Genome Biology and Evolution

Composition and Expression of Conserved MicroRNA Genes in Diploid Cotton (*Gossypium*) Species

Lei Gong[1], Atul Kakrana[2], Siwaret Arikit[3], Blake C. Meyers[2,3], and Jonathan F. Wendel[1,*]

[1]Department of Ecology, Evolution and Organismal Biology, Iowa State University

[2]Center for Bioinformatics and Computational Biology, Delaware Biotechnology Institute, University of Delaware

[3]Department of Plant and Soil Sciences, Delaware Biotechnology Institute, University of Delaware

*Corresponding author: E-mail: jfw@iastate.edu.

Accepted: November 22, 2013

Abstract

MicroRNAs are ubiquitous in plant genomes but vary greatly in their abundance within and conservation among plant lineages. To gain insight into the evolutionary birth/death dynamics of microRNA families, we sequenced small RNA and 5′-end PARE libraries generated from two closely related species of *Gossypium*. Here, we demonstrate that 33 microRNA families, with similar copy numbers and average evolutionary rates, are conserved in the two congeneric cottons. Analysis of the presence/absence of these microRNA families in other land plants sheds light on their depth of phylogenetic origin and lineage-specific loss/gain. Conserved microRNA families in *Gossypium* exhibit a striking interspecific asymmetry in expression, potentially connected to relative proximity to neighboring transposable elements. A complex correlated expression pattern of microRNA target genes with their controlling microRNAs indicates that possible functional divergence of conserved microRNA families can also exist even within a single plant genus.

Key words: miRNA, gene family evolution, biased gene expression, evolutionary divergence.

Introduction

MicroRNAs (miRNAs) are a diverse category of nuclear-encoded small RNAs that play multiple, central functions in eukaryotic development, stress responses, and many other biological processes (Mallory and Vaucheret 2006; Voinnet 2009; Axtell 2013). Primary transcripts (Pri-miRNA) encoded by miRNA genes fold into a stem-loop structure of the precursor transcript (pre-miRNA), which is further cleaved into the mature miRNA duplex, mostly by RNase III domain nucleases (Carthew and Sontheimer 2009). Mature miRNAs in the miRNA–miRNA* duplexes specifically recognize target transcripts via bound Argonaute (AGO) proteins, which facilitate cleavage (between the 10th and 11th nucleotide position from the 5′-end of the miRNA) of bound target genes and/or trigger translation repression via binding to the 3′UTR or coding region of the target mRNAs (Carthew and Sontheimer 2009).

Many miRNA families are conserved across vast phylogenetic scales, with some conserved within entire kingdoms (Zhang et al. 2006; Lee et al. 2007; Grimson et al. 2008). Different lineages, however, also contain miRNA genes with more restricted phylogenetic distributions (Voinnet 2009; Fahlgren et al. 2010), although our understanding of these distributions remains relatively limited. In plants, many miRNA genes are family- or species-specific (Cuperus et al. 2011), suggesting that miRNA genes in plants arose and diverged on scales ranging from family to species. These lineage-specific miRNAs tend to be expressed at low levels and may be transient miRNA genes that evolve neutrally (Axtell 2008; Fahlgren et al. 2010). Conserved miRNAs, which often have higher expression levels, appear to be characterized by a history of duplication and further sub- and/or neofunctionalization and regulate the gene expression of multiple targets (Maher et al. 2006; Chen and Rajewsky 2007; Rubio-Somoza et al. 2009; Debernardi et al. 2012). The foregoing suggests that miRNA gene evolution entails complex and as yet relatively poorly understood birth/death dynamics. Insights into these dynamics may emerge from comparative evolutionary analysis of miRNA gene content and function in two or more closely related species within individual genera.

This perspective motivated the present study of miRNA gene content in the phylogenetically well-understood cotton

© The Author(s) 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution

SMBE

genus (*Gossypium* L.). Among the 45 diploid ($n = 13$) species, 2 diploid clades, the Old World, A-genome and the New World, D-genome, diverged from a common ancestor about 5–10 Ma, subsequently acquiring a nearly 2-fold difference in genome size (Wendel et al. 2009; Wendel et al. 2010; Wendel et al. 2012). During the mid-Pleistocene (~1–2 Ma), representatives of these two divergent genomes became reunited in a common nucleus following hybridization and genome doubling, giving rise to a lineage now represented by modern allopolyploid (AD genome) cottons, which dominate cotton commerce worldwide. *Gossypium arboreum* (A2) and *G. raimondii* (D5) represent reasonably good models of the two diploid progenitors of allopolyploid cotton (Wendel et al. 2009; Wendel et al. 2010). Accordingly, *Gossypium* is a useful model for investigations of the genomic, transcriptomic, and proteomic consequences of polyploidy in plants (Hawkins et al. 2006; Hu et al. 2011; Wendel et al. 2012; Yoo et al. 2013).

Here, we focus on the evolution of conserved miRNA genes in diploid cotton species. This work was enabled by the recent completion of the first high-quality genome assembly for *G. raimondii* (Paterson et al. 2012) along with *G. arboreum* genome assembly (Udall JA and Page JT, unpublished data). We performed deep sequencing of small RNA libraries in conjunction with degradome (5′-end PARE, Parallel Analysis of RNA Ends) and RNA-Seq analyses in both diploid species, in the process describing miRNA compositional diversity and expression, origin of miRNA genes, miRNA gene expression, and miRNA target composition and correlated expression. These analyses reveal stability of conserved miRNA gene families accompanying the divergence of two congeneric species, but that this stability is accompanied by a striking interspecific asymmetry in miRNA gene expression and correlated expression patterns of their regulated target genes.

