,** 172-181.

Sunkar, R., Zhou, X., Zheng, Y., Zhang, W. and Zhu, J.K. (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC plant biology* **8**, 25.

Suo, J., Liang, X., Pu, L., Zhang, Y. and Xue, Y. (2003) Identification of GhMYB109 encoding a R2R3 MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (Gossypium hirsutum L.). *Biochimica et biophysica acta* **1630**, 25-34.

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A.C., Hilbert, J.L., Bartel, D.P. and Crete, P. (2004) Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Molecular cell* **16**, 69-79.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., Liu, K., Wang, B., Song, C., Shi, N., Kohel, R.J., Percy, R.G., Yu, J.Z., Zhu, Y.X. and Yu, S. (2012a) The draft genome of a diploid cotton Gossypium raimondii. *Nature genetics* **44**, 1098-1103.

Wang, S., Wang, J.W., Yu, N., Li, C.H., Luo, B., Gou, J.Y., Wang, L.J. and Chen, X.Y. (2004) Control of plant trichome development by a cotton fiber MYB gene. *The Plant cell* **16**, 2323-2334.

Wang, Z.M., Xue, W., Dong, C.J., Jin, L.G., Bian, S.M., Wang, C., Wu, X.Y. and Liu, J.Y. (2012b) A comparative miRNAome analysis reveals seven fiber initiation-related and 36 novel miRNAs in developing cotton ovules. *Molecular plant* **5**, 889-900.

Wendel, J.F. (1989) New World tetraploid cottons contain Old World cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 4132-4136.

Williams, L., Carles, C.C., Osmont, K.S. and Fletcher, J.C. (2005) A database analysis method identifies an endogenous trans-acting short-interfering RNA that targets the Arabidopsis ARF2, ARF3, and ARF4 genes.

Proceedings of the National Academy of Sciences of the United States of America **102,** 9703-9708.

Xie, F., Frazier, T.P. and Zhang, B. (2010) Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (Panicum virgatum). *Planta* **232,** 417-434.

Xie, F., Sun, G., Stiller, J.W. and Zhang, B. (2011) Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database. *PloS one* **6**, e26980.

Xie, F., Xiao, P., Chen, D., Xu, L. and Zhang, B. (2012a) miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant molecular biology*.

Xie, K., Shen, J., Hou, X., Yao, J., Li, X., Xiao, J. and Xiong, L. (2012b) Gradual increase of miR156 regulates temporal expression changes of numerous genes during leaf development in rice. *Plant physiology* **158,** 1382-1394.

Xing, S., Salinas, M., Hohmann, S., Berndtgen, R. and Huijser, P. (2010) miR156-targeted and nontargeted SBP-box transcription factors act in concert to secure male fertility in Arabidopsis. *The Plant cell* **22,** 3935-3950.

Xue, W., Wang, Z., Du, M., Liu, Y. and Liu, J.Y. (2013) Genome-wide analysis of small RNAs reveals eight fiber elongation-related and 257 novel microRNAs in elongating cotton fiber cells. *BMC genomics* **14,** 629.

Yoshikawa, M., Peragine, A., Park, M.Y. and Poethig, R.S. (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. *Genes & development* **19,** 2164-2175.

Zhang, B. and Wang, Q. (2014) MicroRNA-based Biotechnology for Plant Improvement. *Journal of Cellular Physiology*, n/a-n/a.

Zhang, C., Li, G., Zhu, S., Zhang, S. and Fang, J. (2014) tasiRNAdb: a database of ta-siRNA regulatory pathways. *Bioinformatics*.

Zhang, F., Liu, X., Zuo, K., Sun, X. and Tang, K. (2011) Molecular cloning and expression analysis of a novel

SANT/MYB gene from Gossypium barbadense. *Molecular biology reports* **38**, 2329-2336.

Zhang, H., Wan, Q., Ye, W., Lv, Y., Wu, H. and Zhang, T. (2013) Genome-wide analysis of small RNA and novel microRNA discovery during fiber and seed initial development in Gossypium hirsutum. L. *PloS one* **8**, e69743.

Zhang, Z. and Zhang, X. (2012) Argonautes compete for miR165/166 to regulate shoot apical meristem development. *Current opinion in plant biology* **15**, 652-658.

Table 5-1. Small RNA categorization in cotton *

	Redundant (L)	Redundant (O)	Unique (L)	Unique (O)
Matched (E/G)	360458 (74.34%)	321638 (96.36%)	18354 (85.10%)	39919 (96.54%)
Matched (D)	288339 (59.46%)	137200 (41.11%)	5469 (25.36%)	14436 (34.91%)
miRNA	204376 (42.15%)	89529 (26.82%)	1972 (9.14%)	2376 (5.75%)
Other RNAs	82953 (17.11%)	3206 (0.96%)	761 (3.53%)	286 (0.69%)
Total (≥ 3 copies)	484896	333771	21568	41350

*: the number represented the raw data generated directly from deep sequencing; L: leaf; O: Ovule; Matched (E/G): Matched to EST and GSS of upland cotton; Matched (D): Matched to cotton D genome *G. raimondii* sequence.

Table 5-2. The expression of conserved miRNA families in leaf (L) and ovule (C)

Family	Leaf	Ovule	Fold change (O vs L)	p-value	Significance
miR156	1336.37	10.19	-7.04	0.000	**
miR157	40524.15	158.79	-8.00	0.000	**
miR159	137.99	75.48	-0.87	0.000	
miR160	24.02	15.10	-0.67	0.150	
miR162	7.82	3.46	-1.18	0.206	
miR164	13.97	0.01	-10.45	0.000	**
miR165	10.89	130.23	3.58	0.000	**
miR166	2332.23	8538.67	1.87	0.000	**
miR167	509.52	2859.14	2.49	0.000	**
miR168	118.72	22.92	-2.37	0.000	**
miR169	243.59	19.10	-3.67	0.000	**
miR171	70.67	16.55	-2.09	0.000	**
miR172	306.44	2790.38	3.19	0.000	**
miR390	705.34	336.50	-1.07	0.000	**
miR393	5.03	14.73	1.55	0.039	*
miR394	0.01	5.46	9.09	0.025	*
miR395	5.31	3.09	-0.78	0.480	
miR396	553.66	13.64	-5.34	0.000	**
miR397	1189.72	1.27	-9.87	0.000	**
miR398	8.10	0.01	-9.66	0.005	**
miR403	5.03	0.55	-3.20	0.025	*

miR408	39.11	1.09	-5.16	0.000	**
miR447	1.12	0.01	-6.80	0.317	
miR477	7.26	3.27	-1.15	0.206	
miR479	0.01	5.27	9.04	0.025	*
miR482	15.92	2.73	-2.54	0.002	**
miR530	8.10	0.01	-9.66	0.005	**
miR535	90.23	0.01	-13.14	0.000	**
miR783	0.01	0.55	5.77	1.000	
miR828	0.01	1.09	6.77	0.317	
miR829	0.84	0.01	-6.39	1.000	
miR858	1.40	8.55	2.61	0.020	*
miR894	0.84	3.27	1.97	0.083	
miR1120	1.12	0.01	-6.80	0.317	
miR1122	0.01	0.73	6.18	1.000	
miR1169	0.01	0.73	6.18	1.000	
miR1424	0.01	0.73	6.18	1.000	
miR1853	0.84	0.01	-6.39	1.000	
miR1873	1.12	0.01	-6.80	0.317	
miR2085	0.84	0.01	-6.39	1.000	
miR2118	2.51	0.01	-7.97	0.157	
miR2629	0.01	2.00	7.64	0.157	
miR2911	30.73	12.00	-1.36	0.005	**
miR2947	19.00	2.00	-3.25	0.000	**

miR2948	19.83	0.73	-4.77	0.000	**
miR2949	31.01	469.28	3.92	0.000	**
miR2950	10.34	0.01	-10.01	0.002	**
miR3443	1.40	0.01	-7.13	0.317	
miR3476	18.16	6.91	-1.39	0.014	*
miR3627	2.51	0.01	-7.97	0.157	
miR3699	0.01	0.91	6.51	1.000	
miR3954	8426.35	709.01	-3.57	0.000	**
miR4370	0.01	3.09	8.27	0.083	
miR4371	2.23	0.01	-7.80	0.157	
miR4992	0.01	0.55	5.77	1.000	
miR5059	0.84	0.01	-6.39	1.000	
miR5170	0.84	0.01	-6.39	1.000	
miR5224	0.84	0.01	-6.39	1.000	
miR5260	0.01	0.55	5.77	1.000	
miR5274	2.23	0.01	-7.80	0.157	
miR5649	0.01	0.55	5.77	1.000	
miR5782	0.01	0.55	5.77	1.000	
miR5813	0.84	0.01	-6.39	1.000	
miR6158	0.01	0.55	5.77	1.000	
miR7122	243.87	32.56	-2.90	0.000	**

*: All miRNA families with redundant abundance were normalized to transcript expression levels per million reads (RPM). If the original miRNA expression in a library was zero, the normalized

expression was adjusted to 0.01 according to a previous report (Murakami et al., 2006). miRNA expression fold change in the two libraries was calculated with the formula, Fold change=$\log_2$(ovule / leaf) (Marsit et al., 2006). A 2 × 2 contingency table was used to perform Pearson's chi-squared test for significance of miRNA expression from two samples. Fold change and *p*-value were combined to determine the final miRNA expression significance. We defined expression difference level as following rules: extremely significant (**) if (fold change ≥ 1 or fold change ≤ -1) and *p-value* ≤ 0.01; significant (*) if (fold change ≥ 1 or fold change ≤ -1) and 0.05 ≥ *p*-value > 0.01; otherwise insignificant.

Table 5-3. Fiber-development-related miRNAs, miRNA targets, GO terms and KEGG pathways in cotton

	miRNAs	Targets	Biological process	Molecular function	Cellular component	Pathway
Development	26	47	66	29	30	6
Fiber development	36	116	135	74	42	12
Metabolism	60	147	194	132	61	48
Signal transduction	18	26	38	21	18	11
Stress response	47	94	127	85	39	31
Transcription factor	38	68	89	19	11	0

Table 5-4. Identified tasiRNAs in cotton

miRNA	tasiRNA-derived gene	Gene funciton	Phased siRNAs	Source
ghr-miR390a/b/c/d	AF034130	MYB-like protein (MYB2)	3	Leaf
ghr-miR447a	AI054467	Beta-galactosidase	23	Leaf
ghr-miR156a/c/d	DT458065	Glutamate receptor 3.6	7	Leaf
ghr-miR390a/b/c/d	DW502659	-	6	Leaf/ovule
ghr-miR156b, ghr-miR447a	ES791839	-	64	Leaf
ghr-miR167b/d	ES819493	Auxin response factor 6	14	Ovule
ghr-n3	ES833630	HVA22-like protein j	4	Ovule
ghr-miR167b/d	ES834779	Auxin response factor 6	8	Ovule
ghr-miR2911, ghr-miR4370	ES834802	ATP synthase subunit beta	22	Leaf/ovule
ghr-miR828a	JQ868558	D1 MYB2	3	Leaf
ghr-miR828a	JQ868555	A2 MYB2	3	Leaf

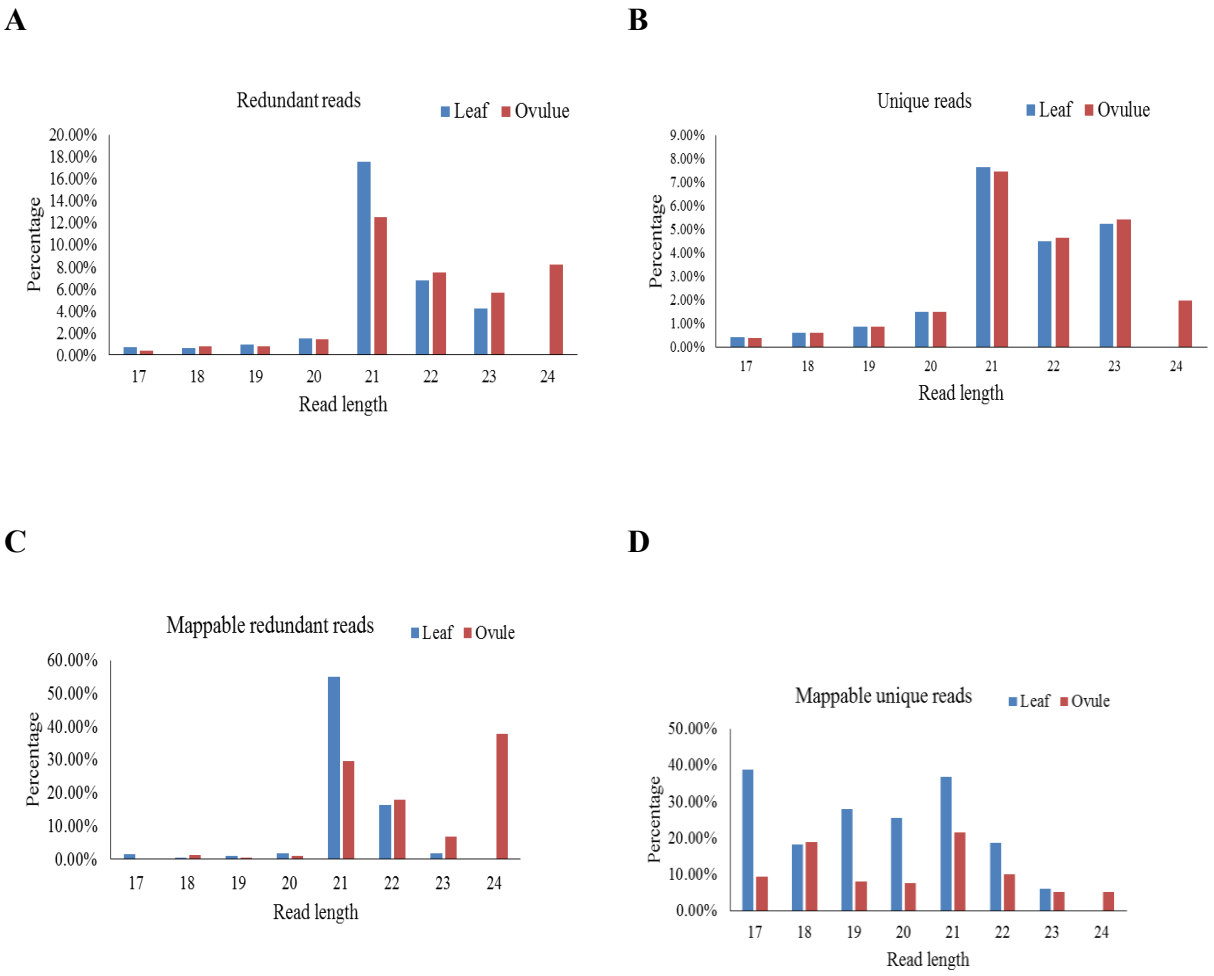


Figure 5-1. Size distribution of redundant and unique small RNA reads in cotton. **A and C:** Size distribution of redundant sRNA reads from leaf and ovule. **B and D:** Size distribution of unique sRNA reads from leaf and ovule. C and D: small RNA reads were fully mapped back to EST and GSS of upland cotton.

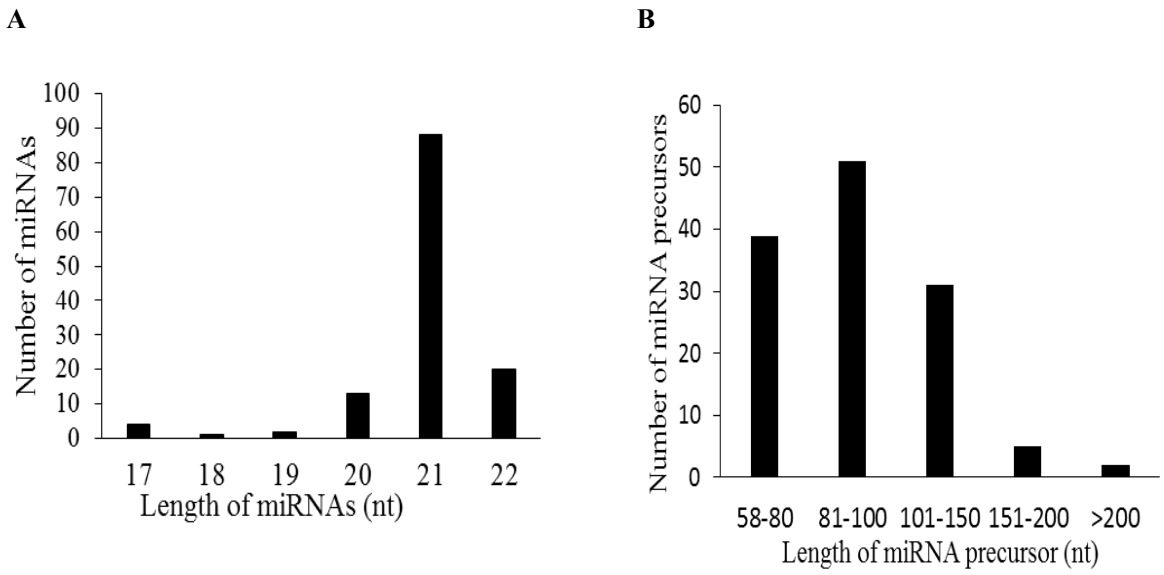


Figure 5-2. Size distribution of identified cotton mature miRNAs (**A**) and their precursors (**B**) from leaf and ovule

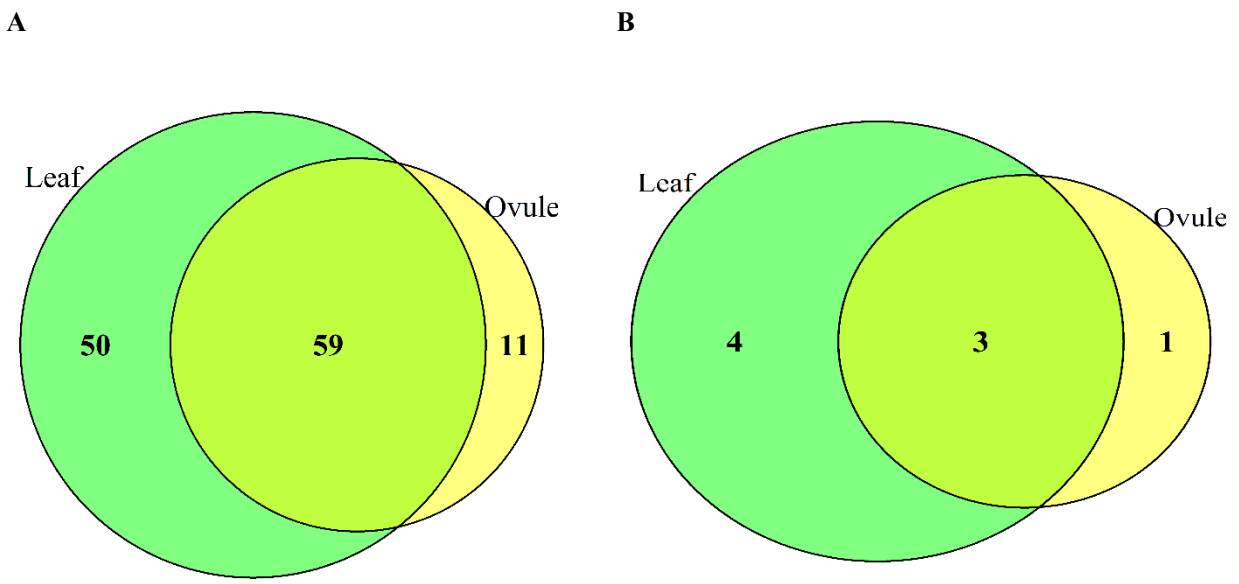


Figure 5-3. Distribution of miRNAs in leaf and ovule in cotton (A: conserved miRNAs; B: novel miRNAs).

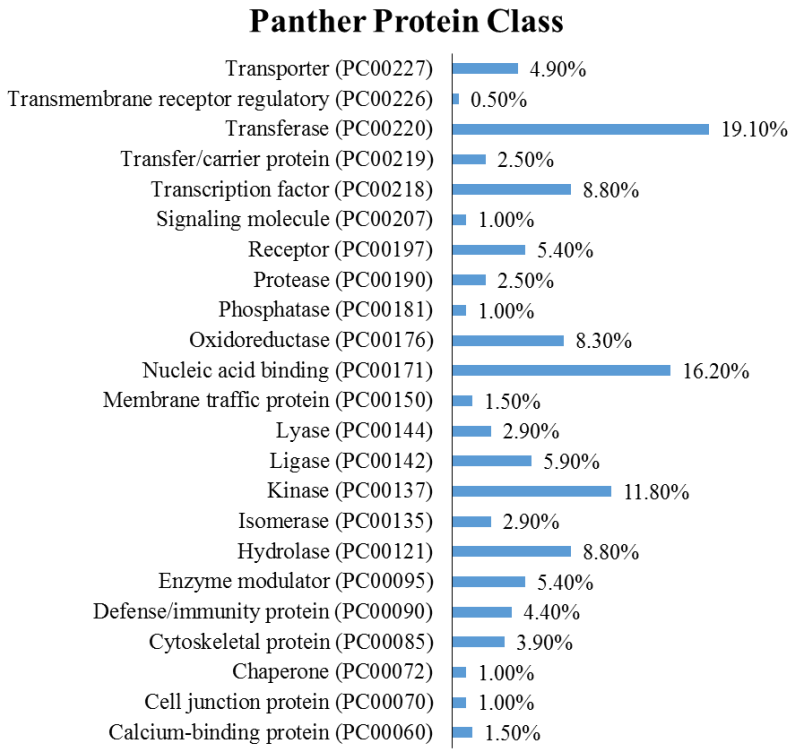


Figure 5-4. Panther-based protein classification of miRNA targets in cotton.