Materials and Methods

Library Construction and Sequencing

Three biological replicates of seedling leaves (3 cm in length, 7th post-cotyledonary) of *G. arboreum* and *G. raimondii* were collected, from which total RNAs were extracted using the Concert Plant RNA Reagent (Invitrogen, Cat. No. 12322-012). Gel size selection of small RNAs and subsequent sequencing library construction were completed as described (Lu et al. 2007). Using the same total RNAs, 5′-end PARE libraries were constructed following Zhai et al. (2013). Small RNA libraries and 5′-end PARE libraries were sequenced on the Illumina GA II sequencer, yielding 36-nt reads at the Sequencing and Genotyping Center at the University of Delaware. FAstaq files of raw sequencing reads are deposited in NCBI SRA database (SRR1029586–SRR1029588 and SRR616255–SRR616257). In small RNA libraries after preliminary processing, which involves adaptor trimming and poor-quality read filtration using our in-house tool "SSRTrim" (Solexa Small RNA Trimmer) with default parameter settings, FASTA-formatted reads were readied for analysis.

miRNA Annotation

Annotation of *G. raimondii* miRNAs in leaves has been described (Paterson et al. 2012), in which all the three replicates were combined into one file. MicroRNA annotation in *G. arboreum* was completed following the same workflow, except that an in-house genome assembly (Udall JA and Page JT, unpublished data) of *G. arboreum* (A2) was utilized as the initial reference genome for mapping the *G. arboreum* small RNA reads. The unpublished genome sequence for *G. arboreum* is freely available at the Comparative Evolutionary Genomics of Cotton website (http://128.192.141.98/CottonFiber/pages/genome/sequence.aspx, last accessed October 23, 2013). Some general quality indexes of this assembly are listed here: coverage of the genome after mapping and assembly (63.2%), number of scaffolds (1,612,870), and N50 scaffold length (2092 bp). Additionally, because mean scaffold length and contiguity are lower for *G. arboreum* than for *G. raimondii*, the allowed maximum copy number of each miRNA family was increased to 35 in the miRDeep-P program (Yang and Li 2011). The miRNA annotation in *G. raimondii* was done a second time using these same parameters to test for any effects of this parameter choice. Following established nomenclature (Meyers et al. 2008), pre-miRNAs with four or fewer nucleotide substitutions in their mature miRNAs were categorized into one gene family.

MicroRNA families with stringent homology (less than four substitutions) to known plant miRNA families in miRBase 20 (Kozomara and Griffiths-Jones 2011) were categorized as "conserved" and were named identically. Thus, miRNA families were tabulated into three categories (I, II, and III), representing, respectively, conserved and shared by both *G. arboreum* (A2) and *G. raimondii* (D5), conserved but detected only in *G. arboreum* (A2), and conserved but detected only in *G. raimondii* (D5).

Target Prediction and Validation

Based on the structural conservation of plant miRNA:target duplexes (Meyers et al. 2008), genome-wide target prediction was carried out using modified Targetfinder 1.6 (http://carringtonlab.org/resources/targetfinder, last accessed December 6, 2013). A wrapper was written in Python to add multiprocessing capabilities to the original Targetfinder 1.6. This enabled target prediction at genome level, which would have been difficult to perform using original Targetfinder. No modification to miRNA-Target scoring schema was made. For the whole-genome assembly of each species, miRNA:target duplex structures with score cut-off less than or equal to 7.0 were considered. These predicted targets were validated using in-house PARE prediction pipeline that

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution SMBE

employs Cleaveland3 algorithm for computing *P* values for miRNA:Target interaction from both geneic and intergenic regions, on the basis of 5′-ends PARE reads mapped to cleavage site (Addo-Quaye et al. 2009). These validated results were further filtered on the basis of PARE reads abundance at cleavage site (≥5), *P* value (<0.05), and abundance ratio of small to large window (≥0.75).

Based on homology with CDS regions of the annotated protein-coding genes in *G. raimondii* (reciprocal BlastN search using our in-house pipeline), 30,744 gene orthologs were determined in our *G. arboreum* genome assembly (unpublished data). The ID of each annotated gene in *G. raimondii* was assigned to corresponding gene ortholog in *G. arboreum*. In both species, if PARE-verified cleavage site lies within boundary of putative/annotated protein-coding gene, then the gene was accepted as PARE-verified target.

Evaluation of miRNA Duplication and Divergence

MicroRNAs in the same family were aligned using a local MAFFT tool version 7.031 (Katoh and Standley 2013), using the E-INS-I algorithm with default parameters because of the miRNA features of conservative stems and variable loops. Pairwise divergence (π) among members in the same family (within-family π) was calculated for each miRNA family in each species using "ape" (version 3.0-8) and "pegas" (version 0.4-4) packages in R workspace (Paradis et al. 2004; Paradis 2010). The density distributions of all within-family π values of different families were constructed (using R) for all conserved miRNA families. For each shared conserved family (Category I), within-family pairwise π values were compared to evaluate whether shared miRNA families maintained the same average evolutionary rate in two lineages after inheritance from their common ancestor. For this analysis, we used the Wilcoxon rank-sum test in R workspace (Wilcoxon 1947), where the FWER (Family-Wise Type I Error Rate) was controlled at 0.01 level using the Bonferroni correction.

Characterization of miRNA Gene Expression in Two Species

For the shared, conserved miRNAs families, their expression difference was evaluated in terms of their expressed mature miRNA read counts. Filtered reads of all three replicates in each species were mapped to the corresponding pre-miRNAs in the same gene family using Bowtie 0.12.7, which only allowed at most 13 multiple mapping positions (the largest number of miRNA gene copies in two species, table 1) and zero mismatch for each read (Langmead et al. 2009). In each replicate, the number of reads covering the mature miRNA regions of all pre-miRNAs in the same gene family was used to represent the expression of that gene family at the level of mature miRNAs. Differential expression of the shared and conserved miRNA families was determined using the Deseq package in R workspace (Anders and Huber 2010). Specifically, to minimize the variance introduced by the library size (or read depth), the initial RPM (reads per million)-normalized counts of each family in all replicates were further normalized using Deseq default normalization. Dispersion values were estimated using the defaulted "Maximum" method. The false discovery rate (FDR = 0.05) was controlled by the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995).

Table 1

Copy Numbers of Conserved miRNA Families in *Gossypium arboreum* (A2) and *G. raimondii* (D5)

Family Name[a]	A2 Copy Number	D5 Copy Number
miR156/157	12	12
miR159/319	3	1
miR160	6	8
miR162	1	1
miR164	5	3
miR165/166	11	11
miR167	6	6
miR169	12	13
miR170/171	8	12
miR172	7	8
miR2111	1	1
miR2947	1	1
miR2948	1	1
miR2949	1	1
miR2950	2	2
miR3476	1	1
miR3627	1	1
miR3441	5	3
miR390	3	3
miR393	4	4
miR394	3	2
miR395	4	1
miR396	5	6
miR397	2	1
miR398	1	2
miR399	6	7
miR403	1	1
miR473/477	2	6
miR479	1	1
miR530	2	3
miR535	2	2
miR827	1	1
miR828	1	1

[a]Families in bold have the same copy numbers in two species; those in italics have only a single member in both species.

Distribution of miRNA Families Relative to Transposons in the Two Species

Genome-wide annotations of transposable elements (TEs) for both genome assemblies were completed using a pipeline as described (Paterson et al. 2012). The TE closest to

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

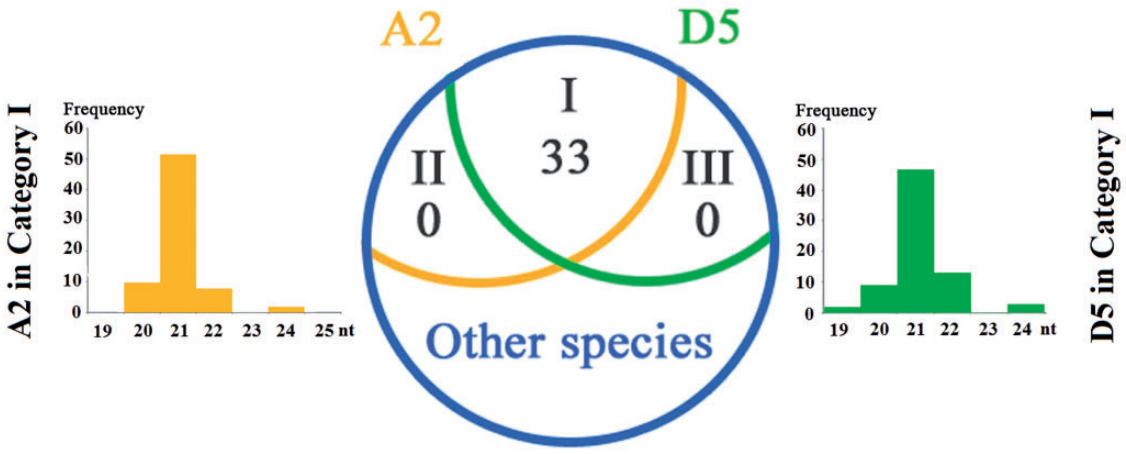

Fig. 1.—Composition and length distribution of conserved miRNAs in *Gossypium arboreum* (A2) and *G. raimondii* (D5). Shown is a diagram of conserved miRNA families (those previously annotated in *Gossypium* species and other plants). Category I: conserved miRNA families shared by both cotton species; Categories II and III: conserved miRNA families uniquely detected in either *G. arboreum* (A2) or *G. raimondii* (D5). Histograms of the length of mature miRNAs also are shown, with sequence lengths and frequencies denoted by the *x* and *y* axes, respectively. Orange, green, and blue colors denote information for *G. arboreum* (A2), *G. raimondii* (D5), and other plant species, respectively.

either end of a pre-miRNAs on the same strand was determined for all miRNA families, and the nucleotide distance between TEs and corresponding pre-miRNAs was calculated. Distances were averaged within each miRNA family, and paired Student's *t*-tests were used to evaluate the statistical significance of determined differences.

Characterization of miRNA Target Gene Expression

For miRNA families having significant differential expression between the two species, the transcriptional expression status of their target orthologous genes were characterized using published RNA-Seq data and methods (Yoo et al. 2013). The number of reads mapped to each target ortholog in each sample replicate was normalized by both total number of reads sequenced in each library using RPM and the Deseq default scaling factor. Between the two species, the difference of the averaged expression values of each ortholog pair was evaluated using one-tailed Student's *t*-test with unequal variance.

Results

Conserved miRNA Gene Family Composition in Diploid Cotton

Deep sequencing of small RNA libraries from *G. arboreum* (A2) and in *G. raimondii* (D5) led to the identification of 33 conserved miRNA families shared with other plant genomes (Category I in fig. 1; supplementary tables S1 and S2, Supplementary Material online), including 122 and 127 miRNA genes in *G. arboreum* and *G. raimondii*, respectively. Notably, no single miRNA family was conserved between only one of the two sequenced cotton species and other plants ($n = 0$ for Categories II and III). For our annotated miRNA families with stringent homology with the *Gossypium* miRNA families already deposited in miRBase 20, we accepted these annotations as conserved miRNA families. Most miRNAs from both species in Category I were 21 nt in length (fig. 1, 72.22% in *G. arboreum* and 62.16% in *G. raimondii*), but there was a minority presence of miRNAs of other lengths in this category.