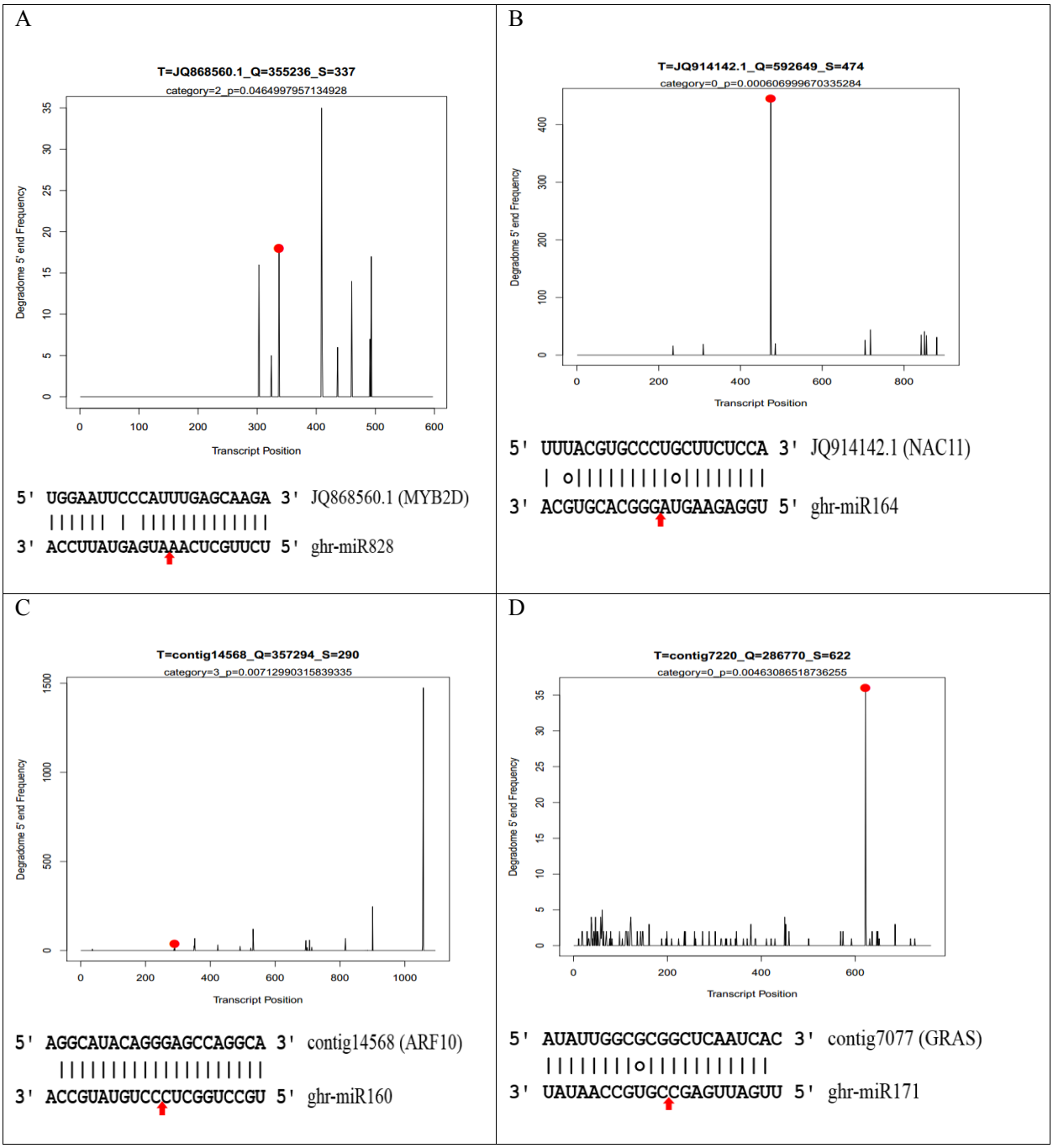


Figure 5-5. Target plots (t-plots) of cotton miRNA targets confirmed by degradome sequencing. A: ghr-miR828 and JQ868560.1 (MYB2D); B: ghr-miR164 and JQ914142.1 (NAC11); C: ghr-miR160 and contig14568 (ARF10); and D: ghr-miR171 and contig7077 (GRAS transcription factor). Red arrow denotes slice site on miRNA target.

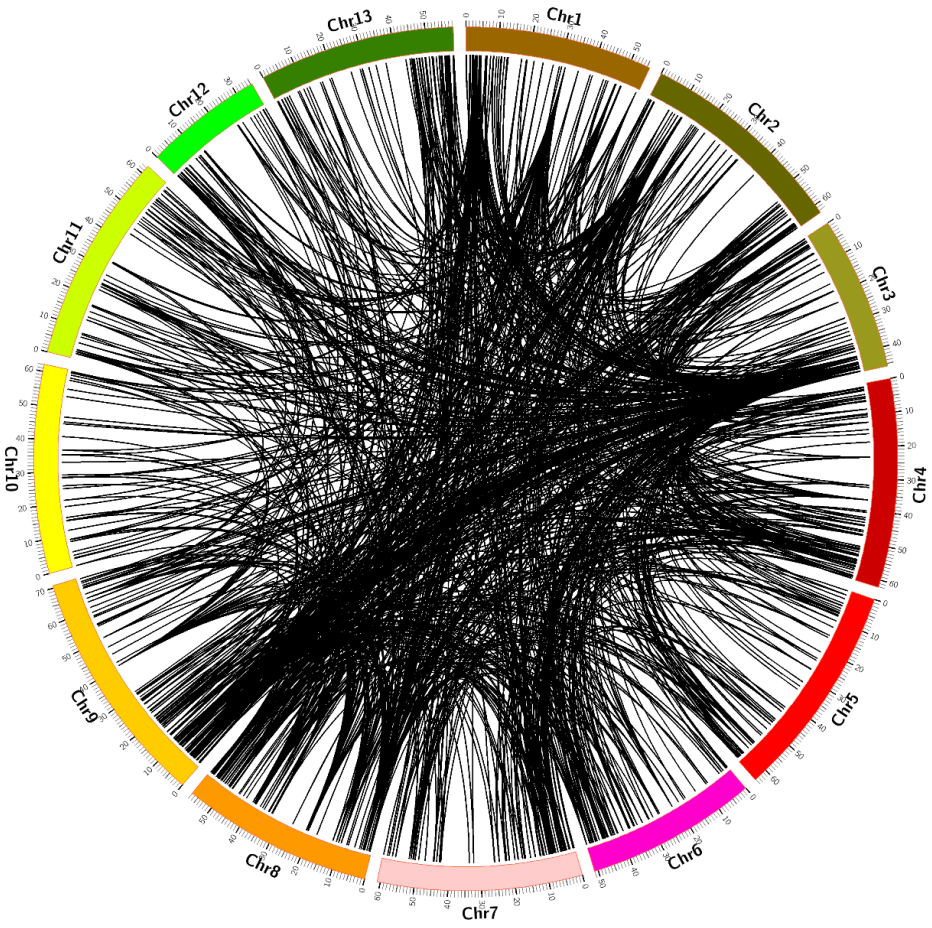
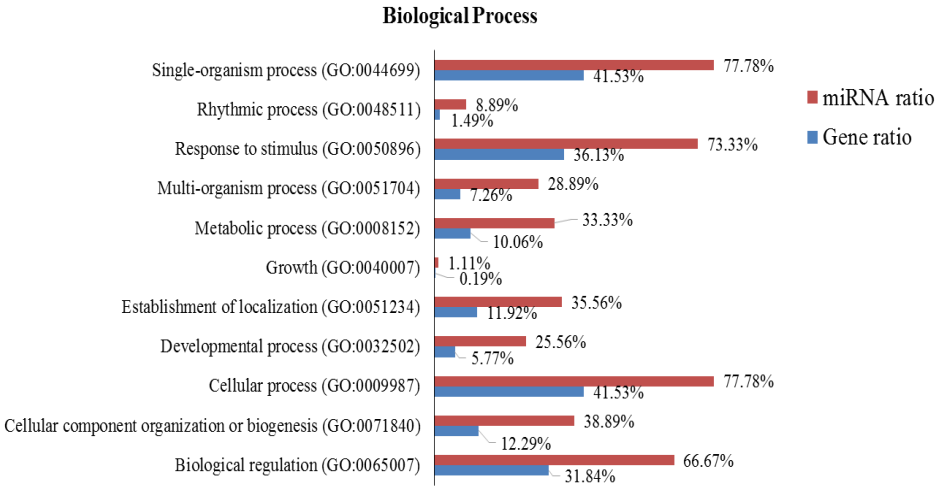
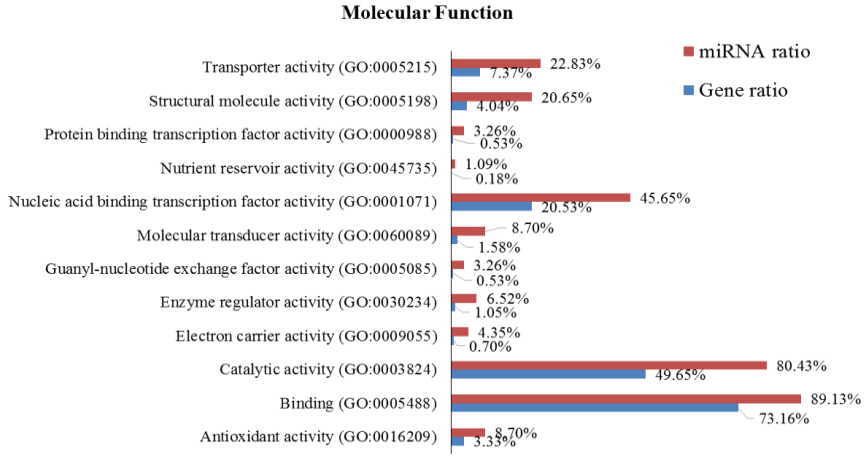


Figure 5-6. Distribution of cotton miRNAs and their targets in cotton D genome

A


Biological Process
Single-organism process (GO:0044699)
77.78%
41.53%
Rhythmic process (GO:0048511)
8.89%
1.49%
Response to stimulus (GO:0050896)
73.33%
36.13%
Multi-organism process (GO:0051704)
28.89%
7.26%
Metabolic process (GO:0008152)
33.33%
10.06%
Growth (GO:0040007)
1.11%
0.19%
Establishment of localization (GO:0051234)
35.56%
11.92%
Developmental process (GO:0032502)
25.56%
5.77%
Cellular process (GO:0009987)
77.78%
41.53%
Cellular component organization or biogenesis (GO:0071840)
38.89%
12.29%
Biological regulation (GO:0065007)
66.67%
31.84%
miRNA ratio
Gene ratio

B


Molecular Function
Transporter activity (GO:0005215)
22.83%
7.37%
Structural molecule activity (GO:0005198)
20.65%
4.04%
Protein binding transcription factor activity (GO:0000988)
3.26%
0.53%
Nutrient reservoir activity (GO:0045735)
1.09%
0.18%
Nucleic acid binding transcription factor activity (GO:0001071)
45.65%
20.53%
Molecular transducer activity (GO:0060089)
8.70%
1.58%
Guanyl-nucleotide exchange factor activity (GO:0005085)
3.26%
0.53%
Enzyme regulator activity (GO:0030234)
6.52%
1.05%
Electron carrier activity (GO:0009055)
4.35%
0.70%
Catalytic activity (GO:0003824)
80.43%
49.65%
Binding (GO:0005488)
89.13%
73.16%
Antioxidant activity (GO:0016209)
8.70%
3.33%
miRNA ratio
Gene ratio

C

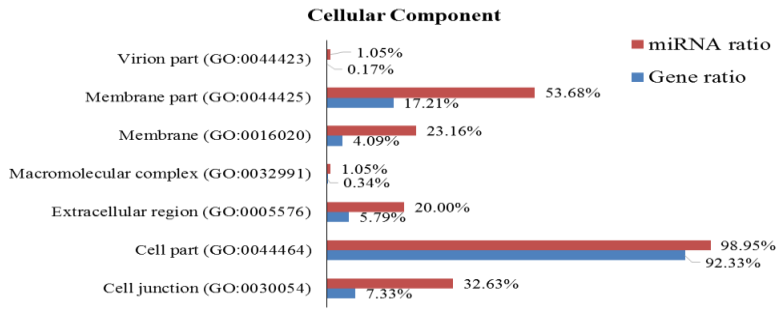


Figure 5-7. Gene ontology-based term classification of cotton miRNA targets (A: Biological process; B: molecular function; C: Cellular component).

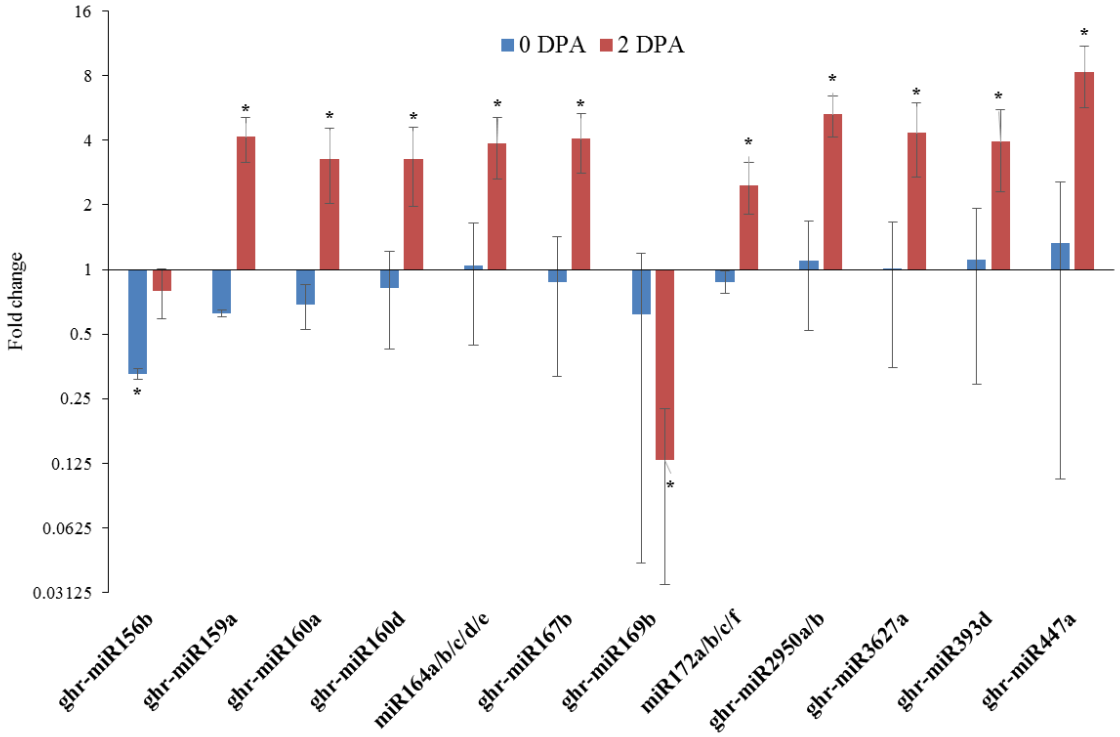


Figure 5-8. Validation and comparison of expression of fiber-development-related miRNAs by qRT-PCR amongst -2, 0, and +2 DPA ovules. miRNA expression is represented as fold change 2^{-}

$^{(\Delta\Delta CT)}$ ($\Delta\Delta CT = \Delta CT_{0\ DPA} - \Delta CT_{-2\ DPA}$ or $\Delta\Delta CT = \Delta CT_{2\ DPA} - \Delta CT_{-2\ DPA}$) relative to -2 DPA miRNA expression. "*" denotes significantly statistical difference at *p-value* ≤ 0.05 by independent student t test.

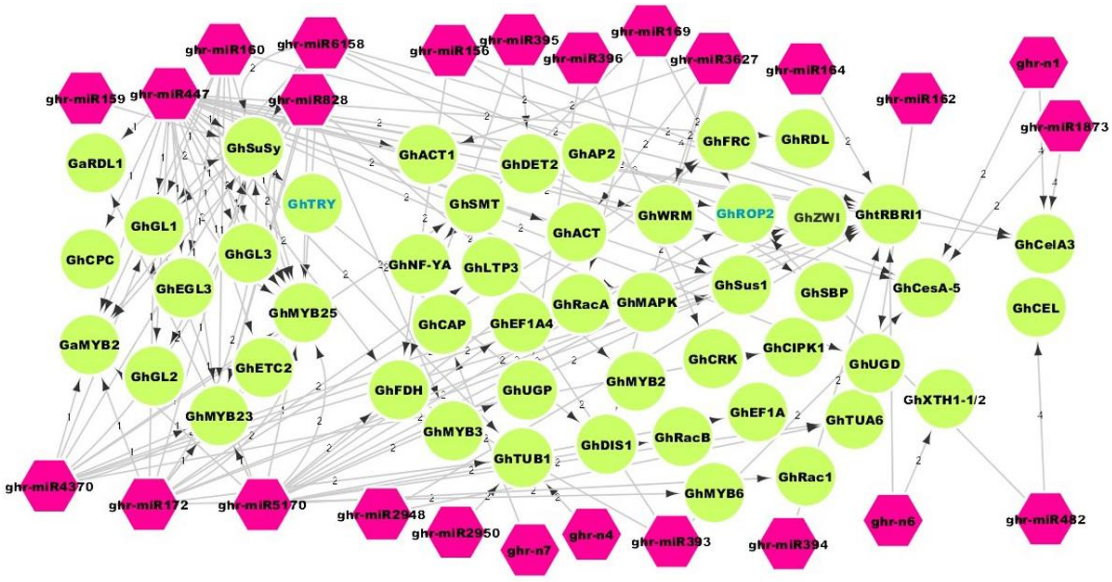


Figure 5-9. Cotton miRNA-mediated interaction network during fiber development stages, including fiber initiation, elongation, and secondary cell wall biosynthesis.

CHAPTER 6: microRNA evolution and expression analysis in polyploidizd cotton genome

Abstract

Upland cotton (*Gossypium hirsutum* L.), the most important domesticated fiber plant in the world, is a tetraploid species, originating from the reunion of two ancestral cotton genomes (A and D genomes) approximately 1-2 Myr ago. It has been reported that a great number of genes were quickly erased or preferentially remained after whole genome duplication (WGD), ultimately leading to morphogenesis evolution. However, microRNAs (miRNAs), as a new class of gene regulators, have not been studied in polyploidization event. Here, we systematically investigated miRNA expansion, expression pattern, miRNA targets amongst three cotton species *G. hirsutum* (AADD), *G. arboreum* (AA), *G. raimondii* (DD). Our results show that certain highly conserved miRNAs were likely to be lost whereas certain were remained after genome polyploidization. Our results also show that cotton-specific miRNAs might undergo remarkably expansion, which results in overall miRNA increase in upland cotton. Based on the sequenced genomes of *G. arboreum* and *G. raimondii*, we are capable for the first time to categorize the origin of miRNAs and coding genes in upland cotton. According to expression comparison of miRNAs and miRNA*s in 6 cotton developmental stages, we found that different genome-derived miRNAs and miRNA*s displayed asymmetric expression pattern, implicating their diverse function in upland cotton phenotype. miRNA target analysis demonstrates that no targeting preference was observed between different genome-derived miRNAs and their target genes in upland cotton. miRNA target comparison in the three cotton species further reveals that origin of

miRNAs and coding genes has no impact on becoming miRNAs and its targets, despite some miRNAs and their targets are extremely conserved in the three cotton species. GO- and KEGG-based analysis of conserved miRNAs show that conserved miRNAs and their targets participate in a series of important biological processes and metabolism pathways. Additionally, A-derived miRNAs might be more responsible for ovule and fiber development.

Introduction

miRNAs are short non-coding RNA regulators involved in post-transcriptional gene regulation by either mRNA degradation or translational repression (Ambros, 2004). They are initially transcribed to primary miRNAs (pri-miRNAs) from intergenic or intron of genome mainly by RNA polymerase II. Pri-miRNAs are then processed successively to miRNA precursors (pre-miRNAs) with a typical stem-loop structure and mature miRNAs. In plants, these processes are conducted by Dicer-like 1 (DCL1). Mature miRNAs are generally considered to be transported to cytoplasm, in which they are incorporated into RNA-induced silencing complex (RISC) where miRNAs cleave the target mRNAs or inhibit their target translation (Bartel, 2004). Extensive studies have demonstrated that miRNAs play crucial roles in a wide range of biological processes, including development, metabolism, cell profiling, and stress response (Ambros, 2004). According to years of study, miRNAs have been identified to widely exist in near all of eukaryotes, owning highly conserved sequences and displaying highly conserved regulatory function in plant or animal kingdom (Meyers et al., 2008). In general, most animal miRNA precursors are in almost identical lengths about 70-80nt, whereas plant miRNA precursors show a much greater length variability (Griffiths-Jones et al., 2003). In general, plant mature miRNAs

are highly conserved rather than their precursors (Griffiths-Jones et al., 2003).

Cotton is one of most important economic crops in the world and is also a powerful single-cell model for research of cell wall and cellulose biosynthesis due to producing excellent nature fiber (Haigler et al., 2012). There were two major events during cotton evolution history: cotton ancestor diverged into A-genome (AA) and D-genome (DD) diploid cotton species about 7-8 Myr ago; about 1-2 Myr ago, the two types of diploid cotton genomes reunioned together, resulting in allotetraploid cotton species (AADD), including the most widely planted upland cotton (*Gossypium hirsutum* L.). miRNA-mediated gene regulation is viewed as an ancient evolutionary mechanism to control gene expression, but miRNA origin is still controversial. Land plants evolved from a group of green algae, perhaps as early as 510 million years ago, resulting in increasing levels of complexity from the earliest algal mats, through bryophytes, lycopods, ferns to the complex gymnosperms and angiosperms (Addo-Quaye et al., 2009). It has been found that some miRNAs existed not only in dicots and monocots but also in ferns, lycipods, and mosses, which indicates these miRNAs are highly conserved in land plants and have ancient origins (Axtell and Bartel, 2005). In fact, besides advanced multicellular plant, miRNAs are also identified in the unicellular green alga *Chlamydomonas reinhardtii*, some of which exhibit differential expression during Chlamydomonas gametogenesis (Zhao et al., 2007). These findings indicates that miRNA pathway is an ancient mechanisms of gene regulation that evolved prior to the emergence of multicellularity. However, no miRNA homolog were found among plants, animals, and green algae, suggesting miRNA may have evolved independently in the lineages leading to animals, plants, and green algae (Zhao et al., 2007). miRNAs are also identified in red alga, *Porphyra yezoensis* that shares common ancestor

with green plants including green algae and land plants (Liang et al., 2010). Interestingly, some of red alga miRNAs are highly conserved with those in land plants, while some of them are conserved with those in green algae (Liang et al., 2010). Therefore, one possible explanation might be used to answer this point. This is some conserved miRNAs in land plants and red alga might be lost in green algae when diverging into green algae and land plants. Another proposed model for miRNA evolution is that miRNA genes might originate from inverted duplication of target gene sequences by forming two adjacent gene segments in either convergent or divergent orientation (Allen et al., 2004). However, a considerable number of plant miRNAs or siRNAs are identical or homologous to transposable elements (TEs), raising a model that miRNAs might come from TEs (Li et al., 2011; Piriyapongsa and Jordan, 2008; Zhang et al., 2011). Genome-wide analysis of rapidly evolving miRNA families showed that some miRNAs might be consequence of the whole genome duplications (WGDs), tandem duplications and segmental duplications followed by their dispersal and diversification, similar to the evolution process of protein-coding genes (Maher et al., 2006). Collectively, as a tetraploid species, upland cotton theoretically inherited two sets of miRNA system from A- and D-genome diploid species. Also, upland cotton miRNAs were likely to undergo the above mentioned evolution events after genome reunion, probably contributing to fiber or other phenotype changes. However, little is known about miRNA origin, expansion, loss, duplication, whether different derived miRNAs exchange with or affect each other, and how different genome-derived miRNAs and different genome-derived coding gene interact in cotton.

Recently, two important representative cotton diploid species, *G. raimondii* (AA) (Paterson et al., 2012) and *G. arboreum* (DD)

(http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AYOE01#contigs) were subsequently sequenced, offering a great opportunity to investigate evolution events in upland cotton. Here, combining with two sequenced genomes, we systematically identified conserved miRNAs and their targets from upland cotton and its two ancestors, *G. raimondii*, and *G. arboreum*. Based on the two genome information, we are the first to categorize miRNAs and coding genes of upland cotton to different genome origin, A, D, or AD. Our results showed the overall miRNAs significantly expanded in upland cotton, whereas some highly conserved miRNAs maintain a relatively stable level. Different genome-derived miRNAs exhibit asymmetric expression pattern in the 6 developmental stages in upland cotton. Our study shed a new light on cotton miRNA evolution in tretraploid background and how cotton miRNAs from different genomes interact with coding genes from different genomes.

Materials and methods

Small RNA sequencing and RNA-seq

The seeds of *G. hirsutum* cv. TM-1 were sterilized with 70% (v/v) ethanol for 60 s, 6% (v/v) bleach for 6-8 min, and then were washed with sterilized water for at least 3 times. The sterilized seeds were planted in 1/2 Murashige and Skoog (MS) medium (pH 5.8) containing 0.8% agar under a 16 h light/8 h dark cycle at room temperature for 10 d. The seedling was replicated for 5 times, each replicates contained 5 seedlings. Ten-day-old cotton seedlings were harvested and immediately frozen in liquid nitrogen. Total RNAs was extracted from each tissue sample using the mirVana miRNA isolation kit (Ambion, Austin, TX) according to the manufacturer's protocol. RNAs were quantified and qualified by Nanodrop ND-1000 (Nanodrop technologies, Wilmington, DE, USA). Each RNA

sample were sent to BGI (Shenzhen, China) for small RNA sequencing.

miRNA data and genome data

Plant miRNAs were downloaded from miRBase (Release 20, http://mirbase.org/); the genome and its annotation data of *G. raimondii* were from http://www.phytozome.net/cotton.php; the whole-genome shotgun sequences of *G. arboreum* were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AYOE01#contigs); the coding genes of *G. arboreum* were downloaded from http://cgp.genomics.org.cn/page/species/download.jsp?category=arboreum (Li et al., 2014); ESTs, GSSs, and mRNAs of *G. hirsutum* were also obtained from miRBase; asides the sequenced seedling data above, 14 other small RNA sequencing dataset (*G. hirsutum* (GSM911189, GSM911190, GSM911191,GSM634228, and GSM911192), *G. raimondii* (SRR616255, SRR616256, and SRR616257), and *G. arboreum* (SRR1029586, SRR1029586, SRR1029586, GSM717572, GSM717571, and GSM717570) and three degradome sequencing dataset (GSM1061853, GSM1008999, GSM1008997) were downloaded from NCBI. RNA family database (Rfam 11.0) were available from http://rfam.sanger.ac.uk/ (Griffiths-Jones et al., 2003); other cotton noncoding RNAs of *G. raimondii* were from http://cgp.genomics.org.cn/page/species/download.jsp?category=raimondii. miRNAs identified by Gong and its co-workers were also integrated into our miRNA analysis (Gong et al., 2013).

miRNA analysis

All of raw small RNA sequenced dataset were cleaned first through trimming adaptor sequences and filtering low-complexity reads. Reads that are 100% mapped to other RNAs (repeat RNA, rRNA, snRNA, snoRNA, and tRNA) were discarded. Cleaned reads were aligned against all of known plant mature miRNAs to categorize conserved reads and non-conserved reads under the cutoff of 3 mismatches. ESTs, GSSs, assembled ESTs and GSSs, *G. ramondii* genome, and *G. arboreum* shotgun sequences were used as data source for miRNA identification in *G. hirsutum* using our developed software, miRDeepFinder (Xie et al., 2012). All of identified miRNA precursors of upland cotton were aligned against *G. raimondii* genome and *G. arboreum* shotgun sequences, respectively. A miRNA was viewed to derive from the subject genome (A-, D-, or AD-derived) if the alignment identity is over 95%, otherwise from the unknown. Expression of miRNAs and miRNA*s were normalized to reads per million reads (RPM).

Annotated mRNA database of upland cotton from NCBI and assembled contigs from our cotton EST database were employed to identify miRNA targets by Target-align with default parameters (Xie and Zhang, 2010). To validate predicted miRNA targets, Cleveland (Version 4.3, http://axtell-lab-psu.weebly.com/cleaveland.html) (Addo-Quaye et al., 2009) was used to do degradome analysis with some modifications, where maximal mismatches between miRNAs and target sites are less than 4 and no mismatches is allowed on the 10th and 11th nucleotides on miRNAs (Meyers et al., 2008).

Coding gene origin categorization

A total of 24,523 upland cotton mRNA sequences from cotton mRNA databases (2,531 sequences) in NCBI and our annotated cotton EST database (21,991 coding

sequences) (Xie et al., 2011) were BLASTNed against cotton A genome (*G. arboreum*) and D genome (*G. raimondii*) under the cutoff of overall identity of 95%, otherwise its origin is 'unknown'. The upland cotton mRNA origin is determined by the alignment with the highest identity between two genomes.

GO- and KEGG-based analysis of miRNAs and their targets in upland cotton

A total of 36 highly conserved miRNA targets were selected to perform BLASTX alignment against *Arabidopsis* protein database (TAIR10, http://www.arabidopsis.org/). Homologous *Arabidopsis* genes of the miRNA targets were submitted to GOrilla (http://cbl-gorilla.cs.technion.ac.il/) for Gene Ontology term enrichment (Eden et al., 2009) with a default *p-value* threshold. For KEGG pathway enrichment analysis, these miRNA targets were BLASTXed against GO protein database to retrieve related KEGG annotation. Finally, hypergeometric test was performed for pathway enrichment significance analysis.

Results

Identification of miRNAs in cotton species

A total of 6 small RNA sequencing libraries of *G. hirsutum* were combined together to systematically identify conserved miRNAs in upland cotton, which had at least three reads in one library and no more than 3 mismatches with a known plant miRNA. It resulted in 5,416 unique conserved reads in total, which subsequently were used to identify miRNA precursors from database of EST, GSS, assembled EST and GSS of upland cotton, *G. raimondii* (D genome), and whole-genome shotgun sequences of *G. arboretum* (A genome), respectively. After removing repeated precursors, finally, 528 conserved miRNA

precursors in 197 miRNA families were identified, including miR156/157, miR160, miR162, miR164, miR159/319, miR172, and miR390 (Figure 6-1 and Supplementary 6-1). Of these upland cotton miRNA precursors, 44, 261, and 223 miRNAs were identified from 6 samll RNA sequence data of upland cotton based on the sequences data of upland cotton, *G. arboreum*, and *G. raimondii,* respectively. Considering miRNA precursor sequences might be altered partially during the cotton evolution, to better track cotton miRNA origin, we assume a miRNA derived from a certain genome if the miRNA precursor could be found to have homologous sequences in the genome (Identify ≥ 95%), even if the homologous counterparts in the genome could not be folded into a stem-loop structure of miRNA. , A total of 523 (99.0%) miRNA precursors of upland cotton were found to have homologous counterparts in A genome, D genome, or both A and D genomes, including 176 A-derived, 106 D-derived, and 241 AD-derived precursors (Figure 6-2A and Supplementary 6-1). In *G. arboreum* (A genome), a total of 318 conserved miRNAswere identified from 6 small RNA sequencing dataset of *G. arboreum* stored in GeneBank, belonging to 77 miRNA families (Supplementary 6-2). Similarly, 327 conserved miRNAs were detected from 3 small RNA sequencing dataset of D genome of *G. raimondii*, consisting of 84 miRNA families (Supplementary 6-3).

The All-to-All alignment of identified miRNA precursors amongst *G. hirsutum*, *G. raimondii*, and *G. arboreum*, showed that 91 miRNAs are able to be found in all of three species, whereas 292, 121 and 166 miRNA precursors were specific to *G. hirsutum*, *G. arboreum* and *G. raimondii*, respectively (Figure 6-3A). Furthermore, 11 miRNA precursors co-exist in *G. arboreum* and *G. raimondii* but not in *G. hirsutum,* implicating they might be from the same cotton ancestor and lost in *G. hirsutum* after genome reunion.

Otherwise, the 11 miRNA precursors were missed in *G. hirsutum* by its limited available genome sequence all of these conserved miRNAs from the three species were totalized into 161 miRNA families, 44 of which are owned by the three species. 47, 21, and 27 miRNA families are specific to *G. hirsutum*, *G. arboreum* and *G. raimondii*, correspondingly (Figure 6-3B). 4 miRNA families are only common to both *G. arboreum* and *G. raimondii.*

miRNA expansion in upland cotton

Upland cotton was the consequence of reunion of cotton A genome and D genome about 1-2 Mya. Meanwhile, miRNAs are highly conserved in plant kingdom and appeared at latest after divergence from chlorophyta to coniferophyta, and before the occurrence of cotton ancestor species as well (Paterson et al., 2012). Theoretically, there were two miRNA sets from cotton A genome and D genome that were integrated into upland cotton. To investigate how miRNAs expanded in upland cotton, 27 representative conserved miRNAs were selected from 23 land plant species including three cotton species, moss, Arabidopsis, rice, and tobacco (Figure 6-4). Overall, upland cotton miRNA family experienced significant expansion relative to its two ancestor species, *G. raimondii* and *G. arboreum*. Mann-Whitney U test showed that upland cotton miRNA families are remarkably larger than those in *G. raimondii* (*p-value*: 0.001038) and *G. arboreum* (*p-value*: 0. 001266). However, no significance difference on miRNA family size was detected between *G. raimondii* and *G. arboreum* (*p-value*: 0.919480). There are at least two major conserved miRNA family groups in cotton that are conserved either in other land plants or only in cotton species. The first group is highly conserved in both eudicotyledons and monocotyledons, including miR156/157, miR159/319, miR160, miR162, miR164,

miR166, miR167, miR168, miR169, miR171, miR172, miR390, miR393, miR394, miR395, miR396, miR398, miR399, and miR482. We also performed Mann-Whitney U test on the first group between *G. hirsutum* and other non-cotton land plants, respectively. Except *Carica papaya*, *Ricinus communis*, *Theobroma cacao*, *Brassica napus*, and *Glycine max*, miRNA family size of *G. hirsutum* is generally similar with other 15 land plants (*p-value* > 0.05) (Supplementary 6-4), indicating this miRNA group might be such critical that they should keep stable in miRNA function and miRNA size in order to maintain basic core biological activities, such as plant development and phase change. The other type of miRNA family group mainly exists in the three cotton species, including miR828, miR2916, miR3441, miR5139, miR5227, miR5255, miR5272, and miR5528 (Figure 6-4). Also, miR2916, miR5139, miR5227, miR5255, miR5272, and miR5528 were significantly expanded and are specific to upland cotton.

Cotton miRNA expression comparison

To investigate whether different genome-derived miRNAs exhibit any differential expression in upland cotton, we investigated the expression of top 50 abundant miRNA family and miRNA star during seedling and fiber at 5 developmental stages by heatmap-based analysis. Overall, miR396, miR166, miR156/157, miR168, miR167, and miR390 were expressed higher than other miRNAs from seedlings to fiber development, indicating they are critical fundamental miRNAs to cotton (Figure 6-5A and Supplementary 6-5). At least 13 miRNA families including miR2118, miR7505, miR408, miR156/157, miR2947, miR171, miR160, miR169, and miR3954, were significantly expressed lower in fiber stages than in seedlings. Moreover, the global miRNA expression in seedling is higher than

that in fiber, but miRNA expression profile was a little changed at the 10-DPA fiber. miRNA* is generally considered to be degraded after unbound from miRNA/miRNA* duplex. However, as miRNAs, miRNA*s have been also found to participate in negative gene regulation sometime with a rich expression abundance (Wu et al., 2013). In addition, a few of miRNA*s demonstrate highly phylogenetic conservation across different species similar to mature miRNAs, implicating their potential function (Guo and Lu, 2010). Based on the analysis of the 6 small RNA sequencing datasets, the expression of the majority of miRNA*s was correlated with the expression of its corresponded mature miRNAs. Heatmap analysis of the miRNA* expression showed that some miRNA*s were differentially expressed amongst seedlings and fiber developmental stages (Figure 6-5B and Supplementary 6-5). For instance, miR172* and miR390* were upregulated in 10-DPA fiber, 3-DPA fiber, and seedlings, and were downregulated in other developmental stages. On the contrary, miR171*, miR2949*, miR3954*, and miR164* were down regulated in seedlings and 10-DPA fiber, but upregulated in other developmental stages.

To determine which genome-derived miRNAs contribute more to miRNA expression and miRNA-mediated gene regulation in upland cotton, we reanalyzed the top 50 abundant miRNA families and their miRNA*s based on the genome origin. Overall, AD-derived miRNA families account for predominant expression and other A- or D-derived miRNA families had sporadic expression (Figure 6-6A). However, at least three A-derived miRNA families including miR7510, miR5225, and miR7506 were expressed higher than those of AD-derived families. Some miRNAs from the three origin were in a similar expression level, like miR172, miRN159/319, miR3954, and miR164. Similarly, the majority of miRNA* expression was contributed by AD-derived miRNA* (Figure 6-

6B) and A- or D-derived miRNA*s made partial expression. D-derived miR399* expression abundance was higher than AD-derived's. miR482* and miR172 from AD, A, or D genome had similar expression level. When it comes to a single individual stage, there is some fluctuation on the composition of expression of different genome-derived miRNAs and miRNA*s (Figure 6-6C and 6-6D). AD-derived miRNA families account for the largest part of miRNA expression on average of 87.80%, followed by D- and A- derived miRNA families on average of 8.70% and 3.51%, respectively. D-derived miRNA families were expressed more than A-derived miRNA families in all of the 6 developmental stages (Figure 6-5C). D-derived miRNA families in 10-DPA fiber and seedlings had the largest expression proportions, 25.52% and 14.32%, respectively. Meanwhile, A-derived miRNA family expression were apparently raised to 13.56% in 10-DPA fiber. Both A- and D-derived miRNA expression proportion underwent a similar alternation tendency during fiber stages, hinting there might be a certain dynamic expression balance amongst A-, D-, and AD-derived miRNAs in order to satisfy the need of fiber development. Similarly, different genome-derived miRNA*s expression in the 6 developmental stages also resembled that of mature miRNAs (Figure 6-6D). 10-DPA fiber and seedlings expressed the two largest D-derived miRNA*s, accounting for 16.78% and 10.85%, respectively. However, the proportion of A-derived miRNA* expression in seedlings is obviously higher than that in A-derived miRNAs. D-derived miRNA* expression seemed to be not so closely correlated with the expression of D-derived miRNAs. Particularly in 0-DPA and 3-DPA fiber, miRNA* expression percentage is higher than their miRNAs', implicating D-derived miRNA*s in the two stages might participate in more regulation.

Upland cotton protein-coding genes and miRNA targets origination

To investigate how different genome-derived miRNAs interact with different genome-derived targets, we first aligned all of protein-coding gene of upland cotton against the genomes of *G. raimonddi* and *G. arboreum*. Using a cutoff of 95% overall alignment identity, 10,780 (43.96%) and 11,244 (45.85%) protein-coding genes were classified to be A-derived and D-derived, respectively (Figure 6-2B and Supplementary 6-6). Only 2,498 (10.18%) genes' origin cannot be ascertained. Overall, most of genes in upland cotton maintain highly conserved with the two ancestor genomes though A genome and D genome have been reunioned for 1-2 million years.

Using the identified 528 conserved miRNAs, 6,864 out of 24,523 (27.99%) protein-coding genes were identified to be targets of 485 miRNAs in upland cotton (Supplementary 6-7). Based on the three degradome sequencing data of upland cotton, we were able to validate 140 miRNA-target pairs, including 69 miRNAs and 117 unique target genes. Then we further investigated the distribution of miRNAs and their targets based on origin of miRNAs and their targets, respectively. 235 AD-derived miRNAs were predicted to the largest part of targets (4,112 genes, 59.91%) (Figure 6-2C). Meanwhile, 140 A-derived and 105 D-derived miRNAs were predicted to target 2,282 (33.29%) and 2,577 (37.54%) genes, individually (Figure 6-2C). Similarly, 2,955 (43.05%) A-derived and 3,143 (45.79%) D-derived genes were found to be targeted by 451 and 485 miRNAs, respectively (Figure 6-2D). Of the 2282 gene targeted by 140 A-derived miRNA, 1,020 and 1,009 genes belong to D-derived and A-derived, respectively (Table 6-1). Likewise, of the 2,577 genes targeted by 105 D-derived miRNA, 1,167 and 1,122 genes belong to D-derived and A-derived, respectively (Table 6-1). Compared with the 140 A-derived miRNAs and their 2,282 targets,

D-derived miRNAs in upland cotton showed more flexibility to target more genes, consisting of 7,094 miRNA-target pairs (Table 6-1 and Supplementary 6-7).

Comparison of miRNA targets in three cotton species

Besides investigating how different genome-derived miRNAs act on different genome-derived coding genes, we further checked how miRNAs from *G. raimondii* (DD) and *G. arboreum* (AA) target the coding genes of upland cotton. As shown in Figure 6-7, the proportions of miRNA-target pairs, miRNAs, and unique targets don't show apparent bias amongst the three cotton species. In addition, we used upland cotton miRNAs to predict miRNA targets in coding genes of *G. raimondii*. There is also no significant difference for A- or D-derived miRNAs of upland cotton to target A-genome coding genes in *G. raimondii* (Figure 6-8). These indicates that at least mature miRNAs from different genomes have no preference to target coding genes, but rather miRNA-target relationship depends on whether the coding genes could carry a target site for the miRNAs. The origin of miRNAs and mRNA targets doesn't influence their regulation relationship.

In addition, we also further investigated whether genes in A genome and D genome that are homologous with miRNA targets of upland cotton could still retain conserved target sites. To this point, we firstly performed All-to-All alignment amongst the coding genes of the three cotton species. A total of 21,028 coding genes of upland cotton were homologous with 14, 975 coding genes of *G. arboreum* and 15,235 coding genes of *G. raimondii*, respectively (Supplementary 6-8). Then we searched miRNA targets in *G. arboreum* and *G. raimondii* with miRNAs in upland cotton. 393 and 383 miRNAs of upland cotton were predicted to target 917 genes in and 1657 genes of *G. raimondii*,

respectively (Supplementary 6-8). We found not too many miRNA target sites are conserved in three cotton species. Of these homologous genes, only 110 genes of *G. arboreum* and 78 genes of *G. raimondii* were also commonly targeted by the same miRNAs with the homologous ones in upland cotton, respectively (Supplementary 6-8). 94 and 65 genes in *G. arboreum* and *G. arboreum* own the same target sites with the homologous ones in *G. hirsutum*, respectively. 48 homologous genes in the three cotton species retain extremely conserved target site*s*, indicating highly conserved regulatory function for these miRNAs and their targets. For example, AGO1 and AP2 in the three cotton species have the same target sites for miRNA168 and miRNA172, respectively (Figure 6-9A and 6-9B). However, 37 and 70 homologous genes in *G. arboreum* and *G. raimondii* were found to have different or no miRNA target sites. For instance, target sites of miR156-LIGULELESS 1 protein and miR5227-Caffeic acid 3-O-methyltransferase exist in *G. hirsutum* and *G. arboreum*, but not in *G. raimondii* (Figure 6-9C and 6-9D). On the contrary, target sites of miR530-Basic helix-loop-helix (bHLH) family protein and miR395-Cellulose synthase exist in *G. hirsutum* and *G. raimondii*, but not in *G. arboreum* (Figure 6-9E and 6-9F). Furthermore, 19 and 15 homologous genes in *G. arboreum* and *G. raimondii* display some minor variation, respectively, despite they share the same miRNAs with the ones in upland cotton (Supplementary 6-8).

GO- and KEGG-based analysis of different genome-derived coding genes and miRNA targets in upland cotton

As no significant preference for different genome-derived miRNAs to target different genome-derived coding genes in upland cotton, we then asked what function for

different genome-derived miRNAs are responsible for. To this end, we selected 36 highly conserved miRNAs to do *GO- and KEGG*-based function analysis, since their expression accounted for 69.13% in the whole miRNA expression repertoire, including miR156/157, miR159/319, miR160, miR162, miR395, miR396, miR397, and miR398. Our GO analysis showed that A-derived miRNAs and their targets were significantly enriched to the biological process of ovule development (GO:0048481) (*p-value*: 0.0004) (Figure 6-10A), involving miRNAs like ghr-miR172i, ghr-miR5270f, ghr-miR167f, and ghr-miR5528b, and their targets like ARF8, AGL6, ARF6, and AP2 (Supplementary 6-9). That might indicate A-derived miRNAs and their targets execute more function in ovule development than other origin types of miRNAs. D-derived miRNAs and their targets significantly overlapped with 9 biological processes including inositol phosphate metabolic process (GO:0043647), nitrogen compound transport (GO:0071705), anatomical structure development (GO:0048856) (Figure 6-10B and Supplementary 6-9). For AD-derived miRNAs and their targets, 19 biological processes were overlapped in a significant statistic level, including protein targeting to membrane (GO:0006612), shoot system development (GO:0048367), system development (GO:0048731), and biological regulation (GO:0065007) (Figure 6-10C and Supplementary 6-9). In addition, these miRNA targets were significantly enriched to 36 KEGG pathways (*p-value* $\leq$ 0.05), including photosynthesis (ATH00195), biosynthesis of plant hormones (ATH01070), fatty acid biosynthesis (ATH00061), glycolysis / gluconeogenesis (ATH00010), and fructose and mannose metabolism (ATH00051) (Supplementary 6-10).