We tabulated copy numbers for each miRNA family in Category I. Eighteen families ($18/33 = 54.55\%$) had the same number of copies in the two cotton species (table 1), of which 11 ($11/18 = 61.11\%$) contained only a single miRNA in both genomes. Overall, there was no significant difference in copy number for conserved miRNAs families in the two diploid *Gossypium* species (paired *t*-statistic $= -1.02$, df $= 31$, *P* value $= 0.32$).

Evolution of miRNA Family

To understand the origins and evolutionary histories of conserved miRNAs in *Gossypium*, we tabulated observations from other land plants, using sequences deposited in miRBase 20 (supplementary table S3, Supplementary Material online). To diagnose the origin of each miRNA family during plant evolution in those most curated sequenced species (from Phytozome 9.0.1, http://www.phytozome.net/, last accessed December 6, 2013), we mapped miRNA family presence/absence onto the green plant tree (illustrated in fig. 2). MicroRNA families with possible alternative classifications (miR156/157, miR159/319, miR165/166, miR170/171, and miR473/477) were excluded from this analysis (Meyers et al. 2008). As illustrated (fig. 2), cotton miRNA families may be classified into four groups: 1) Those detected in multiple

Genome Biology and Evolution

SMBE

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution SMBE

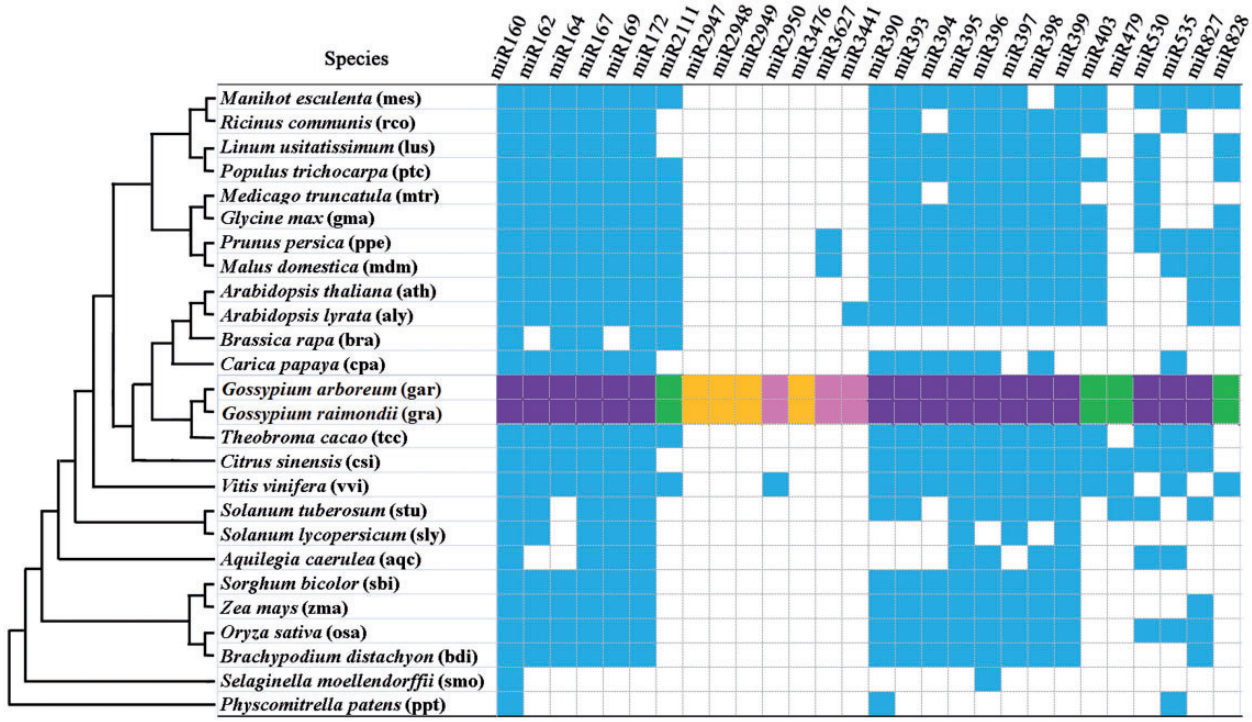

Fig. 2.—Conserved miRNA families in sequenced land plant species. In each sequenced land plant species (rows, with abbreviated miRBase three-letter names parenthesized) deposited in Phytozome 9.0.1, their miRNA families shared with conserved miRNA families (miRNA families with no alternative classifications) identified in *Gossypium arboreum* (A2) and *G. raimondii* (D5) (columns) are tabulated. Each blue and white shaded cell represents presence and absence of a miRNA family in non-*Gossypium* species, respectively. Based on the phylogenetic tree (left), the cells shaded in other colors denote the groups of *Gossypium* miRNA families in terms of their origins and evolutionary histories: purple shading marks miRNA families with ancient origins that predate angiosperms; green shading denotes families that have their origins near the base of the eudicots; orange shading denotes families that have a recent ancestry which includes at least *Gossypium*; and pink shading marks miRNA families with independent sporadic occurrences in both *Gossypium* and other taxa.

eudicots and monocots, indicating an ancient origin that predates angiosperms (purple shading); the absence of some of these families in some lineages therefore is most parsimoniously explained by miRNA gene loss, allowing for the possible explanation of incomplete genome sequence data and incomplete miRNA annotations. 2) Four families (green shading), including miR2111, miR403, miR479, and miR828, are shared with other eudicots, including the relatively basal *Vitis vinifera*; these miRNAs likely have their origins near the base of the eudicots. 3) The four cotton-specific miRNA families, identified and deposited previously in miRBase 20 (miR2947, miR2948, miR2949, and miR3476, yellow shading in fig. 2), likely evolved in the more recent ancestry of *Gossypium*, which is confirmed by the lack of homologous loci in the genome sequences of all other species in figure 2; ascertaining the phylogenetic extent of occurrence of these miRNA families will require sampling more genera in the Malvaceae and perhaps beyond. 4) Three additional families (miR2950, miR3627, and miR3441, pink shading) have a sporadic occurrence in both *Gossypium* and a few other species. For example, miR2950 was only detected in *Vitis vinifera* and *Gossypium*, the most parsimonious explanation of this being an independent origin. Alternatively, this miRNA may have been incorrectly annotated in one of the species.