Discussion

miRNA conservation and divergence in cotton

It has been reported that cotton allopolyploidy reuniting 1-2 Myr ago led to about 30–36-fold duplication of ancestral angiosperm (flowering plant) genes in two elite cotton species, *G. hirsutum* and *G. barbadense* (Paterson et al., 2012). miRNAs as a class of small endogenous RNAs that direct post-transcriptional repression of coding genes, were estimated to target at least 1/3 protein-coding genes in human (Ambros, 2004). Therefore, in theory, miRNAs are presumably expected to expand 30-36 fold in accordance with the gene expansion after the reunion of A and D genome in upland cotton. However, our miRNA family comparison in 17 land plants showed that no significant change was observed on these highly conserved miRNAs, such as miR156/157, miR159/319, miR160, miR162, miR164 (Figure 6-4). Also, we didn't find any remarkable difference of these conserved miRNA families amongst other non-cotton land plants. It seems to be a paradox that conserved miRNA evolution was too slow to be consistent with coding gene fast expansion in cotton. A couple of possible explanations might be employed to answer the plausible paradox. First of all, as single-gene (Birchler, 2012), conserved miRNAs would decrease the polyploid effects using selective or partial loss of miRNAs. At the same time, these miRNAs were preferentially retained with the retention of their targets during polyploidization event. Secondarily, miRNA targets in two cotton ancestor species were also so conserved that there is always a balance between a miRNA and their target sites on coding genes. It is likely for some genes to be non-functional, resulting in some miRNA functional degeneration and subsequently miRNA loss. Thirdly, in addition to conserved miRNAs, we also identified many cotton-specific miRNAs, like miR2916, miR5139, miR5227, miR5255, miR5272, and miR5528 (Figure 6-4). These cotton-specific miRNA

qualities in *G. arboreum* or *G. raimondii* are dramatically less than those in *G. hirsutum*, indicating they underwent fast multiple duplication. Their expansion might bridge a miRNA regulation gap between conserved miRNA stability and fast gene duplication in upland cotton. However, according to the comparison of miRNAs in *G. arboreum* and *G. raimondii,* no independent loss/gain of conserved miRNAs, similar copy number of conserved miRNA families, and similar family-wide nucleotide diversities were observed in the two diploid cotton species (Gong et al., 2013). Collectively, at least conserved miRNA family size in upland cotton is independent to size and ploid of genome, somewhat contrasting with coding gene evolution pattern. Fast expanded cotton-specific miRNAs might distinguish upland cotton from their ancestor species, contributing in import phenotype change, like fiber quality and density.

miRNA expression in upland cotton

Completely sequencing *G. arboreum* and *G. raimodii* genomes allow us for the first time to systematically investigate the miRNA origins and their expression in allotetraploid upland cotton. Our result shows that cotton A-derived miRNAs and D-derived miRNAs display asymmetric expression, as well as for the corresponding miRNA*, despites both AD-derived miRNAs and miRNA*s predominate the whole expression of miRNAs and miRNA*s (Figure 6-6C and 6-6D). Some of A- and D-derived miRNAs exist in the same miRNA family and were expressed in a different level. This indicates that different genome-derived miRNAs might have different regulation tasks in upland cotton. Moreover, we also compared miRNA family expression of cotton seedling amongst *G. hirsutum*, *G. arboreum*, and *G. raimondii* (Supplementary 6-11). It turned out that the majority of

miRNA families display differential expression when compared with each other, largely consisting with asymmetric expression of miRNAs between *G. arboreum* and *G. raimondii* reported by Gong and its co-workers (Gong et al., 2013). There are some typically upregulated or downregulated miRNA families in upland cotton when compared with those in *G. arboreum*, or *G. raimondii,* such as miR172, miR156/157, miR394, and miR535 (Figure 6-11). To date, there is no report of how a miRNA family expression is altered after genome ploidization. In coding genes, extensive changes to patterns of parental gene expression (also known as "transcriptome shock") always occurs in allopolyploid formation (Adams and Wendel, 2005). For example, some genes were rapidly silenced in newly formed Arabidopsis allotetraploids (Comai et al., 2000). But it was argued to be likely led by interspecific hybridization rather than polyploidization (Wang et al., 2006). According to comparison of floral gene expression in allohexaploid *Senecio cambrensis* with that in its parent species, *S. vulgaris* (tetraploid) and *S. squalidus* (diploid), and their triploid F1 hybrid, *S. x baxteri,* many transcriptome-shocked genes observed in *S. x baxteriis* is ameliorated after genome duplication in the first generation of synthetic *S. cambrensis*, and this altered expression pattern is stably passed to the subsequent offspring generations (Hegarty et al., 2006). However, in our study, nearly half conserved miRNA expression in upland cotton were upregulated and downregulated when compared with that in the parental species, respectively (Supplementary 6-9). Therefore, to illuminate what cause the miRNA expression pattern after genome duplication in upland cotton, more studies including transcriptome comparison or miRNA target expression analysis are needed.

miRNAs and their targets in upland cotton

In our miRNA expression data of upland cotton, D-derived miRNA expression is generally higher than that of A-derived miRNAs in the selected 6 developmental stages (Figure 6-6C). Also, D-derived miRNAs had a sudden increase at 10-DPA fiber and then was decreased back to normal level as other 4 fiber developmental stages. Considering 10 DPA is the critical transition stage in which fiber cells is developing from initiation to elongation (Qin and Zhu, 2011), it is very interesting for us to trace the function of D-derived miRNAs. Three miRNAs with dramatical upregulation in 10-DPA fiber were ghr-miR164b, ghr-miR172b, and ghr-miR390b. Our target analysis shows that a total of 44 genes were targeted by the three miRNAs, including some well-known transcription factors, like NAC11 (JQ914142.1) and AP2 (contig24554 and contig10679). miR164 was firstly identified in Arabidopsis to negatively control lateral root emergency via targeting a five NAM/ATAF/CUC (NAC) domain–encoding mRNA, NAC1 (Guo et al., 2005). Actually, subsequent studies unraveled that miR164 is a versatile regulator that get involved in a series of biology processes by targeting NAC genes, such as negatively regulating drought resistance in rice (Fang et al., 2014), negatively modulating resistance of wheat to stripe rust (Feng et al., 2014), and the proper formation and separation of adjacent embryonic, vegetative, and floral organs in Arabidopsis (Mallory et al., 2004). We also identified D-derived ghr-miR164b to target NAC11 in upland cotton. Additionally, ghr-miR164b's expression was gradually upregulated from 0-DPA to 10 DPA and then were suddenly downregulated back to the level similar with that in 0-DPA, 3 DPA, and 7 DPA (Supplementary 6-1). Meanwhile, our ghr-miR164b's expression pattern is also consistent with the counterpart in microarray-based miRNA analysis result from Pang and its

colleagues (Pang et al., 2009). In addition, they also found the expression of miR164 of *G. arboretum* was barely detected in the ovules at 0 and +3 DPA, but miR164 was expressed at 0 or +3 DPA in two other fiber-producing cotton species, *G. thurberi* and N1N1 mutant (Pang et al., 2009). Interestingly, we detected that gar-miR164a-f (the same mature sequence with ghr-miR164b) in *G. arboreum* were expressed in a similar level with ghr-miR164b at 8 DPA fiber (Supplementary 6-2). Thus, we infer that miR164 might be activated earlier in upland cotton than in *G. arboreum*. In fact, asides D-derived ghr-miR164b, we also found there are other miR164s that have the same mature miRNAs with ghr-miR164b but with different precursors (Supplementary 6-1). Therefore, it is too early for us to clarify which miR164 targets NAC11 in upland cotton. Furthermore, we found NAC11 (JQ914142.1) is an A-derived gene (Supplementary 6-6), which might enable us to raise a hypothesis that A-derived NAC11 might be involved in improving fiber development by acquiring the regulation of miR164 in upland cotton.

In addition to the highly expressed D-derived ghr-miR164b at 10 DPA, D-derived ghr-miR172 that was predicted to target an AP2 domain-containing transcription factor was also found to highly express in the fiber at 10 DPA. AP2, a class of multifunctional transcription factors, was initially found to play a role in the ABC model of flower development in Arabidopsis (Riechmann and Meyerowitz, 1998). Later on, AP2 was uncovered to participate in other regulations including lateral root development (Kang et al., 2013), response to biotic and abiotic stress (Jin et al., 2013), and promoting fruit ripening process (Licausi et al., 2010). In cotton, AP2/EREBP family was found to be highly expressed at 1-5 DPA in the dominant naked-seed cotton mutant relative to the wild type TM-1, indicating a negative role of AP2/EREBP in fiber development (Wan et al.,

2014). Overexpression of a cotton gene of AP2/EREBP family in Arabidopsis was shown to dwarf the transgenic plant and delay the mean bolting time by about 10 days (Zhou et al., 2013). In fact, miR172 has been shown to cause early flowering and disrupt the specification of floral organ identity by downregulating AP2-like target genes in Arabidopsis and tobacco (Aukerman and Sakai, 2003; Mlotshwa et al., 2006). Therefore, based on our results and previous findings, we speculated that ghr-miR172 might be a positive factor for fiber development by negatively regulating AP2 genes in upland cotton.

Conclusion

This is the first study to systematically identify miRNAs and their targets in upland cotton with data from *G. hirsutuem*, *G. araboreum*, and *G. raimondii*, and to classify the genome origin of miRNAs and coding genes. Our results demonstrate that miRNA evolution in upland cotton likely utilized a pattern different from that of coding gene in polyploidziation event, in which traditional conserved miRNAs maintain a stable level and cotton-specific miRNAs acquired a striking expansion. Different genome-derived miRNAs have asymmetric expression pattern, implicating their diverse function to reconstruct upland cotton phenotype after the reunion of A and D genomes. According to miRNA target analysis, no targeting bias was observed between different genome-derived miRNAs and their target genes in upland cotton. Also, miRNA target comparison amongst *G. hirstutum*, *G. arboreum*, and *G. raimondii*, shows that the possibility that a coding gene could become a miRNA target gene is only determined by that whether the gene owns a target site for the miRNAs, rather than being impacted by origin of miRNA or coding gene. Additionally, some miRNAs and their targets are highly conserved among the three cotton species, while

some show minor variation. Together, our findings might broaden our understanding on how miRNAs carry out their function in the context of genome duplication, as well on how cotton miRNAs play roles in fiber development.

Supporting Information

Supplementary 6-1: Conserved miRNAs identified in upland cotton

Supplementary 6-2: Conserved miRNAs identified in *Gossypium arboreum*

Supplementary 6-3: Conserved miRNAs identified in *Gossypium raimondii*

Supplementary 6-4: Mann-Whitney U test for miRNA family size

Supplementary 6-5: Expression of miRNAs and miRNA stars in upland cotton

Supplementary 6-6: Genome origin analysis of protein-coding genes in upland cotton

Supplementary 6-7: miRNA targets of upland cotton

Supplementary 6-8: Common miRNA targets in *Gossypium hirsutum*, *Gossypium arboreum and Gossypium raimondii*

Supplementary 6-9: GO ontology classification of miRNAs and their targets in upland cotton.

Supplementary 6-10: Enriched KEGG pathways of cotton miRNA and their targets in upland cotton

Supplementary 6-11: miRNA expression comparison of seedlings in *Gossypium hirsutum*, *Gossypium arboreum and Gossypium raimondii*

Reference

Adams, K.L., and Wendel, J.F. (2005). Novel patterns of gene expression in polyploid plants. Trends in genetics : TIG *21*, 539-543.

Addo-Quaye, C., Miller, W., and Axtell, M.J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics *25*, 130-131.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nature genetics *36*, 1282-1290.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350-355.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. The Plant cell *15*, 2730-2741.

Axtell, M.J., and Bartel, D.P. (2005). Antiquity of microRNAs and their targets in land plants. The Plant cell *17*, 1658-1673.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Birchler, J.A. (2012). Genetic Consequences of Polyploidy in Plants. 21-32.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y., and Byers, B. (2000). Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. The Plant cell *12*, 1551-1568.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics *10*, 48.

Fang, Y., Xie, K., and Xiong, L. (2014). Conserved miR164-targeted NAC genes negatively regulate drought resistance in rice. Journal of experimental botany *65*, 2119-2135.

Feng, H., Duan, X., Zhang, Q., Li, X., Wang, B., Huang, L., Wang, X., and Kang, Z. (2014). The target gene of tae-miR164, a novel NAC transcription factor from the NAM subfamily, negatively regulates resistance of wheat to stripe rust. Molecular plant pathology *15*, 284-296.

Gong, L., Kakrana, A., Arikit, S., Meyers, B.C., and Wendel, J.F. (2013). Composition and Expression of Conserved MicroRNA Genes in Diploid Cotton (Gossypium) Species. Genome biology and evolution *5*, 2449-2459.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. Nucleic acids research *31*, 439-441.

Guo, H.S., Xie, Q., Fei, J.F., and Chua, N.H. (2005). MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. The Plant cell *17*, 1376-1386.

Guo, L., and Lu, Z. (2010). The fate of miRNA* strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule? PloS one *5*, e11387.

Haigler, C.H., Betancur, L., Stiff, M.R., and Tuttle, J.R. (2012). Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Front Plant Sci *3*, 104.

Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. Current biology : CB *16*, 1652-1659.

Jin, X., Xue, Y., Wang, R., Xu, R., Bian, L., Zhu, B., Han, H., Peng, R., and Yao, Q. (2013). Transcription factor OsAP21 gene increases salt/drought tolerance in transgenic Arabidopsis thaliana. Molecular biology reports *40*, 1743-1752.

Kang, N.Y., Lee, H.W., and Kim, J. (2013). The AP2/EREBP gene PUCHI Co-Acts with LBD16/ASL18 and LBD18/ASL20 downstream of ARF7 and ARF19 to regulate lateral root development in Arabidopsis. Plant & cell physiology *54*, 1326-1334.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., Zou, C.*, et al.* (2014). Genome sequence of the cultivated cotton Gossypium arboreum. Nature genetics *46*, 567-572.

Li, Y., Li, C., Xia, J., and Jin, Y. (2011). Domestication of transposable elements into MicroRNA genes in plants. PloS one *6*, e19212.

Liang, C., Zhang, X., Zou, J., Xu, D., Su, F., and Ye, N. (2010). Identification of miRNA from Porphyra yezoensis by high-throughput sequencing and bioinformatics analysis. PloS one *5*, e10698.

Licausi, F., Giorgi, F.M., Zenoni, S., Osti, F., Pezzotti, M., and Perata, P. (2010). Genomic and transcriptomic analysis of the AP2/ERF superfamily in Vitis vinifera. BMC genomics *11*, 719.

Maher, C., Stein, L., and Ware, D. (2006). Evolution of Arabidopsis microRNA families through duplication events. Genome Res *16*, 510-519.

Mallory, A.C., Dugas, D.V., Bartel, D.P., and Bartel, B. (2004). MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs. Current biology : CB *14*, 1035-1046.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X.,

Carrington, J.C., Chen, X., Green, P.J., *et al.* (2008). Criteria for annotation of plant MicroRNAs. The Plant cell *20*, 3186-3190.

Mlotshwa, S., Yang, Z., Kim, Y., and Chen, X. (2006). Floral patterning defects induced by Arabidopsis APETALA2 and microRNA172 expression in Nicotiana benthamiana. Plant molecular biology *61*, 781-793.

Pang, M., Woodward, A.W., Agarwal, V., Guan, X., Ha, M., Ramachandran, V., Chen, X., Triplett, B.A., Stelly, D.M., and Chen, Z.J. (2009). Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). Genome biology *10*, R122.

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., *et al.* (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature *492*, 423-427.

Piriyapongsa, J., and Jordan, I.K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. Rna *14*, 814-821.

Qin, Y.M., and Zhu, Y.X. (2011). How cotton fibers elongate: a tale of linear cell-growth mode. Current opinion in plant biology *14*, 106-111.

Riechmann, J.L., and Meyerowitz, E.M. (1998). The AP2/EREBP family of plant transcription factors. Biological chemistry *379*, 633-646.

Wan, Q., Zhang, H., Ye, W., Wu, H., and Zhang, T. (2014). Genome-wide transcriptome profiling revealed cotton fuzz fiber development having a similar molecular model as Arabidopsis trichome. PloS one *9*, e97313.

Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C.,

Doerge, R.W., Comai, L., *et al.* (2006). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics *172*, 507-517.

Wu, X., Bhayani, M.K., Dodge, C.T., Nicoloso, M.S., Chen, Y., Yan, X., Adachi, M., Thomas, L., Galer, C.E., Jiffar, T., *et al.* (2013). Coordinated targeting of the EGFR signaling axis by microRNA-27a*. Oncotarget *4*, 1388-1398.

Xie, F., Sun, G., Stiller, J.W., and Zhang, B. (2011). Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database. PloS one *6*, e26980.

Xie, F., Xiao, P., Chen, D., Xu, L., and Zhang, B. (2012). miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant molecular biology.

Xie, F., and Zhang, B. (2010). Target-align: a tool for plant microRNA target identification. Bioinformatics *26*, 3002-3003.

Zhang, Y., Jiang, W.K., and Gao, L.Z. (2011). Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana: an update of the inverted duplication model. PloS one *6*, e28073.

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G.J., Wang, X.J., and Qi, Y. (2007). A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. Genes & development *21*, 1190-1203.

Zhou, Y., Xia, H., Li, X.J., Hu, R., Chen, Y., and Li, X.B. (2013). Overexpression of a cotton gene that encodes a putative transcription factor of AP2/EREBP family in Arabidopsis affects growth and development of transgenic plants. PloS one *8*, e78635.

Table 6-1. Different genome-derived miRNAs target different genome-derived coding genes in upland cotton.

miRNA origin	Target origin	miRNA-target pairs	Unique miRNAs	Unique targets
A	D	1385	127	1020
A	A	1303	114	1009
A	Unknown	360	97	253
AD	D	3069	221	1899
AD	A	2911	228	1748
AD	Unknown	787	186	465
D	A	3223	104	1122
D	D	3135	105	1167
D	Unknown	736	100	288
Unknown	A	72	5	71
Unknown	D	64	5	64
Unknown	Unknown	25	5	23

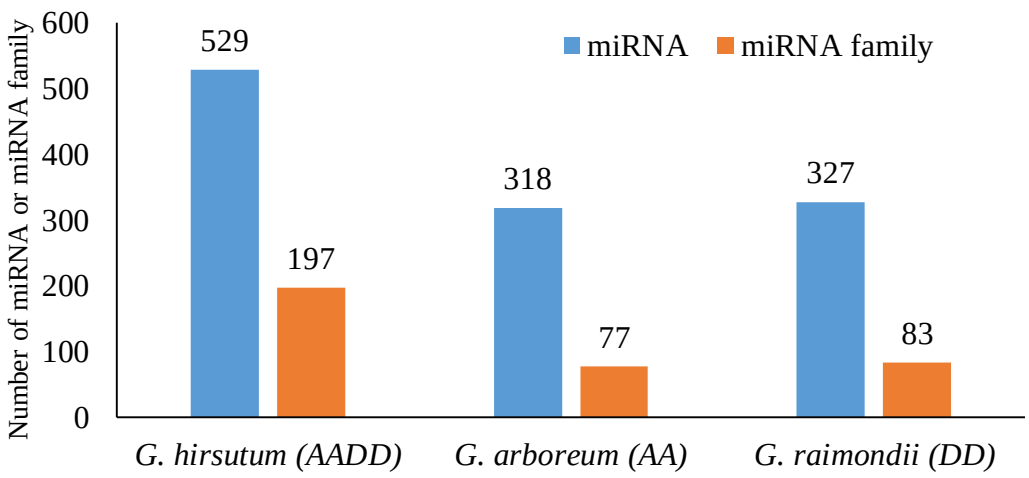


Figure 6-1. Conserved miRNAs and miRNA families identified in *G. hirsutum* (AADD), *G. arboreum* (AA), and *G. raimondii* (DD).

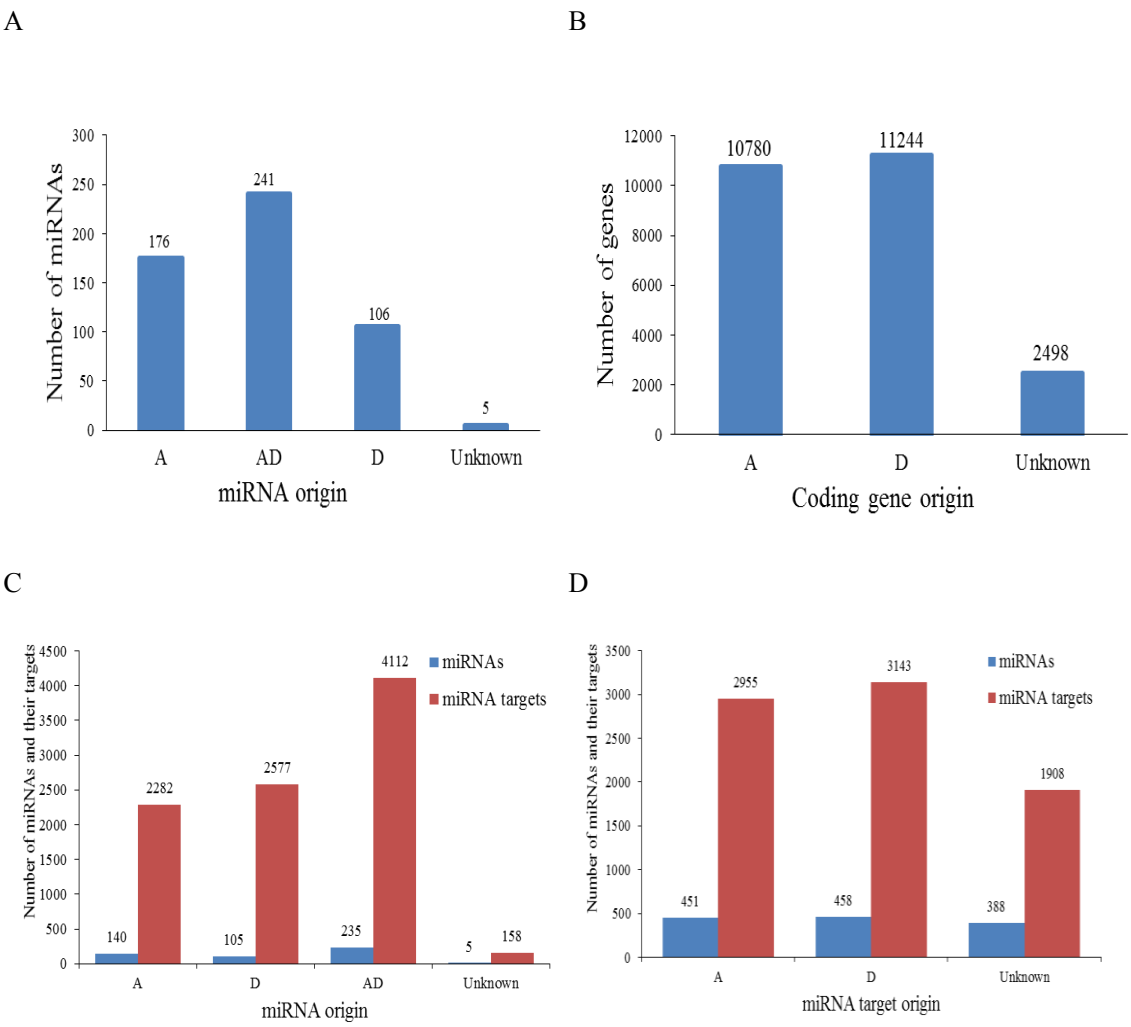


Figure 6-2. The distribution of origin of miRNAs and protein-coding genes in *G. hirsutum* (A: miRNAs and B: protein-coding gene) and the distribution of miRNAs and their targets based on miRNA origin and miRNA target origin.

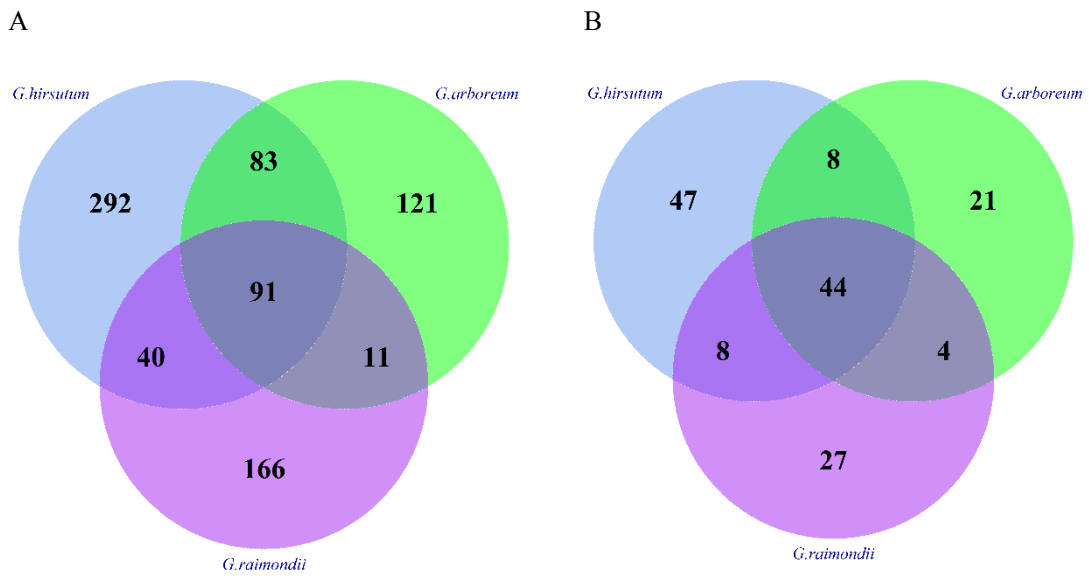
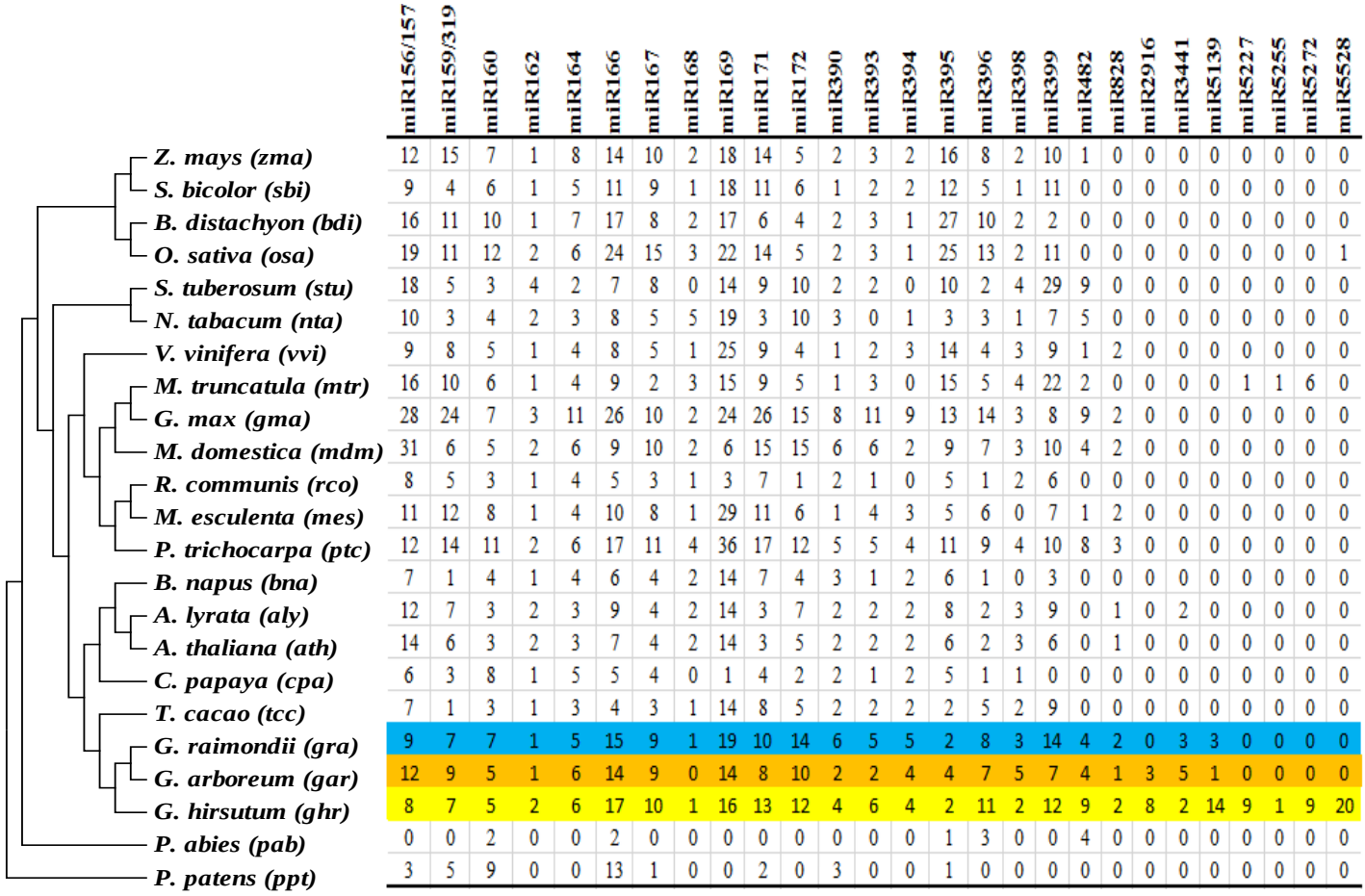


Figure 6-3. The distribution of conserved miRNAs (A) and miRNA families (B) in *G. hirsutum*, *G. arboreum*, and *G. raimondii*.



	miR156/157	miR159/319	miR160	miR162	miR164	miR166	miR167	miR168	miR169	miR171	miR172	miR390	miR393	miR394	miR395	miR396	miR398	miR399	miR482	miR828	miR2916	miR3441	miR5139	miR5227	miR5255	miR5272	miR5528
Z. mays (zma)	12	15	7	1	8	14	10	2	18	14	5	2	3	2	16	8	2	10	1	0	0	0	0	0	0	0	0
S. bicolor (sbi)	9	4	6	1	5	11	9	1	18	11	6	1	2	2	12	5	1	11	0	0	0	0	0	0	0	0	0
B. distachyon (bdi)	16	11	10	1	7	17	8	2	17	6	4	2	3	1	27	10	2	2	0	0	0	0	0	0	0	0	0
O. sativa (osa)	19	11	12	2	6	24	15	3	22	14	5	2	3	1	25	13	2	11	0	0	0	0	0	0	0	0	1
S. tuberosum (stu)	18	5	3	4	2	7	8	0	14	9	10	2	2	0	10	2	4	29	9	0	0	0	0	0	0	0	0
N. tabacum (nta)	10	3	4	2	3	8	5	5	19	3	10	3	0	1	3	3	1	7	5	0	0	0	0	0	0	0	0
V. vinifera (vvi)	9	8	5	1	4	8	5	1	25	9	4	1	2	3	14	4	3	9	1	2	0	0	0	0	0	0	0
M. truncatula (mtr)	16	10	6	1	4	9	2	3	15	9	5	1	3	0	15	5	4	22	2	0	0	0	0	1	1	6	0
G. max (gma)	28	24	7	3	11	26	10	2	24	26	15	8	11	9	13	14	3	8	9	2	0	0	0	0	0	0	0
M. domestica (mdm)	31	6	5	2	6	9	10	2	6	15	15	6	6	2	9	7	3	10	4	2	0	0	0	0	0	0	0
R. communis (rco)	8	5	3	1	4	5	3	1	3	7	1	2	1	0	5	1	2	6	0	0	0	0	0	0	0	0	0
M. esculenta (mes)	11	12	8	1	4	10	8	1	29	11	6	1	4	3	5	6	0	7	1	2	0	0	0	0	0	0	0
P. trichocarpa (ptc)	12	14	11	2	6	17	11	4	36	17	12	5	5	4	11	9	4	10	8	3	0	0	0	0	0	0	0
B. napus (bna)	7	1	4	1	4	6	4	2	14	7	4	3	1	2	6	1	0	3	0	0	0	0	0	0	0	0	0
A. lyrata (aly)	12	7	3	2	3	9	4	2	14	3	7	2	2	2	8	2	3	9	0	1	0	2	0	0	0	0	0
A. thaliana (ath)	14	6	3	2	3	7	4	2	14	3	5	2	2	2	6	2	3	6	0	1	0	0	0	0	0	0	0
C. papaya (cpa)	6	3	8	1	5	5	4	0	1	4	2	2	1	2	5	1	1	0	0	0	0	0	0	0	0	0	0
T. cacao (tcc)	7	1	3	1	3	4	3	1	14	8	5	2	2	2	2	5	2	9	0	0	0	0	0	0	0	0	0
G. raimondii (gra)	9	7	7	1	5	15	9	1	19	10	14	6	5	5	2	8	3	14	4	2	0	3	3	0	0	0	0
G. arboreum (gar)	12	9	5	1	6	14	9	0	14	8	10	2	2	4	4	7	5	7	4	1	3	5	1	0	0	0	0
G. hirsutum (ghr)	8	7	5	2	6	17	10	1	16	13	12	4	6	4	2	11	2	12	9	2	8	2	14	9	1	9	20
P. abies (pab)	0	0	2	0	0	2	0	0	0	0	0	0	0	0	1	3	0	0	4	0	0	0	0	0	0	0	0
P. patens (ppt)	3	5	9	0	0	13	1	0	0	2	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6-4. 27 representative conserved miRNA families in 23 land plants. All miRNAs of three cotton species were identified from small RNA sequencing data (see the methods). Other land land plant miRNAs are available from miRNA database, miRBase (Release 20). The number in each well stands for related miRNA family size.

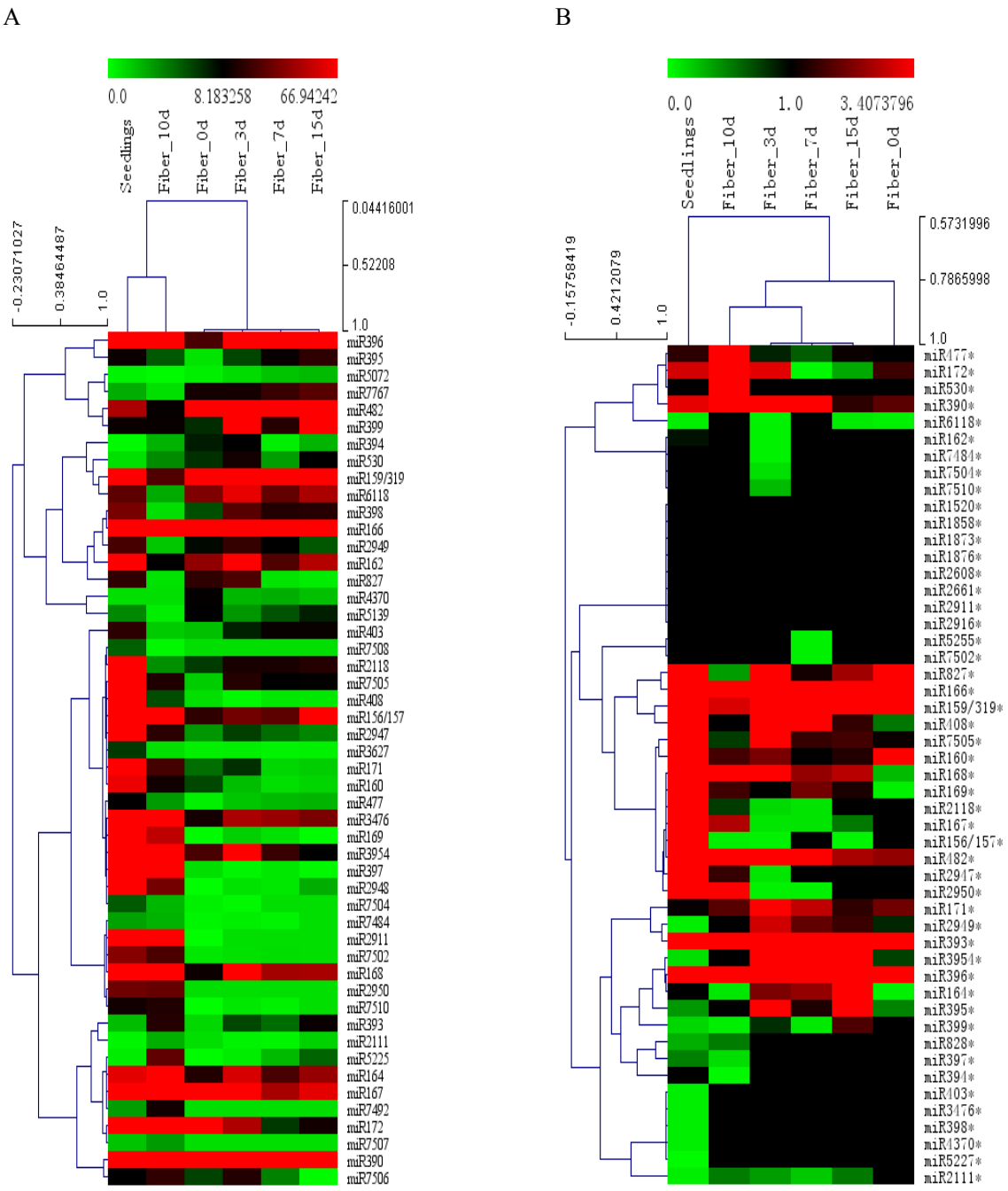
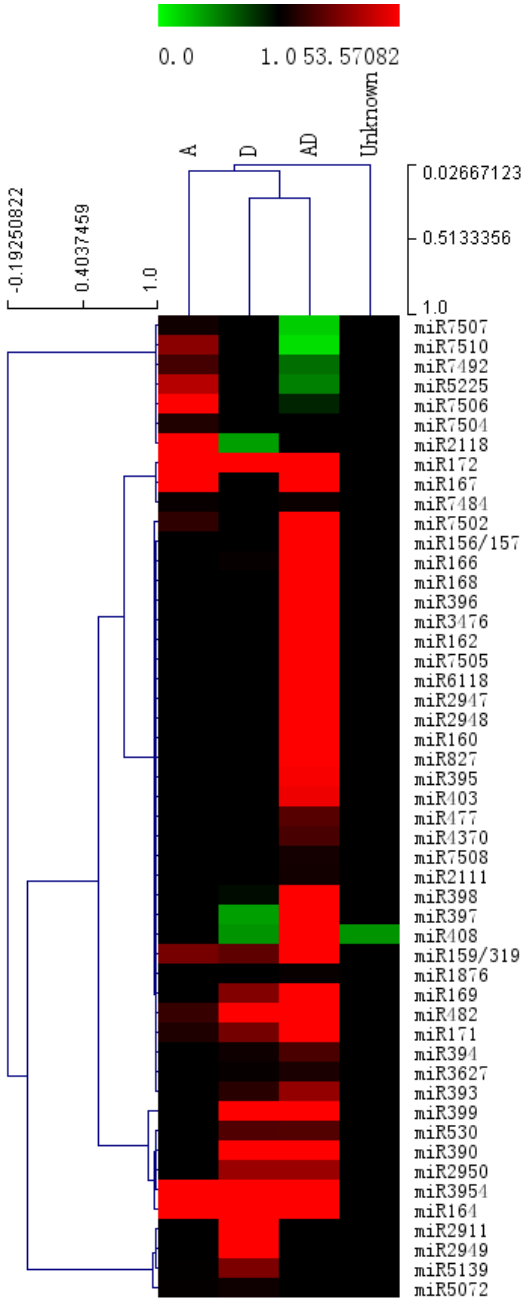
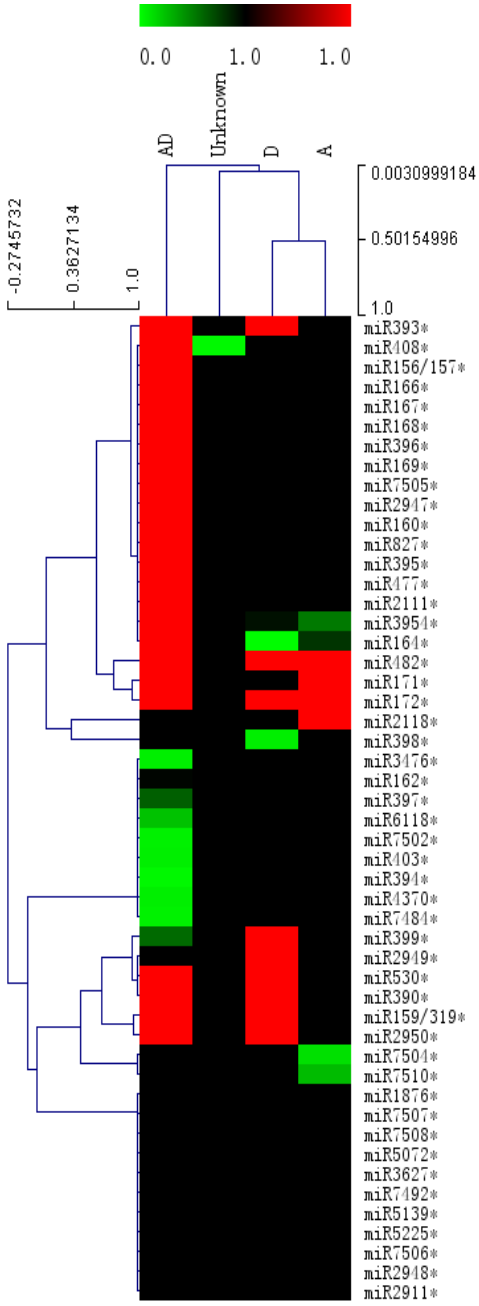


Figure 6-5. Heatmap analysis of top 50 abundant miRNA expression (A) and miRNA* expression (B).

A


0.0
1.0 53.57082
A
D
AD
Unknown
-0.19250822
0.4037459
1.0
0.02667123
0.5133356
1.0
miR7507
miR7510
miR7492
miR5225
miR7506
miR7504
miR2118
miR172
miR167
miR7484
miR7502
miR156/157
miR166
miR168
miR396
miR3476
miR162
miR7505
miR6118
miR2947
miR2948
miR160
miR827
miR395
miR403
miR477
miR4370
miR7508
miR2111
miR398
miR397
miR408
miR159/319
miR1876
miR169
miR482
miR171
miR394
miR3627
miR393
miR399
miR530
miR390
miR2950
miR3954
miR164
miR2911
miR2949
miR5139
miR5072

B


0.0
1.0
1.0
AD
Unknown
D
A
-0.2745732
0.3627134
1.0
0.0030999184
0.50154996
1.0
miR393*
miR408*
miR156/157*
miR166*
miR167*
miR168*
miR396*
miR169*
miR7505*
miR2947*
miR160*
miR827*
miR395*
miR477*
miR2111*
miR3954*
miR164*
miR482*
miR171*
miR172*
miR2118*
miR398*
miR3476*
miR162*
miR397*
miR6118*
miR7502*
miR403*
miR394*
miR4370*
miR7484*
miR399*
miR2949*
miR530*
miR390*
miR159/319*
miR2950*
miR7504*
miR7510*
miR1876*
miR7507*
miR7508*
miR5072*
miR3627*
miR7492*
miR5139*
miR5225*
miR7506*
miR2948*
miR2911*

C

D

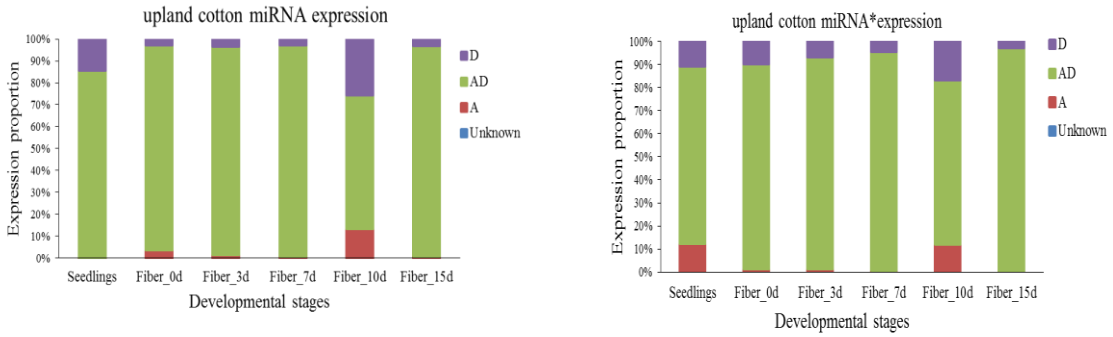


Figure 6-6. Expression analysis of miRNAs and miRNA*s of upland cotton in 6 developmental stages based on different genome derivation. **A and B:** Heatmap analysis of top 50 abundant different genome-derived miRNAs (A) and miRNA*s (B); C and D: Distribution of expression of miRNAs (C) and miRNA*s (D) in 6 developmental stages.

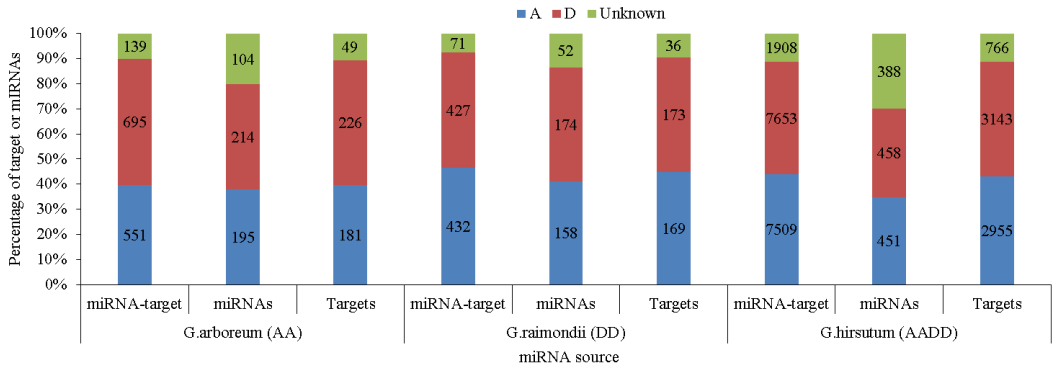


Figure 6-7. Distribution of miRNA-target pairs, miRNAs, and targets based on protein-coding genes of upland cotton that were used to predict targets with miRNAs from *G. arboreum*, *G. raimondii*, and *G. hirsutum* (A: A-derived miRNA targets, D: D-derived miRNA targets, and unknown-derived miRNA targets). Each number on the bar stands for the number of item on first X axis under different genome-derived miRNA targets.

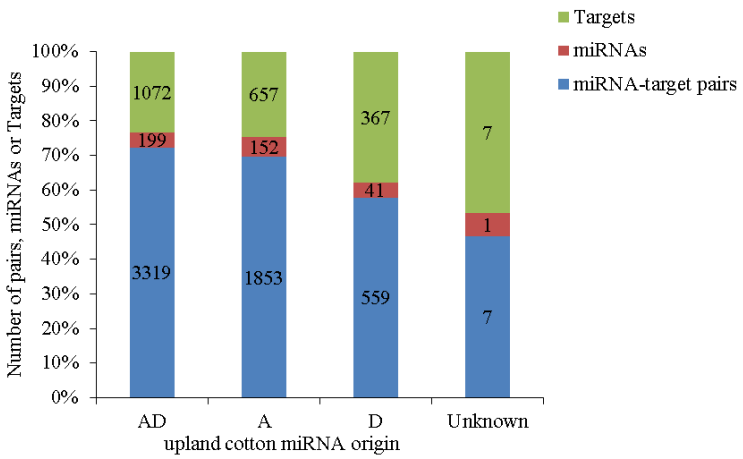


Figure 6-8. Distribution of miRNA targets predicted on coding genes of *G. raimondii* (DD) with miRNAs of upland cotton (AADD).