To evaluate whether shared miRNA families maintained the same average evolutionary rate in two lineages after inheritance from their common ancestor, we calculated pairwise nucleotide divergences (pairwise π) within each miRNA family within each species and fitted these data onto smoothed histograms (fig. 3). As shown, curves in both species were similar, located within a range of π values from 0 to 0.55 and with nonsignificant *P* values (*P* values > 0.01) by Wilcoxon rank-sum tests for all shared families. In addition, in both species, the majority of π values were clustered in a range of values larger than 0.2; however, a peak at 0.16 in *G. arboreum* and a peak at 0.24 in *G. raimondii* suggested two possible evolutionary duplication events of miRNA families (fig. 3).

Comparative Expression and Distribution of miRNA Genes

Expression of mature miRNAs from the shared, conserved families was compared. Overall, 23 of the 33 shared,

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution

SMBE

conserved miRNA families (23/33 = 69.70%) were differentially expressed in the two species (fig. 4). In the other 10 families (miR393, miR399, miR473/477, miR530, miR827, miR828, miR2111, miR2947, miR2949, and miR3823), the two cotton species exhibited similar expression levels. Most noticeable in figure 4 is the result that among the differentially expressed families, all but one (miR398) were expressed more highly in *G. arboreum* than in *G. raimondii*.

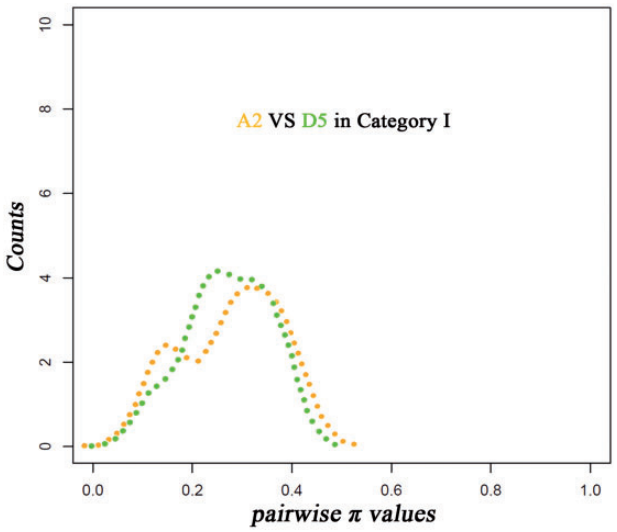

Fig. 3.—Density curves of miRNA sequence divergence (pairwise π) fitted on histograms. The *x* and *y* axes indicate divergence for all pairwise comparisons in each family and the density at a given divergence point, respectively. Orange and green colors denote data for *Gossypium arboreum* (A2) and *G. raimondii* (D5), respectively.

In an effort to account for why there was such asymmetric expression among miRNA families in the two species, the locations of miRNA families relative to their nearest TEs were analyzed (table 2). Although some exceptional miRNA families that had higher expression in *G. arboreum* than in *G. raimondii* were closer to the nearest TE in *G. arboreum* (A2), the overall distances of miRNAs to their nearest neighbor TE in *G. arboreum* was statistically higher than in *G. raimondii* (D5) (paired Student's *t*-test, $P < 0.05$). In addition, the miRNAs with no differential expression have similar proximities to their nearest TEs in the two species (paired Student's *t*-test, $P > 0.05$). All of these results implicate a possible inverse relationship between the expression of miRNA genes and their distance to the nearest neighboring TEs. The miR398 family, with uniquely higher expression in *G. raimondii* (D5) than in *G. arboreum* (A2), consistently showed a more distant distribution from its nearest neighboring TEs in *G. raimondii* (D5) (table 2).

Comparative Expression of miRNA-Targeted Genes

Following complementarity-based target prediction, 100% of the annotated miRNA families in both species were believed to form potential miRNA:target duplexes. Among these, both species had a high proportion of families with PARE-verified cleaved targets (16/33 = 48.49% in *G. arboreum* and 22/33 = 66.67% in *G. raimondii*) (supplementary tables S4 and S5, Supplementary Material online). The lower rate of PARE verification in *G. arboreum* and lack of PARE verification for one-third of the families may reflect the relatively low coverage of the 5′-end PARE reads in *G. arboreum* at the cleavage sites on predicted miRNA:target duplexes. There were 14