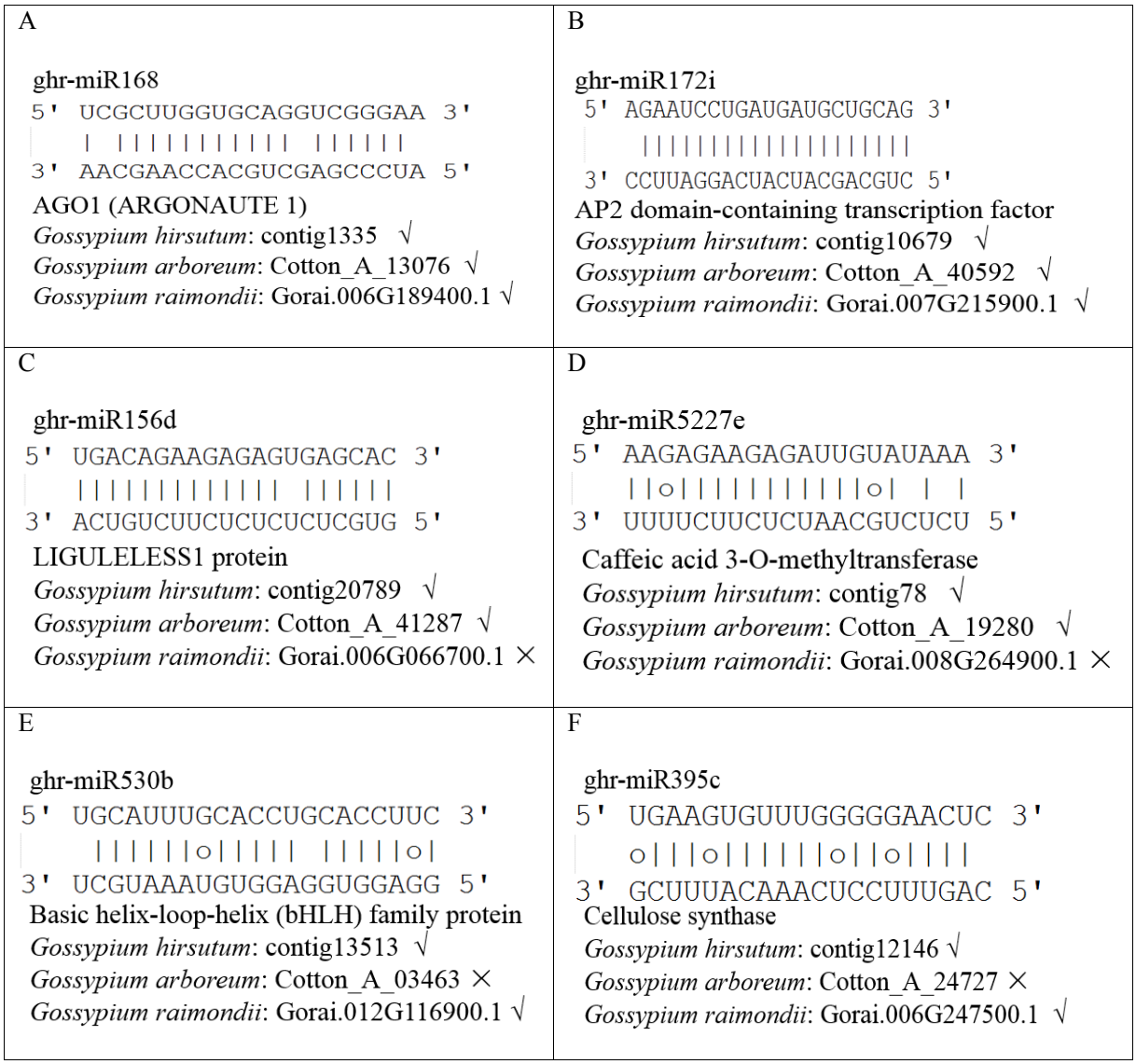
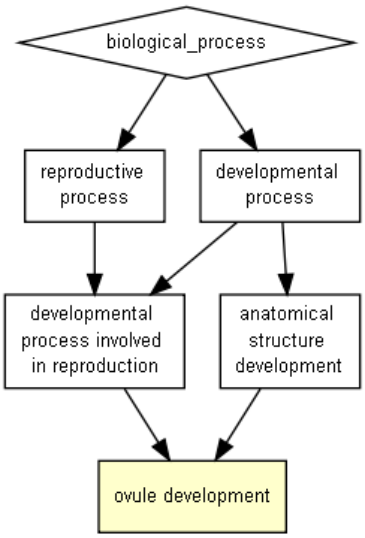
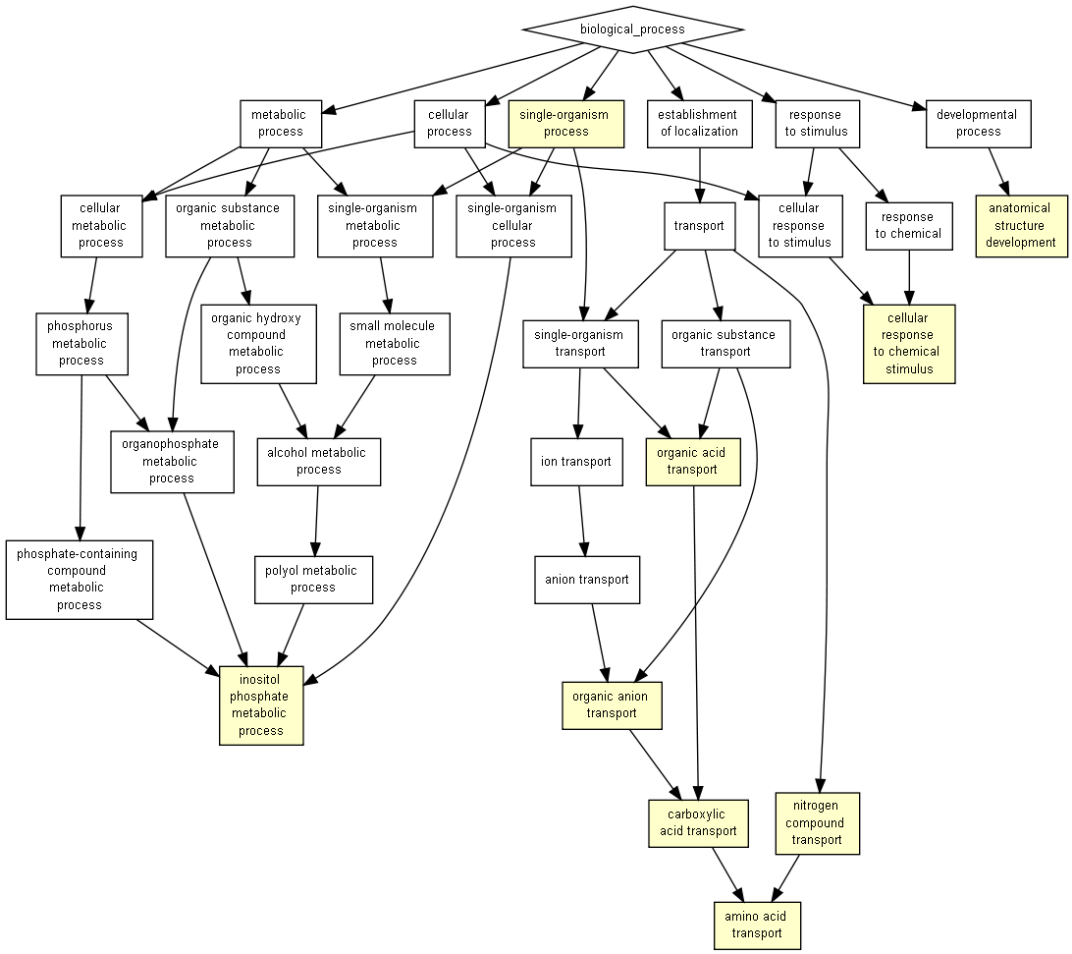


Figure 6-9. Alignment of miRNAs and their targets of *G. hirsutum*, *G. arboreum*, and *G. raimondii*. The upper sequences are miRNAs and the bottom ones are miRNA target sequences. In each item, genes from *G. arboreum* and *G. raimondii* in each item are the most homologous to the gene of *G. hirsutum*. Three genes in each items are the most homologous with each other. '√' denotes the gene has the target site on this gene. Oppositely, '×' denotes no target site no this gene.

A


biological_process
reproductive process
developmental process
developmental process involved in reproduction
anatomical structure development
ovule development

B


biological_process
metabolic process
cellular process
single-organism process
establishment of localization
response to stimulus
developmental process
cellular metabolic process
organic substance metabolic process
single-organism metabolic process
single-organism cellular process
transport
cellular response to stimulus
response to chemical
anatomical structure development
phosphorus metabolic process
organic hydroxy compound metabolic process
small molecule metabolic process
single-organism transport
organic substance transport
cellular response to chemical stimulus
organophosphate metabolic process
alcohol metabolic process
ion transport
organic acid transport
phosphate-containing compound metabolic process
polyol metabolic process
anion transport
inositol phosphate metabolic process
organic anion transport
carboxylic acid transport
nitrogen compound transport
amino acid transport

C

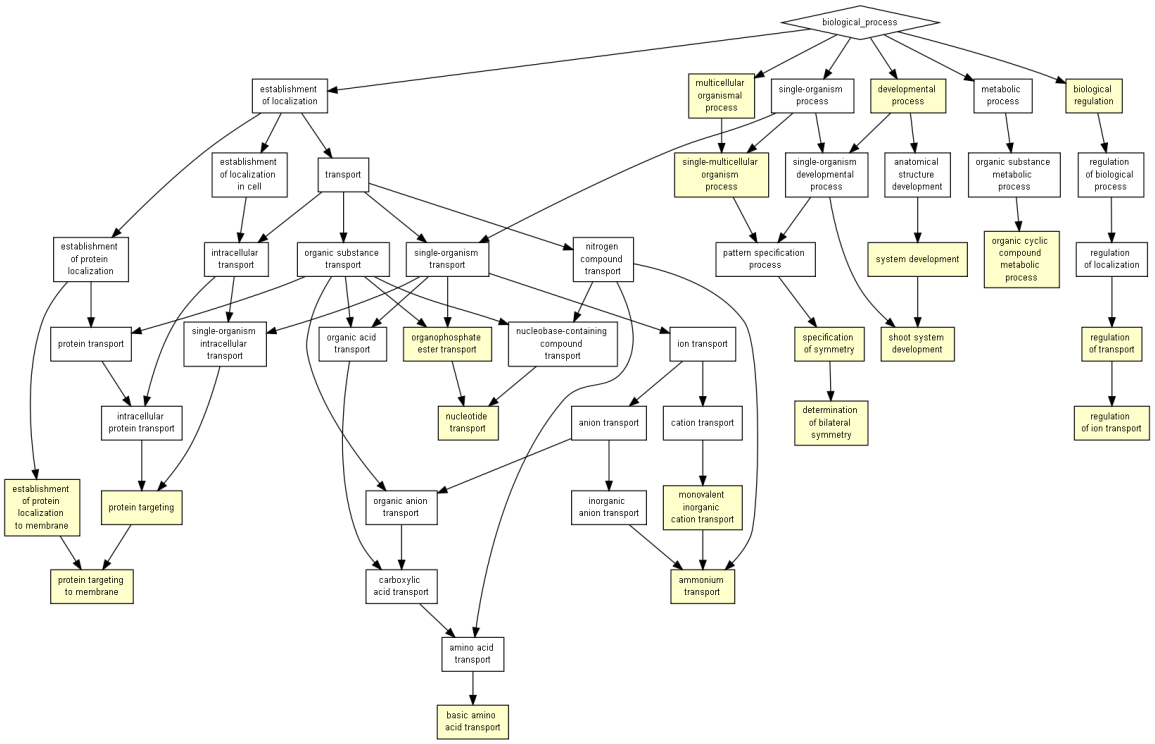


Figure 6-10. Biological process enrichment of Gorilla-based GO term analysis on different genome-derived miRNA targets (A: A-derived miRNA targets; B: D-derived miRNA targets; and C: AD-derived miRNA targets).

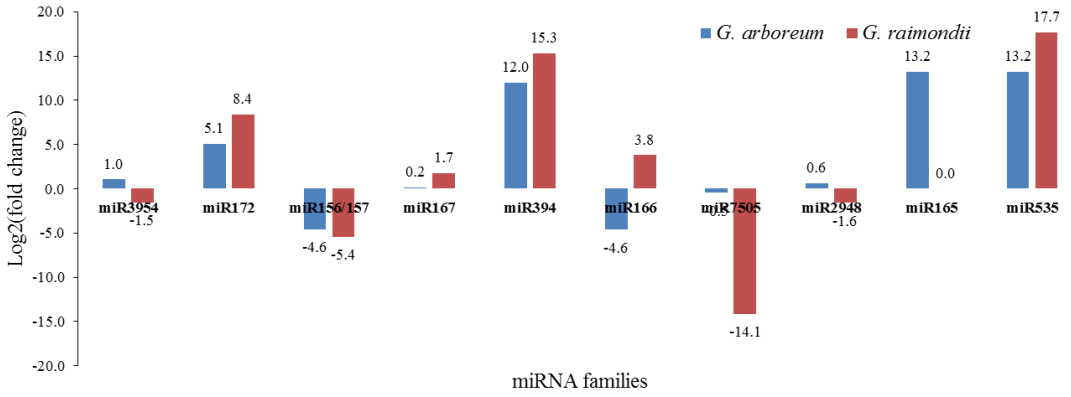


Figure 6-11. Differential expression of miRNA families in the seedlings of *G. hirsutum*, *G. arboreum*, and *G. raimondii*. Y axis denotes Log2 transformation of the expression fold changes (*G. arboreum vs. G. hirsutum* and *G. raimondii vs. G. hirsutum*).

CHAPTER 7: Global microRNA modification in cotton (*Gossypium hirsutum L.*)

Abstract

MicroRNAs (miRNAs) are small noncoding RNAs participating in versatile biological processes via post-transcriptionally gene regulation. However, how miRNAs are modified or degraded remains unknown, despites years of studies have been unraveled much details of miRNA biogenesis and function. Here, we systematically investigated miRNA modification using six small RNA sequencing libraries generated from cotton seedling as well as cotton fiber at 5 developmental stages. Our results show that 1-2-nt truncation and addition on both 5' and 3' ends of miRNAs are the major modification forms. The 5' and 3' end miRNA modification was almost equal in the 6 development stages. Truncation was more common than addition on both 5' and 3' end. Structure analysis of the 5' and 3' ends of miRNAs and isomiRs shows that uridine is the preferential nucleotide at the first position of both 5' and 3' ends. According to analysis of nucleotides truncated and tailed from miRNAs, both miRNAs and isomiRs share a similar positional structure distribution at their 5' and 3' ends, respectively. Furthermore, opposite to previous reports, cytodine is more frequently truncated and tailed from the two ends of isomiRs, implying existence of a complex cytodine balance in isomiRs. Comparison of isomiR expression shows differential miRNA modification amongst the 6 developmental stages in terms of selective modification form, development-dependent modification, and differential expression abundance. Our results globally uncovered miRNA modification features in cotton, which could

contribute us to understanding miRNA's post-mature modification and its regulatory function.

Introduction

MicroRNAs (miRNAs) are a class of short regulatory RNAs, which play critical roles in development, metabolism, and cell profiling as well as stress response in both plants and animals. miRNAs participate in post-transcriptional gene regulation through partially complementary binding with target mRNA sequences, resulting in the cleavage or translation repression of target mRNAs (Ambros, 2004; Voinnet, 2009). Most of primary miRNAs (pri-miRNAs) are initially transcribed from intergenic or intron regions on genome majorly by RNA polymerase II. In animal cells, the pri-miRNAs are first processed into a hairpin-structured precursor miRNA (pre-miRNA) by a microprocessor complex containing ribonuclease type III enzyme (Drosha) and a RNA-binding protein (DGCR8). Then pre-miRNAs are transported from nucleus to cytoplasm with the assistance of Exportin-5 and Ran-GTP (Lund et al., 2004). Cytoplasmic pre-miRNA is cleaved by another ribonuclease III enzyme (Dicer), generating an imperfect miRNA:miRNA* duplex. Mature miRNA is released from the duplex and then incorporated into a RNA-induced silencing complex (RISC) to act on its target mRNA (Ambros, 2004). However, plant miRNA:miRNA* duplex is generated in nucleus within two steps catalyzed by the plant Dicer homolog, Dicer-like 1 (DCL1).

In fact, to be a functional miRNA, plant miRNAs need to undergo a series of modifications including methylation, uridylation, adenylation, untemplated nucleotide addition, truncation, and tailing (Ji and Chen, 2012; Kim et al., 2010; Zhai et al., 2013). miRNA modification is found as an extensively existing phenomenon that could affect miRNA stability, miRNA diversity, or miRNA targeting specificity. For example, in plants, *HUA1 ENHANCER1* (*HEN1*) was the first methyltransferase identified to add a 2-O-methyl group on 3' end of miRNAs, in order to enhance miRNA stability and protect it from 3' -to- 5' truncation and 3' uridylation that finally cause miRNA degradation (Kim et al., 2010). On the contrary, *HEN1 SUPPRESSOR1* (*HESO1*) was subsequently identified as a terminal nucleotidyl transferase, which prefers 3' uridylation of miRNAs and promotes miRNA degradation (Zhao et al., 2012). A recent study offered evidence that plant miRNAs 3' truncation and tailing are AGO1-dependent, which occur either during the process of AGO1 loading or in the assembled RISC (Zhai et al., 2013). Some miRNA families show dramatic diversity in methylation, truncation and uridylation (Zhai et al., 2013). As a result of miRNA modification, many miRNAs can generate multiple variants apart from the reference sequences since there are missing or extra nucleotides at both 5' and 3' end of mature miRNAs. High-throughput sequencing uncovers that individual miRNAs genes are broadly give rise to many miRNA isoforms (isomiRs) that differ in length (Zhai et al., 2013; Zhang et al., 2013). It has been revealed that the combined effects of 5' trimming and 3' nucleotide addition can alter specificity by which miRNAs associate with different Argonatue proteins (Zhai et al., 2013). 3' nucleotide addition is proposed to be a mechanism for

regulating miRNA stability when miRNAs lose methylation after post-transcriptional modification of miRNAs (Zhang et al., 2013).

However, little is known about the global miRNA modification and their corresponded function. Therefore, in this study, we employed small RNA sequencing data of upland cotton from 6 developmental stages to systematically investigation miRNA modification on both 5' and 3' end, which includes truncation and addition. Our results showed global miRNA modification in cotton. We also revealed many unique features of cotton miRNA modification, which could allow us a new scope to study miRNA modification and function in cotton or even plant kingdom.

Results

Identification of conserved miRNAs and isomiRs

Considering potential sequencing errors might mislead to identification of false positive miRNA, only short sequences in reads of ≥ 3 were kept for searching miRNA precursors from EST, GSS, A genome, and D genome. Finally, after removing the repeated and coding precursor candidates, a total of 528 conserved miRNA precursors were obtained in upland cotton, affiliating to 197 miRNA families, such as miR156/157, miR166, miR172, miR164, miR395, miR398, miR399 (Supplementary 7-1). Based on criteria for identifying isomiRs mentioned above, 526 of 528 (99.6%) conserved miRNAs were identified to have 36,208 isomiRs per million reads per developmental stage on average (Data not shown). Of the 528 miRNAs, 21-nt miRNAs (33.1%)

accounted for the largest part, followed by those in length of 24 (17.0%) and 19 (11.7%) nt (Figure 7-1A). Coincidently, amongst 10-day-old seedlings and 5 fiber stages of cotton (see Methods), 21-nt miRNAs were also the most abundant in the length frequency distribution of these miRNAs (Figure 7-1B). Interestingly, length distribution and nucleotide frequency distribution of miRNAs were some different from those of isomiRs including 5' truncation, 5' addition, 3' truncation, and 3' addition. Most of isomiRs in seedling and 10-DPA fiber were in length of 22 nt, whereas 21-nt isomiRs were the largest part in the 0-, 3-, 7-, and 15-DPA fibers, respectively (Figure 7-1C). To look at which one in 4 types of isomiR variants predominates in isomiR length distribution, we then investigated the expression distribution of 4 sorts of isomiR variants. It turned out the expression of addition and truncation of the 5' end were generally similar to that of addition and truncation of the 3' end in 0-, 3-, 7-, and 15-DPA fibers (Figure 7-2). However, there were some different expression distribution in seedling and 10-DPA fiber, 5' addition and 3' truncation represent the most sufficient part in seedlings and 10-DPA fiber, respectively (Figure 7-2). The 5' addition in seedling might be the main contributor that enables isomiRs 1-nt longer than miRNAs. In addition, during the 5 fiber stages, both of the 5' and 3' truncations were higher than the 5' and 3' additions, correspondingly (Figure 7-2).

Nucleotides structure of the 5' and 3' ends of miRNAs and isomiRs

As far as current understanding on miRNAs, plant mature miRNAs are protected from degradation by methylation on the 3' end of terminal nucleotides of miRNAs (Guo and Lu, 2010; Wu et al., 2013). There is a plausible fact that 3' adenylation and uridylation of miRNAs promote miRNAs stability and degradation, respectively (Adams and Wendel, 2005; Comai et al., 2000; Lu et al., 2009). However, the idea is not supported by the finding that both 3' adenylation and uridylation miRNA were proposed to favor degradation of miRNAs in Drosophila and human (Wang et al., 2006). Although the mechanism is still unclear, we believe there might be certain balance between degradation and stability of miRNAs in terms of miRNA terminal modification including adenylation and uridylation. As a result, the balance would enable terminal nucleotides of miRNAs with a preferential nucleotide composition that protects miRNAs from degradation. Likewise, as miRNAs, isomiRs might be also acted by a similar balance. To this end, we first investigated the positional structure distribution of start and end nucleotides of miRNAs and isomiRs. As shown in the Figure 3, the first nucleotides of 5' end of miRNAs were preferential to be “U”, accounting for more than 90% in seedling and 5 fiber stages (0 -15 DPA) (Figure 7-3A). Except in seedling, the secondary nucleotides of 5' end of miRNAs in the 5 fiber stages were mainly “C”. “G” was the richest one for both the third and fourth nucleotides of the 5' end of miRNAs in all of the 6 developmental stages (Figure 7-3A). On the 3' end of miRNAs, only miRNAs in seedling mainly ended with “U”, whereas “C” was the largest part at the first nucleotide in the 5 fiber stages (Figure 7-3B). Interestingly, all of six stages reached a consensus on the 3' secondary nucleotides that

"U" was the absolutely predominant nucleotide (Figure 7-3B). From 3rd to 5th nucleotides, "U" was generally dominant one in the 6 developmental stages, except the one that the 3'-end fourth nucleotide of miRNAs in seedling were mainly "A". Overall, terminal nucleotides of miRNAs displayed a high heterogeneity in the six stages, despites some tended to be similar.

For isomiRs, the nucleotides on the 5' end shared a similar composition pattern (Figure 7-3C) with those of miRNAs. The first nucleotide was predominant by "U", and the 5' secondary nucleotide was mainly "C" in 0-15 DPA fibers rather than "U" in seedling (Figure 7-3C). The major nucleotides on the third and fourth position at 5' end of isomiRs were also "G", whereas "A" was main nucleotide on both the third and fourth positions at the 5' end in the 10-DPA fiber. Similarly, the 3' end nucleotides were generally similar to those of miRNAs (Figure 7-3D). However, the proportion of "U" at the fourth nucleotide of the 3' end of isomiRs was significantly higher than that of miRNAs. Moreover, much difference of the 3' end of isomiRs was found at the 1st, 3rd, and 5th nucleotides in 10-DPA fiber, in comparison to those of miRNAs (Figure 7-3D).

Structure of truncated and added nucleotides on the 5' and 3' ends of miRNAs

In analysis of terminal nucleotide structure of the 5' and 3' ends of miRNAs and isomiRs, certain nucleotide bias was observed to display rich diversity amongst seedling and 5 different fiber stages. We then asked whether there is also nucleotide bias on the 5' and 3' ends of miRNAs, which is prone to be truncated or tailed by related

enzymes for small RNA modification (Hegarty et al., 2006; Ramachandran and Chen, 2008; Zhai et al., 2013). To this point, we analyzed the first 5 nucleotides that are truncated or tailed on both 5' (counting from the most 5' end) and 3' (counting from the most 3' end) ends of isomiRs (Figure 7-4). Our results show that a similar positional structure distribution of truncated nucleotides existed on the 5' and 3' ends of miRNAs (Figure 7-4A and 7-4B). Overall, "C" was more frequently truncated from the 5' and 3' ends of miRNAs than other nucleotides in the 6 development stages. Contrarily, in the first 5 truncated nucleotides, the nucleotide on the secondary position strikingly tended to "U" in all of the 6 developmental stages. In addition, there was also some exception amongst these stages. For instance, in the 10-DPA fiber, the third position of truncated nucleotides was likely to be "G" and "A" on the 5' and 3' end of miRNAs, respectively. Different from other 5 developmental stages, the first nucleotide truncated from the 3' end of miRNAs in 10-DPA fiber was mainly "G" and then "A". Also, its general ratios of "A" and "G" of the 5' truncated nucleotides appeared a peak when compared with the other 4 fiber developmental stages (Figure 7-4A).

In the truncated nucleotides on miRNAs, "C" was more frequently added to the 3' end of miRNAs (Figure 7-4D). But the secondary position of tailed nucleotide on 3' end of miRNAs is more preferential to be "U". "G" is the largest one that was added to the secondary position of the 3' added nucleotides only in 10-DPA fiber (Figure 7-4D). Compared with the nucleotides added to 3' end of miRNAs, both "U" and "C" seemed to be more often tailed to the 5' end of miRNAs (Figure 7-4B). However, the first two nucleotides on both 5 and 3 added nucleotides in seedling were predominant to be "U"

(Figure 7-4B). Also, “G” was the most abundant one on the third position of the 5' added nucleotides of miRNAs in seedling, whereas the counterpart in the 0-15 DPA fiber were mainly “U” (Figure 7-4B).

Length and frequency distributions of truncated and tailed nucleotides on the 5' and 3' ends of miRNAs

We further investigated the length and frequency distributions of nucleotides that were truncated and tailed to the 5' and 3' ends of miRNAs, correspondingly. generally speaking, 1- or 2-nucleotide tailing and trimming on the 5' and 3' ends of miRNAs accounted for the largest part in the length variation of terminally modified miRNAs amongst all of the 6 developmental stages (Figure 7-5A/C and Figure 7-6A/C), while “U” and “C” mainly consisted of the terminal nucleotides truncated and added to the 5' and 3' ends of miRNAs in upland cotton (Figure 7-5B/D and Figure 7-6B/D). For the 5' truncated nucleotides in seedling, the 1-nt and 2-nt truncation almost were in the same proportion (Figure 7-5A). But for the 5'-end tailed nucleotides in seedling, 1-nt addition made up the largest part (Figure 7-5C). Nevertheless, except 10-DPA fiber, the ratio of 2-nt modification to 1-nt modification in both 5' truncated and added nucleotides were largely similar in the other 4 fiber stages (Figure 7-5A and 7-5C). The proportions of 1-nt truncation and addition on 5' end of miRNAs in 10-DPA fiber were apparently raised in comparison to that in other 4 fiber developmental stages (Figure 7-5A and 7-5C). Likewise, we also observed that the ratio of “U” to “C” in both 5'

truncated and added nucleotides of miRNAs was development-specific. Amongst 0-15 DPA fibers, the ratios for 5' truncation and addition were around 1, whereas they were about 5 on average in seedling (Figure 7-5B and 7-5D).

In the first 5 nucleotides that truncated and added to the 3' end of a miRNA, "C" was the most abundant nucleotide amongst the 6 different developmental stages, followed by "A" and "U". In the 5 fiber developmental stages, there were "A" and "G" peaks of the 3' truncated and added nucleotides at 10-DPA fiber (Figure 7-6B and 7-6D). Interestingly, the expression alternation between the 3' truncated and added miRNAs in seedling was significantly opposite to that in the 3-DPA fiber (Figure 7-6B and 7-6D). The majority of 3' added miRNAs were expressed lower than the 3' truncated miRNAs in the 5 different fiber stages.

Differential modification in cotton conserved miRNAs

Besides global divergence of miRNA modification, many conserved miRNAs showed differential modification in terms of truncation and addition on the 3' and 5' ends of miRNAs. First of all, some miRNAs might be selectively modified by one or more forms of truncation and addition. For instance, only 3' truncation, 3' addition, and 5' truncation were detected in the isomiRs of ghr-miR157a amongst the 6 different developmental stages (Figure 7-7A, 7-7C, and 7-7E). ghr-miR172i were merely 3' truncated (Figure 7-8A). Second, miRNA modification is dependent on development stage. The 3' and 5' truncation of ghr-miR157a was only observed in the young seedling,

0-DPA fiber, 10-DPA fiber, and 15-DPA fiber, whereas the 3' addition of ghr-miR157a were found in all of the 6 developmental stages (Figure 7-7A, 7-7C, and 7-7E). The 3' truncation of ghr-miR172i was observed in all of the 5 fiber stage but not in the young seedling (Figure 7-8A). Third, miRNAs modification was also differentially expressed in different developmental stages. The 3' truncation, 3' addition, and 5' truncation of ghr-miR157a existed much richer in seedling relative to the 5 different fiber stages (0-15 DPA) (Figure 7-7B, 7-7D, and 7-7F). However, the 3' truncation of ghr-miR172i was commonly observed in the 5 fiber stages, and experienced a sharp fluctuation at 7-DPA fiber (Figure 7-8B). Interestingly, for many miRNAs, their expression was positively correlated with the expression of their isomiRs in a significant level (Figure 7-7B, 7-7D, and 7-7F). In addition to differential modification on isomiRs, as shown in Figure 7-8(B, D, F) and Figure 7-9A, some modified nucleotide forms were observed to be highly conserved in different developmental stages indicating some miRNA modification are not a random event. We infer that conserved isomiRs might be a common consequence of miRNA post-mature modification or might be necessary to get involved in an unknown regulation.

Discussion

Terminal "U" protects miRNAs and isomiRs from degradation?

Based on the 6 small RNA sequencing data of upland cotton at the different developmental stages, we were able to estimate that mean miRNA and isomiR reads

were 53,127 and 36,208 reads per million (RPM) in a developmental stage, respectively. Accordingly to those numbers, miRNAs are theoretically richer and more stable than their isomiRs. According to analysis of positional structure distribution of cotton miRNA nuceotides, both of the first nucleotides on the 5' and 3' end of miRNAs were prone to start or end with "U", respectively. However, compared with miRNAs, the 3' end of isomiRs was preferential to be "C", despites its 5' end also mainly started with "U". It was reported that 3' adenlylaiton increases stability and 3' uridylation enhances miRNAs degradation in plants and animals (Adams and Wendel, 2005; Comai et al., 2000; Lu et al., 2009), despites of some exceptional cases that both 3' uridylation and adenlyation were proposed to promote miRNA degradation in *Drosophila* and humans (Wang et al., 2006). Moreover, HUA1 ENHANCER1 (HEN1), is a methyltransferase that functions in adding a 2-O-methyl group to the 3'-most terminal nucleotide of miRNAs and small interfering RNAs (Hegarty et al., 2006; Qin and Zhu, 2011). miRNAs were found to be heavily truncated and in a high uridylation level in *hen1* mutant (functional loss) in Arabidopsis. Most truncation of miRNAs was observed prior to uridylation (Zhai et al., 2013). According to global detection of the first nucleotides on two ends of plant and animal known miRNAs, we also found U is the predominant nucleotide on the 5' and 3' ends of miRNAs (Figure 7-9). T-test showed that the 5' end of miRNAs significantly starts with "U" in both kingdoms. Our observation on positional structure of cotton miRNAs and isomiRs was also consistent with the finding in plant and animal known miRNAs. Therefore, we infer that uridylation might be a consequence of modification to miRNA or isomiRs probably in order to avoid

degradation. Furthermore, uridylation is not only important for the 3' end of miRNAs and isomiRs, but also critical for their 5' ends. The inference is different from previous report that the 3'-end adenylation of miRNAs might stabilize the 3' end (Lu et al., 2009), since no significant adenylation to the 3' end of miRNAs and isomiRs was observed in either one of our 6 cotton developmental stages. Combining the distribution of the 3' end nucleotide and the stability between miRNAs and isomiRs, we might also make further inference that the 3' end is weaker than the 5' end in small RNAs and small RNA degradation-related enzymes more readily act on small RNAs from 3'-5'. The subsequent uridylation in *hen1* mutant might compromise miRNAs or isomiRs from heavy truncation in absence of small RNA-specific methyltransferase. Amongst the 6 developmental stages, we found 10-DPA fiber displayed the most different terminal nucleotides on 3' end of miRNAs and isomiRs, implicating an important development landmark where the activities of small RNA degradation or modification-related enzymes are dramatically adjusted. However, these inferences are needed to be supported by further experiment evidence, like transcritpome comparison between 10-DPA fiber and other fiber stages and gene knockout validation to key small RNA degradation or modification-related enzymes.

Cytidine balance between the truncation and addition of the two ends of isomiRs?

Zhang and its co-workers found “C” addition to the 3' end of a miRNA extensively exists in plant kingdom (Zhang et al., 2013). We also found “C” is generally

added to the 3' end of miRNAs in cotton (Figure 7-4D). Interestingly, “C” addition is not preferentially added to the secondary position of the 3' end of miRNAs, but rather “U”. “U-C” (isomiR-UC) is the most popular form at the 3' end of isomiRs in cotton. However, previous reports showed that continuous “A”, “U”, and their combinations are the most abundant nucleotide forms added to the 3' end of animal miRNAs (Ji and Chen, 2012; Zhou et al., 2012). Zhang and its co-workers discovered “A+U+X” combination is main form in the isomiRs of 22 conserved miRNAs in larch (Zhang et al., 2013). In addition to the 3' end of cotton isomiRs, “U” and “C” were also more frequently added to the 5' end of cotton isomiRs (Figure 7-4B). Surprisingly, we found that “C” is also frequently observed in the truncated nucleotides on both 5' and 3' ends of cotton miRNAs (Figure 7-4A and 7-4C), indicating that there might be a cytidine balance between truncation and addition of two ends of isomiRs in cotton. Furthermore, cytidine balance might also exist in the truncation and addition of the 3' end of miRNAs (Figure 7-6B and 7-6D). Adding and trimming cytidine might have special sense in maintaining integrity and stability of isomiRs. Besides cytidine balance, uridine balance is also observed in truncation and addition of 5' end of miRNAs (Figure 7-5B and 7-5D).

Conclusion

To date, understanding on miRNA modification remains poor. Utilizing small RNA sequencing data of upland cotton in the 6 different developmental stages, we

systematically analyzed truncation and addition of the 5' and 3' ends of miRNA in cotton. Opposite to the previous finding that 3'-end modification is higher or important than that of 5'-end, in general, both two forms equally exist in cotton, indicating both of them are also equally important for miRNA modification. “U” preference was found on the 3' and 5' end of miRNAs and isomiRs in cotton, as well in known miRNAs of plants and animals. In addition, complex cytodine balance was also observed, contrary to previous understanding that cytodine addition is preferential on the 3' end of miRNAs. Our findings might offer an idea in cotton on how to evaluate stability of a small RNA or design an effective artificial small RNA in cotton regarding the possible small RNA modification. In addition, differential miRNA modification amongst the 6 different developmental stages might imply inner environment change in cotton that might cause differential activities of small RNA modification-related enzymes. Together, our study provided a new insight into understating miRNA modification and related regulation.

Materials and Methods

Material preparation and Small RNA sequencing

The seeds of *Gossypium hirsutum* cultivar TM-1 were sterilized and cultivated in 1/2 Murashige and Skoog (MS) medium (pH 5.8) under a 16 h light/8 h dark cycle at room temperature (22 ±1℃). Ten-day-old seedlings from 5 biological replicates were harvested and immediately frozen in liquid nitrogen. Total RNAs was extracted from each tissue sample using the mirVana miRNA isolation kit (Ambion, Austin, TX)

according to the manufacturer's protocol. RNAs' quantity and quality were assayed by Nanodrop ND-1000 (Nanodrop technologies, Wilmington, DE, USA). RNA sample were submitted to BGI (Shenzhen, China) for small RNA sequencing.

miRNA identification

Besides the small RNA sequencing data of seedling mentioned above, 5 other small RNA sequencing data of upland cotton were downloaded from NCBI (GSM911189: 0-DPA fiber, GSM911190: 3-DPA fiber, GSM911191: 7-DPA fiber, GSM686015: 10-DPA fiber, and GSM911192: 15-DPA fiber). Finally, using our developed software miRDeepFinder and previously reported pipeline (Xie et al., 2014; Xie et al., 2012), 6 small RNA sequencing data of upland cotton were combined together to identify conserved miRNAs from EST and GSS of upland cotton, A genome (whole-genome shotgun sequences of *G. arboreum* from NCBI (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AYOE01#contigs), and D genome of *G. raimondii* from JGI (http://www.phytozome.net/cotton.php).

Identification of miRNA isoforms (isomiRs)

IsomiRs were identified from the 6 small RNA libraries of upland cotton according to the methods described in the previous reports (Zhai et al., 2013; Zhang et al., 2013). Briefly, short sequences are retrieved and considered as isomiRs if: 1) it matches perfectly to the mature miRNA sequence and trims one or more nucleotides at

the 5' or 3' end; 2) it maps perfectly to the mature miRNA sequence and have one or more mononucleotide stretches at their 3' end that do not overlap to the precursor sequence. Nucleotide additions are classified to three types, ‘unambiguous’, ‘ambiguous’, and ‘mixed’. For ‘unambiguous’, if 3' end extra nucleotides of isomiRs cannot map back to the miRNA precursor, they are likely added after maturation. Otherwise, the addition belongs to ‘ambiguous’. If an ambiguous addition is followed by an unambiguous addition, it is sorted as ‘mixed’. The type, order, and frequency distribution of truncated and added nucleotides in the 6 developmental stages were recorded for further comparison analysis, respectively.

Statistical analysis

Expression of miRNAs and isomiRs were normalized to reads per million reads (RPM). The correlation analysis and T-test were conducted with SPSS software.

Supporting Information

Supplementary 7-1. Cotton miRNA expression in 6 different developmental stages

Reference

Adams, K.L., and Wendel, J.F. (2005). Novel patterns of gene expression in polyploid plants. Trends in genetics : TIG *21*, 539-543.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350-355.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y., and Byers, B. (2000). Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. The Plant cell *12*, 1551-1568.

Guo, L., and Lu, Z. (2010). The fate of miRNA* strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule? PloS one *5*, e11387.

Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2006). Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. Current biology : CB *16*, 1652-1659.

Ji, L., and Chen, X. (2012). Regulation of small RNA stability: methylation and beyond. Cell research *22*, 624-636.

Kim, Y.K., Heo, I., and Kim, V.N. (2010). Modifications of small RNAs and their associated proteins. Cell *143*, 703-709.

Lu, S., Sun, Y.H., and Chiang, V.L. (2009). Adenylation of plant miRNAs. Nucleic acids research *37*, 1878-1885.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95-98.

Qin, Y.M., and Zhu, Y.X. (2011). How cotton fibers elongate: a tale of linear cell-growth mode. Current opinion in plant biology *14*, 106-111.

Ramachandran, V., and Chen, X. (2008). Degradation of microRNAs by a family of exoribonucleases in Arabidopsis. Science *321*, 1490-1492.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. Cell *136*, 669-687.

Wang, J., Tian, L., Lee, H.S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L.*, et al.* (2006). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics *172*, 507-517.

Wu, X., Bhayani, M.K., Dodge, C.T., Nicoloso, M.S., Chen, Y., Yan, X., Adachi, M., Thomas, L., Galer, C.E., Jiffar, T.*, et al.* (2013). Coordinated targeting of the EGFR signaling axis by microRNA-27a*. Oncotarget *4*, 1388-1398.

Xie, F., Stewart, C.N., Jr., Taki, F.A., He, Q., Liu, H., and Zhang, B. (2014). High-throughput deep sequencing shows that microRNAs play important roles in switchgrass responses to drought and salinity stress. Plant biotechnology journal *12*, 354-366.

Xie, F., Xiao, P., Chen, D., Xu, L., and Zhang, B. (2012). miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant molecular biology.

Zhai, J., Zhao, Y., Simon, S.A., Huang, S., Petsch, K., Arikit, S., Pillay, M., Ji, L., Xie,

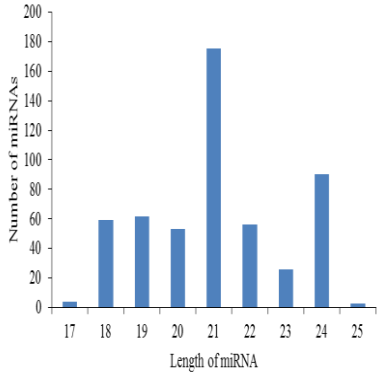
M., Cao, X., *et al.* (2013). Plant microRNAs display differential 3' truncation and tailing modifications that are ARGONAUTE1 dependent and conserved across species. The Plant cell *25*, 2417-2428.

Zhang, J., Zhang, S., Li, S., Han, S., Wu, T., Li, X., and Qi, L. (2013). A genome-wide survey of microRNA truncation and 3' nucleotide addition events in larch (Larix leptolepis). Planta *237*, 1047-1056.

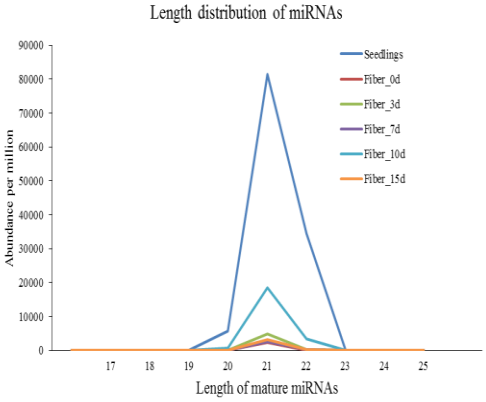
Zhao, Y., Yu, Y., Zhai, J., Ramachandran, V., Dinh, T.T., Meyers, B.C., Mo, B., and Chen, X. (2012). The Arabidopsis nucleotidyl transferase HESO1 uridylates unmethylated small RNAs to trigger their degradation. Current biology : CB *22*, 689-694.

Zhou, H., Arcila, M.L., Li, Z., Lee, E.J., Henzler, C., Liu, J., Rana, T.M., and Kosik, K.S. (2012). Deep annotation of mouse iso-miR and iso-moR variation. Nucleic acids research *40*, 5864-5875.

A



B



C

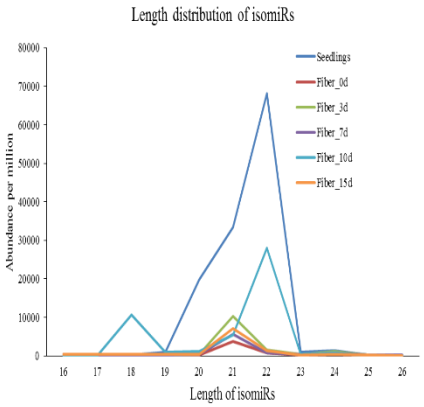


Figure 7-1. Length distribution of miRNAs and isomiRs in cotton seedlings and fiber at 5 developmental stages (Fiber_0d: 0-DPA fiber; Fiber_3d: 3-DPA fiber; Fiber_7d: 7-DPA fiber; Fiber_10d: 10-DPA fiber; and Fiber_15d: 15-DPA fiber). A: Length distribution of unique miRNAs; B: Length frequency distribution of miRNAs; and C: Length frequency distribution of isomiRs.

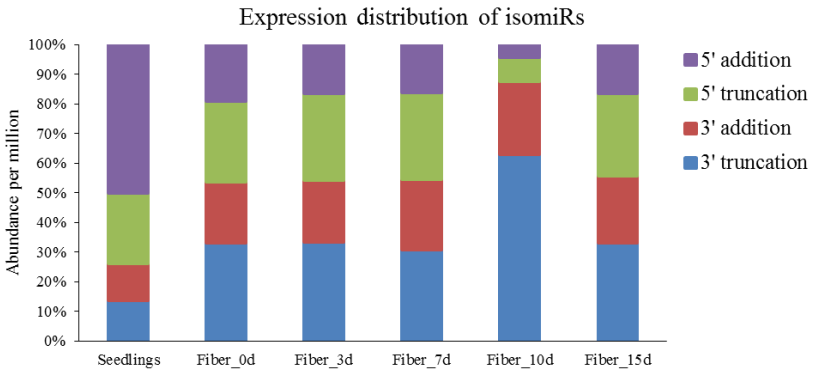
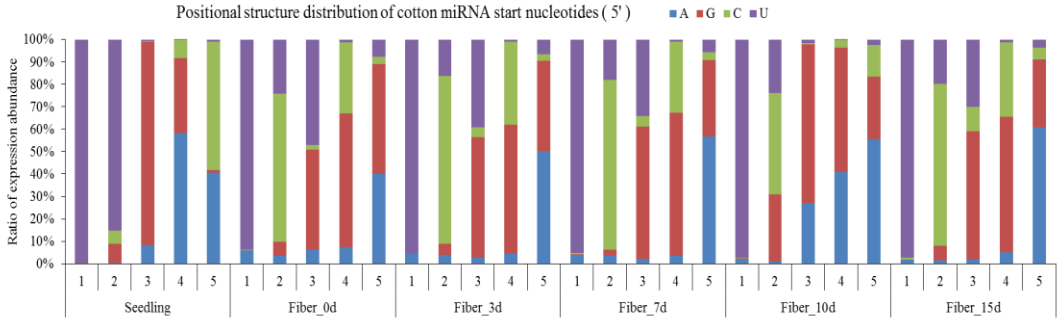
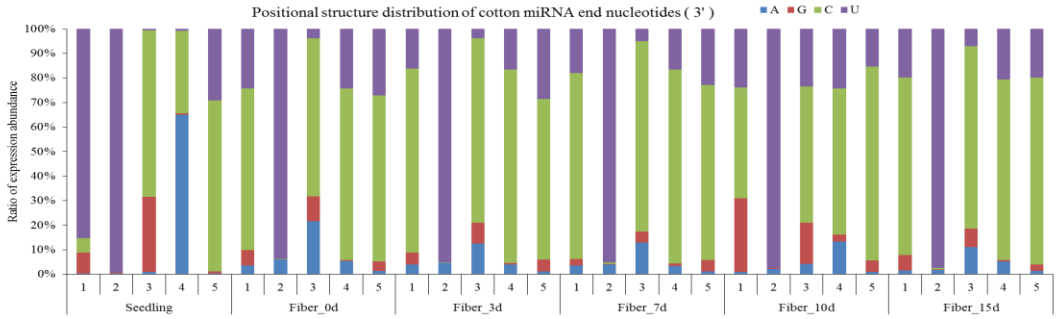


Figure 7-2. Expression distribution of cotton isomiRs including 5' addition, 5' truncation, 3' addition, and 3' truncation.