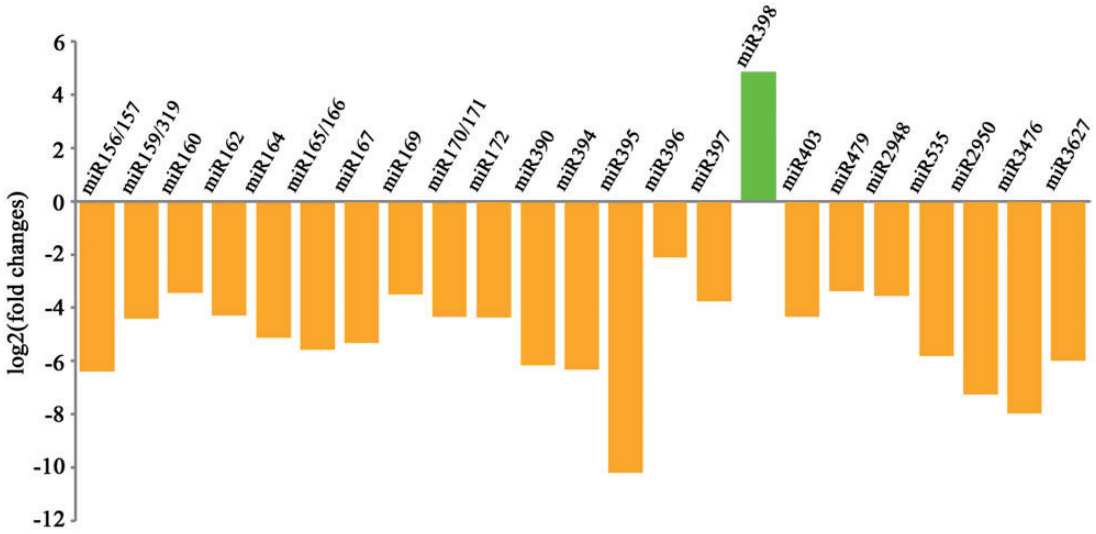

Fig. 4.—Differential expression of miRNA gene families in *Gossypium arboreum* (higher expression in *G. arboreum*; orange) and *G. raimondii* (higher expression in *G. raimondii*; green). MiRNA gene families with significant differential expression are shown. Log2 transformations of the expression fold changes (*G. raimondii* vs. *G. arboreum*) are represented by bars. *y* axis denotes the levels of transformed expression fold changes.

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution SMBE

Table 2
Average Distance between miRNA Families and Their Nearest Neighboring TEs in *Gossypium arboreum* (A2) and *G. raimondii* (D5)

Differentially Expressed miRNA Families[a]	Average Distance (bp) between miRNA and Nearest TE in A2	Average Distance (bp) between miRNA and Nearest TE in D5[b]
miR156/157	1,648	953
miR159/319	1,505	2,870
miR160	2,969	353
miR162	3,769	1,540
miR164	1,648	1,115
miR165/166	1,519	1,680
miR167	1,688	1,522
miR169	1,771	1,513
miR170/171	1,807	1,129
miR172	1,555	610
miR390	6,381	51
miR394	2,763	3,976
miR395	1,744	1
miR396	754	1,351
miR397	2,242	742
miR398	154	313
miR403	789	1
miR479	1,795	1
miR2948	348	655
miR535	1,634	923
miR2950	1,447	1,620
miR3476	3,431	1
miR3627	621	1

Nondifferentially Expressed miRNA Families	Average Distance (bp) between miRNA and Nearest TE in A2	Average Distance (bp) between miRNA and Nearest TE in D5[b]
miR393	870	792
miR399	1,102	909
miR473/477	782	723
miR530	678	597
miR827	903	678
miR828	1,902	2,210
miR2111	1,823	2,081
miR2947	2,109	1,834
miR2949	729	690

[a]Families that show higher expression in *G. arboreum* (A2) and are more distant from the nearest TE than in *G. raimondii* (D5) are shown in normal characters; those in italics do not have the inverse relationship. miR398 (in bold) is the sole family with higher expression in D5.

[b]To facilitate statistical testing, the average distances of families with miRNAs residing in the TEs are denoted as "1."

overlapping miRNA families with PARE-verified targets in both species (supplementary table S6, Supplementary Material online).

To explore the potential functional implication of the asymmetric differential expression of conserved miRNA families (fig. 4), we studied the expression of their target genes in *G. arboreum* (A2) and *G. raimondii* (D5) (fig. 5 and supplementary table S7, Supplementary Material online). For miRNA398, as discovered in *Arabidopsis thaliana*, miR398 mainly targets three kinds of genes via transcriptional cleavage: cytosolic CSD1 (AT1G08830) and chloroplast-localized CSD2 (AT2G28190), COX5b-1 (AT3G15640), and CCS1 (AT1G12520) (Bonnet et al. 2004; Jones-Rhoades and Bartel 2004; Sunkar and Zhu 2004; Beauclair et al. 2010; Zhu et al. 2011). After searching by sequence homology, their corresponding homologs in cotton species were determined (fig. 5*a*). Through Targetfinder prediction and PARE validation, all homologous genes in *G. raimondii* (D5) were also verified as target genes (fig. 5*a*). Given that the two cotton species have on average only about 1–2% sequence divergence in their protein-coding genes (Senchina et al. 2003; Flagel et al. 2012), it is reasonable to expect that their corresponding ortholog genes should also be targets of the same miR398 family. RNA-Seq data revealed that most of the miR398 target genes did not show significant differential expression in the two cotton species; however, the CSD2 gene (Gorai.009G090300) in *G. arboreum* (A2) was expressed at a significantly higher level than its ortholog in *G. raimondii* (D5) (fig. 5*a*; $P < 0.001$), suggesting a correlation between higher expression of miR398 and responsive repression of CSD2.

In addition to miR398, we examined gene expression in *G. arboreum* (A2) and *G. raimondii* (D5) of putative targets of the 21 miRNA families that exhibit higher miRNA expression in *G. arboreum* (A2) (fig. 4). To ensure a conservative list of target genes by each family, only the target gene homologs with PARE-verified cleavage sites identified in both species were included in this analysis. Using this stringency criterion, seven miRNA families with higher RNA-sequencing expression in *G. arboreum* (A2) survived this filter, for which there were 19 protein genes targeted (fig. 5*b*). Based on sequence homology with homologs in *A. thaliana*, the targeted genes were categorized into different functional groups. Notably, genes targeted by each miRNA family invariably have similar putative functions. For example, genes targeted by miRNA160 were all auxin response factors (fig. 5*b*). Among the nine targeted genes with significant differential expression between the two cotton species, five gene homologs (Gorai.013G267100, Gorai.007G038100, Gorai.003G139800, Gorai.004G002100, and Gorai.005G098700) also had lower expression in *G. arboreum* (A2) than in *G. raimondii* (D5) (negative correlation with expression of their controlling miRNAs), but the other four genes (Gorai.002G181700, Gorai.010G046000, Gorai.006G008700, and Gorai.010G048800) were expressed significantly higher in *G. arboreum* (A2) (fig. 5*b*).

Discussion

It has long been apparent that miRNAs comprise a diverse assemblage of related sequences, which vary in their phylogenetic distribution and relative breadth of conservation among various plant families, yet there remain few studies

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution

SMBE

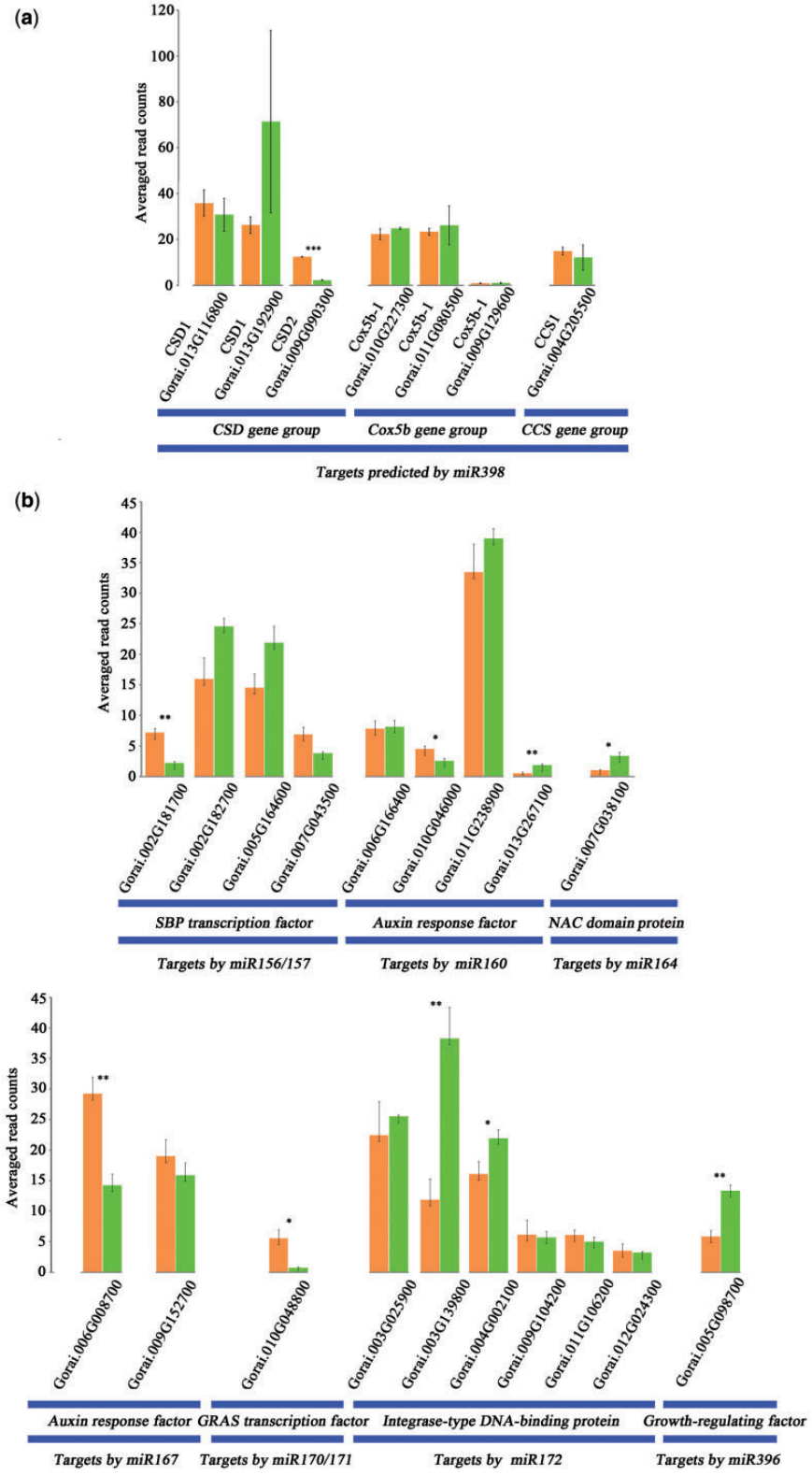

Fig. 5.—Expression of gene homologs targeted by miRNA families with differential expression between *Gossypium arboreum* and *G. raimondii*. Expressed read counts of genes and standard errors (relative to vertical *y* axis) are shown (orange = *G. arboreum* [A2]; green = *G. raimondii* [D5]). Gene IDs in *G. raimondii* (annotation file at Phytozome 9.0.1) are listed at the bottom. Single, double, and triple asterisks denote significantly different expression at $\alpha = 0.05$, 0.01, and 0.001, respectively. Expression comparisons of the genes targeted by miR398 and other miRNA families (miR156/167, miR160, miR164,

(continued)