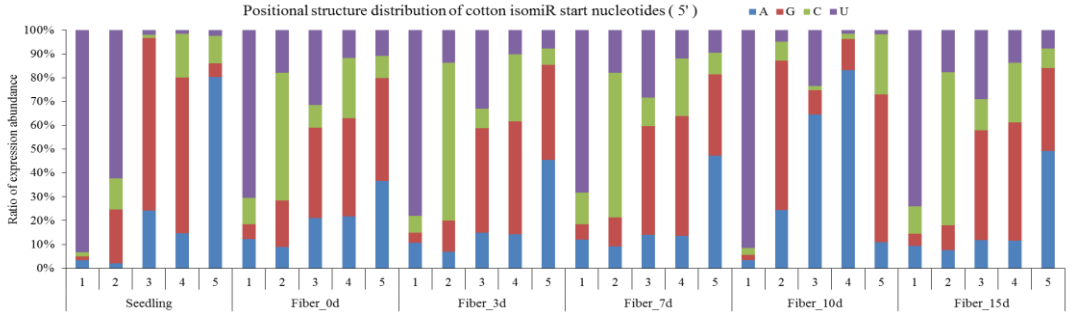
A


Positional structure distribution of cotton miRNA start nucleotides (5')
A G C U
Ratio of expression abundance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5
Seedling
Fiber_0d
Fiber_3d
Fiber_7d
Fiber_10d
Fiber_15d

B


Positional structure distribution of cotton miRNA end nucleotides (3')
A G C U
Ratio of expression abundance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5
Seedling
Fiber_0d
Fiber_3d
Fiber_7d
Fiber_10d
Fiber_15d

C


Positional structure distribution of cotton isomiR start nucleotides (5')
A G C U
Ratio of expression abundance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5
Seedling
Fiber_0d
Fiber_3d
Fiber_7d
Fiber_10d
Fiber_15d

D

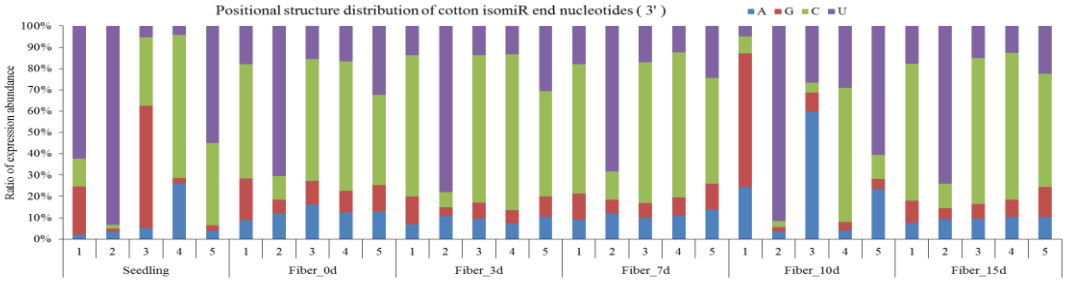
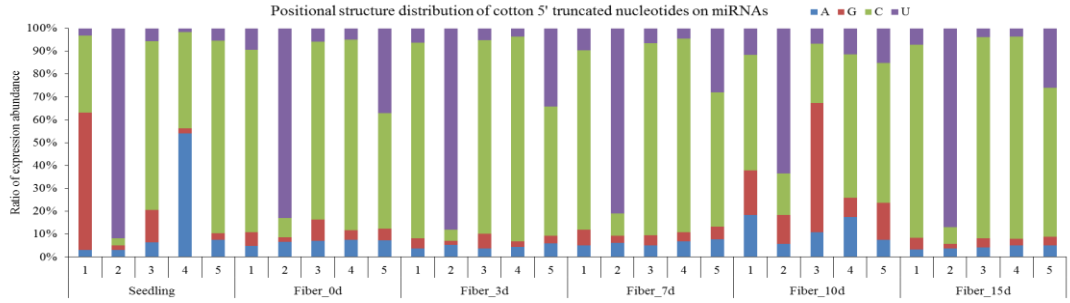
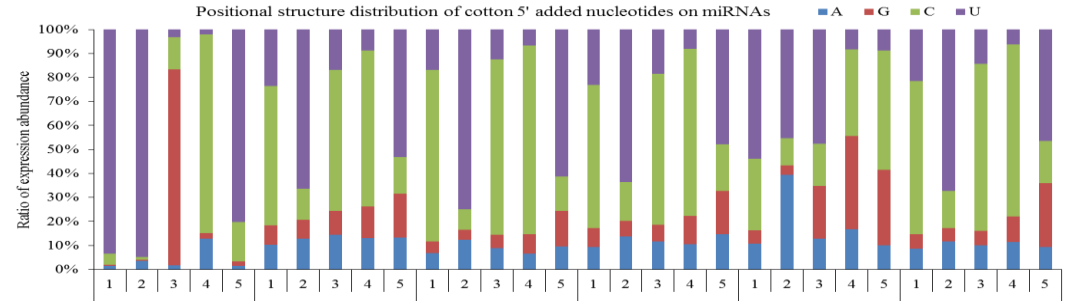


Figure 7-3. Positional structure distribution of the start (5') and end (3') nucleotides of miRNAs and isomiRs in cotton (A: start (5') nucleotides of miRNAs; B: end (3') nucleotides of miRNAs; C: start (5') nucleotides of isomiRs; B: end (3') nucleotides of isomiRs). 1-5 on X axis denote the 1st-5th nucleotide starting from the most 5' end or 3' end of miRNA and isomiRs.

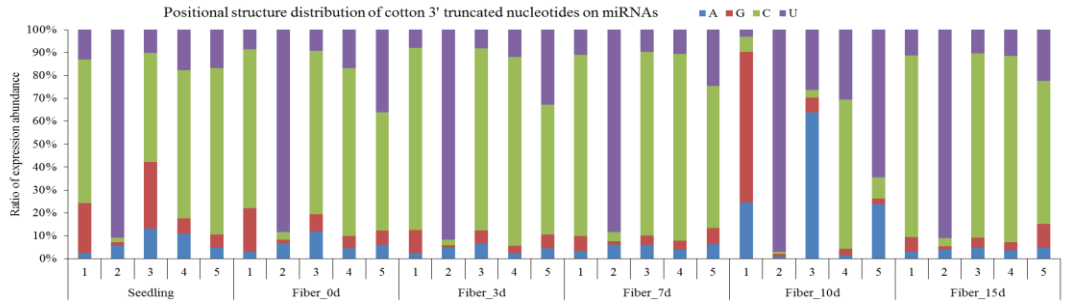
A



B



C



D

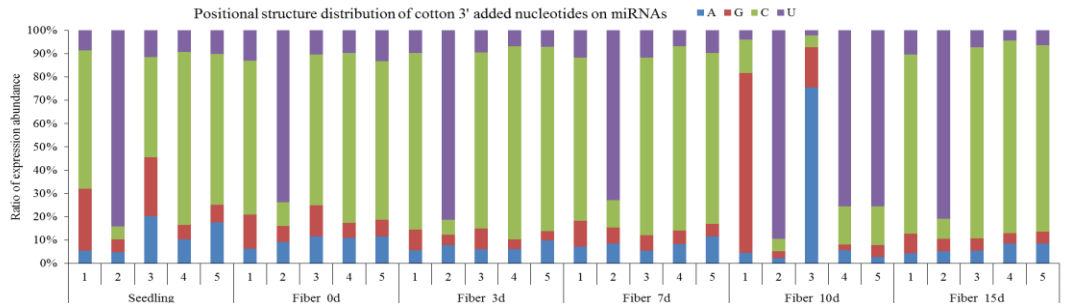


Figure 7-4. Positional structure distribution of cotton truncated or added nucleotides on the 5' and 3' end of miRNAs. A: truncated nucleotides on 5' end of miRNAs; B: added nucleotides on 5' end of miRNAs; C: truncated nucleotides on 3' end of miRNAs;

D: added nucleotides on 3' end of miRNAs. 1-5 on X axis denote the 1^{st}-5^{th} nucleotides that are truncated or added on the most 5' or 3' end of miRNAs.

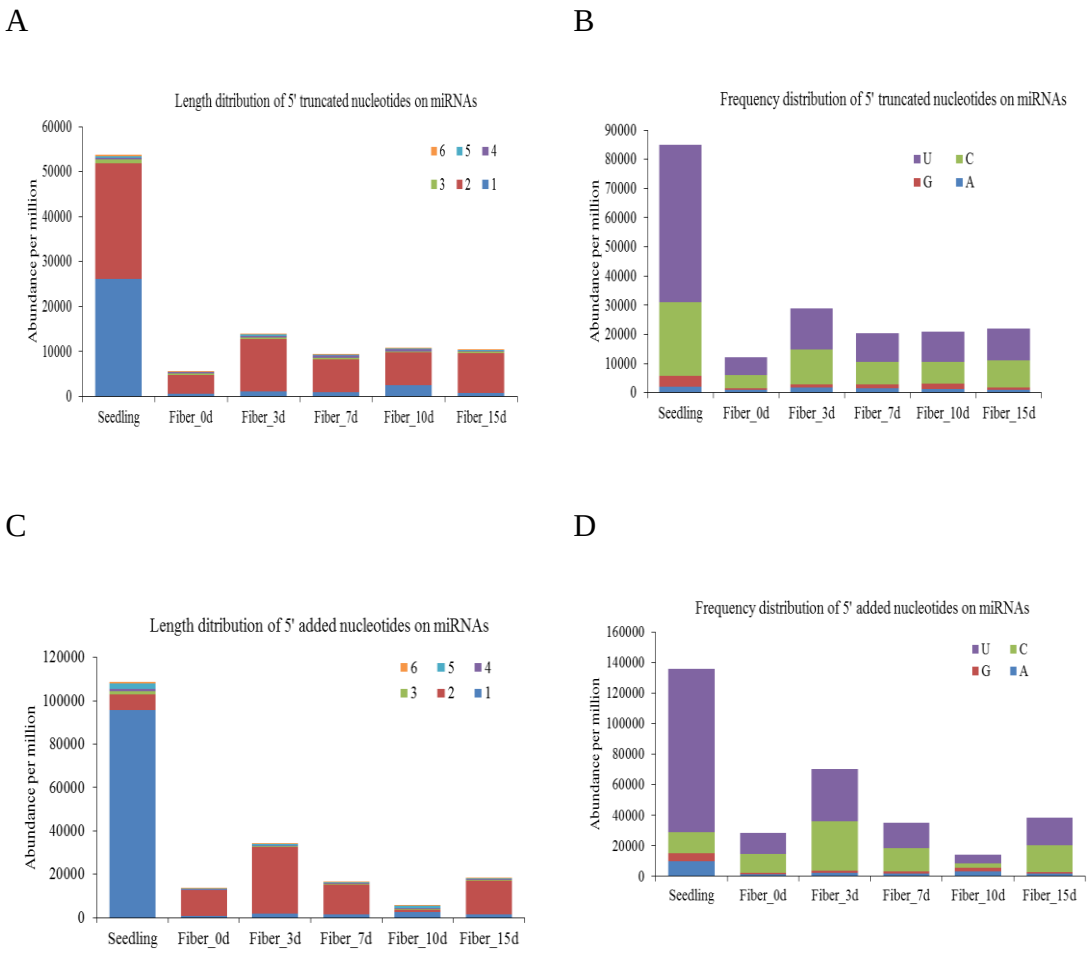


Figure 7-5. Length distribution and frequency distribution of truncated and added nucleotides on 5' end of miRNAs. A: Length distribution of truncated nucleotides on 5' end of miRNAs; B: Frequency distribution of truncated nucleotides on 5' end of miRNAs; C: Length distribution of added nucleotides on 5' end of miRNAs; D: Frequency distribution of added nucleotides on 5' end of miRNAs. Group 1-6 on A and C denote the number of truncated or added nucleotide on the 5' end of miRNAs, respectively.

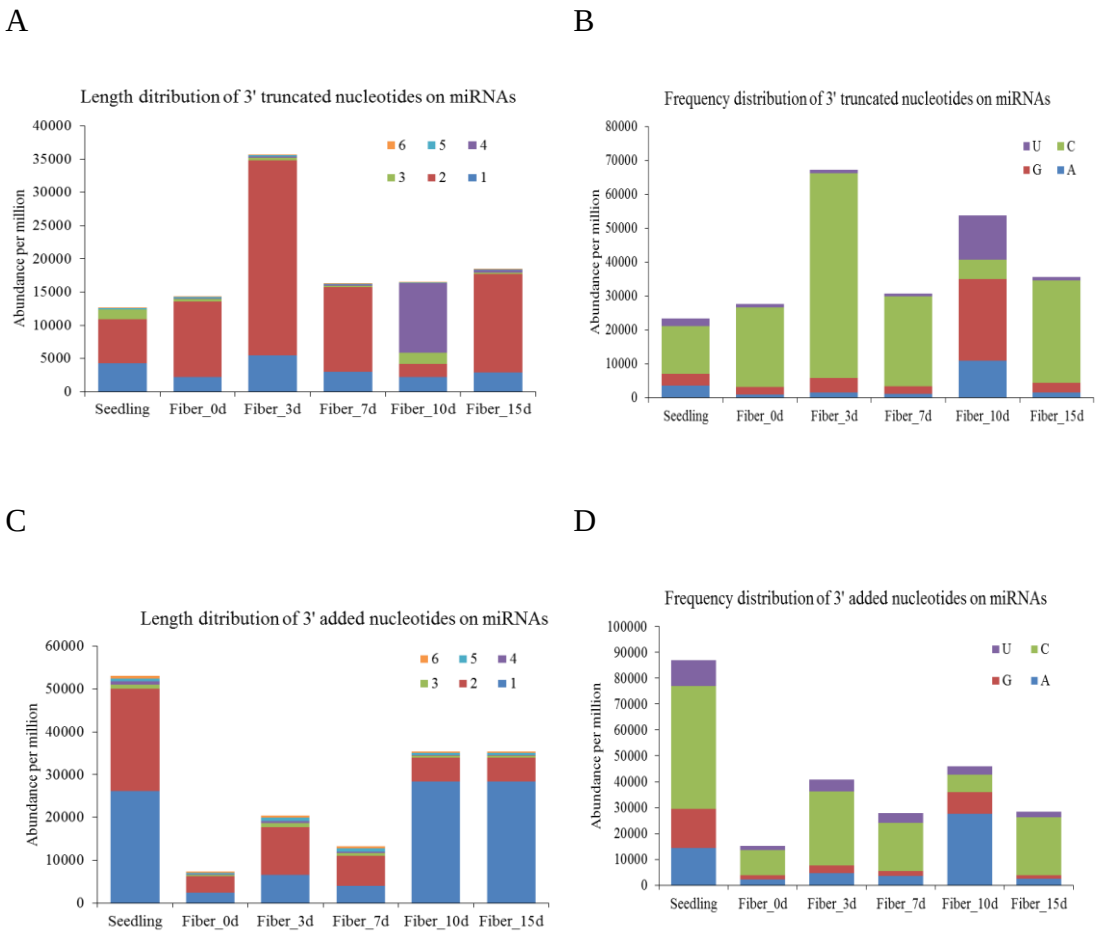
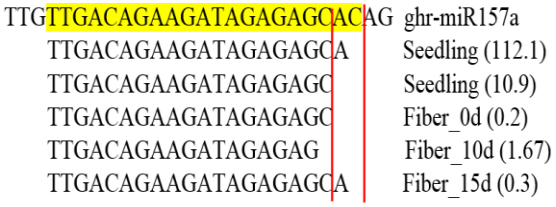
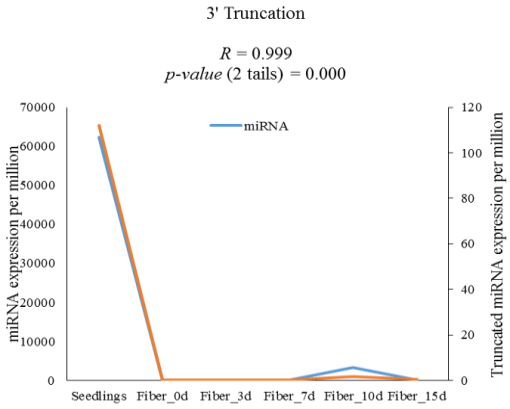


Figure 7-6. Length distribution and frequency distribution of truncated and added nucleotides on 3' end of miRNAs. A: Length distribution of truncated nucleotides on 3' end of miRNAs; B: Frequency distribution of truncated nucleotides on 3' end of miRNAs; C: Length distribution of added nucleotides on 3' end of miRNAs; D: Frequency distribution of added nucleotides on 3' end of miRNAs. Group 1-6 on A and C denote the number of truncated or added nucleotides on the 3' end of miRNAs, respectively.

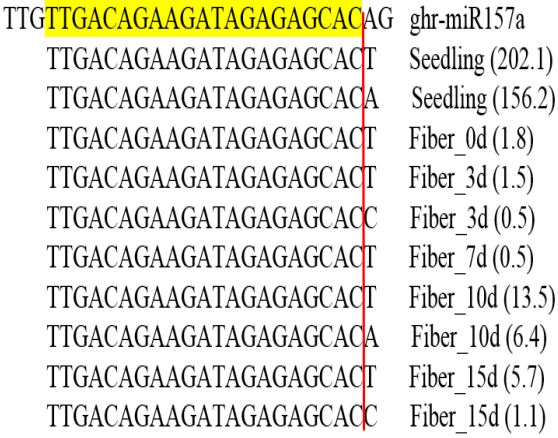
A



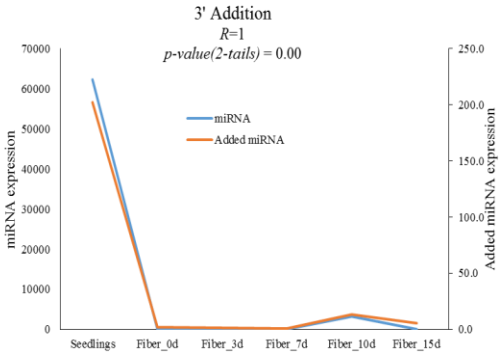
B



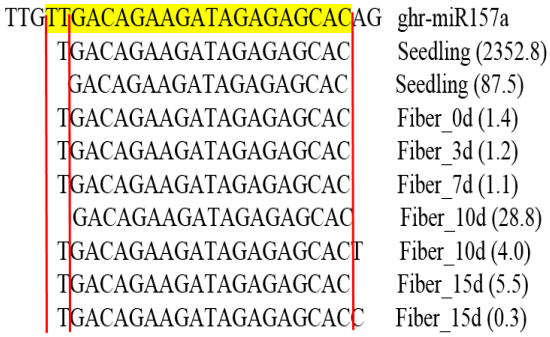
C



D



E



F

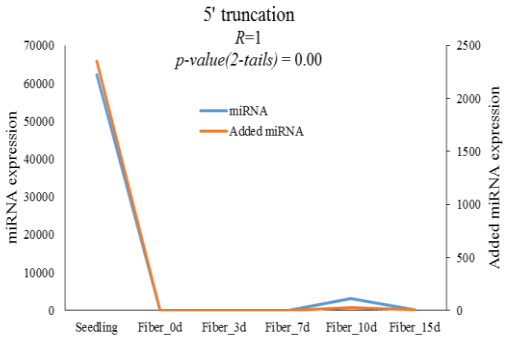


Figure 7-7. Nucleotide modification to cotton ghr-miR157a. A: A schematics of 3' truncated isomiRs derived from ghr-miR157a; B: Expression of ghr-miR157a and its 3'

truncated isomiRs; C: A schematics of 3' tailed isomiRs derived from ghr-miR157a; D: Expression of ghr-miR157a and its 3' tailed isomiRs; E: A schematics of 5' truncated isomiRs derived from ghr-miR157a; and F: Expression of ghr-miR157a and its 5' truncated isomiRs. The number in bracket in A, C, and E denotes the normalized abundance of related isomiRs. Only top 2 abundant isomiRs were listed in each developmental stage.

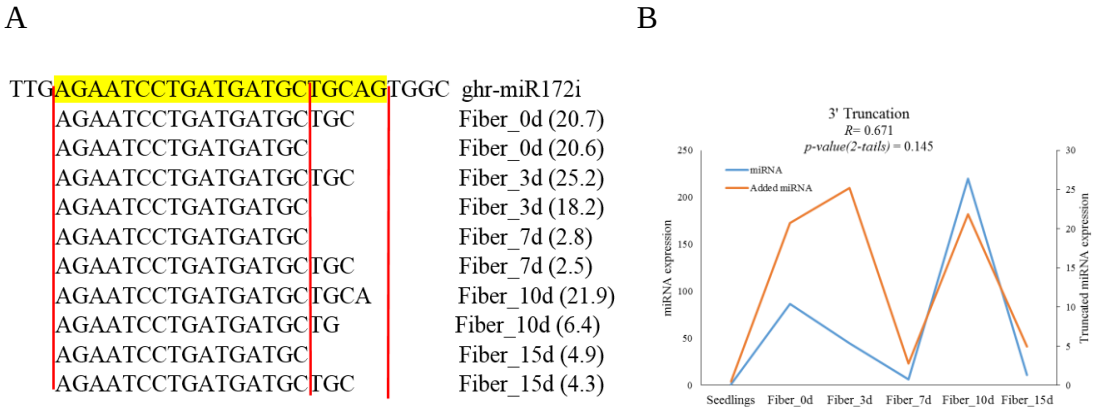


Figure 7-8. Nucleotide modification to cotton ghr-miR172i. A: A schematics of 3' truncated isomiRs derived from ghr-miR172i; and B: Expression of ghr-miR172i and its 3' truncated isomiRs. The number in bracket in A, C, and E denotes the normalized abundance of related isomiRs. Only top 2 abundant isomiRs were listed in each developmental stage.

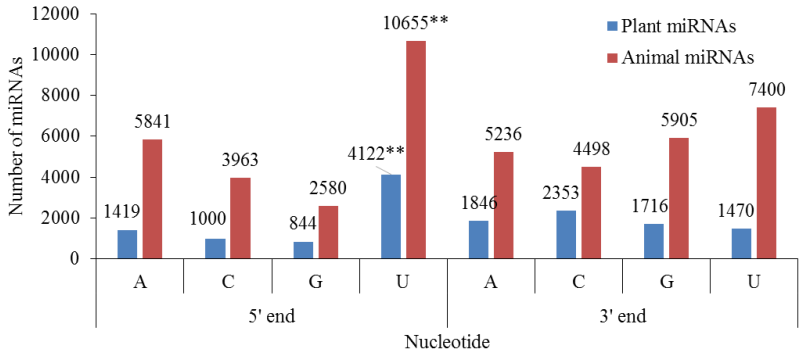


Figure 7-9. Frequency distribution of the first nucleotide on 5' and 3' ends of plant and animal known miRNAs.