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution SMBE

of the genesis of this pattern. To gain insight into the evolutionary dynamics of miRNAs, we employed a phylogenetic comparative framework involving two closely related cotton species, *G. arboreum* and *G. raimondii*, whose divergence time is reasonably well understood and which have genomes that vary nearly 2-fold in size. To accomplish this, we performed deep sequencing of small RNA libraries combined with analyses of miRNA family composition, biogenesis history, miRNA expression, and composition and expression of miRNA-targeted genes.

miRNA Gene Family Conservation

The 33 miRNAs families that were conserved between cotton and other species add to our understanding that most miRNA families are ancient and stable over vast evolutionary timescales (fig. 2). Specifically, for the two diploid cotton species, there was no independent loss/gain of conserved miRNA families (fig. 1), similar copy numbers of conserved families (table 1), and similar family-wide nucleotide diversities in both species (fig. 3). For comparison, in two sequenced *Arabidopsis* species (*A. thaliana* and *A. lyrata*) with clear miRNA annotations, there are 36 shared families but also 2 unshared families (miR447 exists only in *A. thaliana* and *Vitis vinifera* and miR1886 exists only in *A. thaliana* and *Solanum tuberosum*) (Fahlgren et al. 2010). Thus, conserved miRNA families, as expected, are not particularly evolutionary labile at the level of a single genus. This same conclusion appears to hold for copy numbers within miRNA families, noting the insignificant copy number variation of conserved families in the two cotton species (fig. 3 and table 1).

Evaluation of the phylogenetic distribution of each conserved miRNA family detected in *Gossypium* provides additional insights into the patterns of gain and loss during land plant evolution. As noted earlier, most conserved miRNA families in *Gossypium* are ancient, with many arising prior to the origin of flowering plants (Jones-Rhoades and Bartel 2004; Cuperus et al. 2011; Axtell 2013). As shown in figure 2, however, new families of miRNAs arise and may be lost in a lineage-specific fashion at various phylogenetic depths, some tracing to the root of the eudicots or the root of asterids. It will be of interest to continue to explore the phylogenetic distribution of miRNA families, both to unravel the timing and nature of family origin and loss in different lineages and also to set the stage for generating insight into the possible functional or adaptive significance of these patterns.

Asymmetric Expression of miRNAs: Mechanisms and Possible Functional Consequences

Given the centrality of miRNAs in regulation of important physiological and developmental processes, it is of interest to explore differences in miRNA expression among different species. Here, 10 of the 33 conserved miRNA families were expressed at equivalent levels in both cotton species, even after their isolation in different lineages for 5–10 million years. These data suggest that these ancient miRNA families are functionally as well as evolutionarily stable. For the remaining 23 miRNA families conserved between *Gossypium* and other plants, there was a striking asymmetry in collective expression (fig. 4), with all but one (miR398) having higher expression in *G. arboreum* than in *G. raimondii*. Although multiple factors regulate miRNA expression in plants (Xie et al. 2010), one possible factor is differential accumulation of TEs, a feature that characterizes the two cotton genomes studied here (Hawkins et al. 2006). Because TEs often have repressive effects on proximal genes (Wang et al. 2013) via promoter disruptions, spread of epigenetically induced silencing, and antisense transcription (Kashkush et al. 2003; Zhang et al. 2008; Hollister and Gaut 2009; Ahmed et al. 2011), we studied the correlation between miRNA adjacency in cotton to nearby TEs. Our results are suggestive in this regard, but perhaps not compelling, with a statistically significant inverse relationship between the expression of miRNA genes and their distance to the nearest neighboring TEs but with some notable exceptions (table 2). The miR398 family, which uniquely exhibited higher expression in *G. raimondii* (D5) than in *G. arboreum* (A2), was also physically more distant from its nearest neighbor TE in *G. raimondii* (D5) (table 2). Given the fact that *G. raimondii* (D5) genome is half the size of that of *G. arboreum* (A2), which is almost entirely due to less TE content in D5 than A2 (Hawkins et al. 2006; Grover and Wendel 2010), the possible effects of more loaded 24 nt siRNAs from overrepresented TEs on lower expression of D5 miRNAs can be excluded.

To explore whether asymmetric expression of conserved miRNAs had functional implications, we also analyzed expression of their downstream target genes (fig. 5). As confirmed repeatedly in both plants and animals (Axtell and Bartel 2005; Wang and Li 2009), and as recently observed in developing anthers of *G. hirsutum* (Wei et al. 2013), most target genes displayed a consistently negative expression correlation with their interacting regulatory miRNAs (fig. 5). For example, in leaves of diploid cottons (data presented here) and anthers in

Fig. 5.—Continued

miR167, miR170/171, miR172, and miR396) are illustrated in panels (*a*) and (*b*), respectively. Target genes were categorized into different functional groups based on homology with known homologs in *Arabidopsis thaliana*. Shown are the targeted genes encoding cytosolic CSD1 and chloroplast-localized CSD2 (copper/zinc superoxide dismutases), COX5b-1 (one subunit of the mitochondrial cytochrome c oxidase), CCS1 (the copper chaperone for Cu/Zn-SODs), SBP transcription factor (squamosa promoter-binding protein-like transcription factor), GRAS transcription factor (transcription factors in GAI, RGA, and SCR family in plant growth and development), and RAP2.7 (Integrase-type DNA-binding protein).

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution

SMBE

G. hirsutum (Wei et al. 2013), interactions of miR398:CSD2 (Gorai.009G090300), miR160:ARF16 (Gorai.013G267100), miR164:NAC100 (Gorai.007G038100), miR172:RAP2.7 (Gorai.003G139800), miR172:RAP2.7 (Gorai.004G002100), and miR396:GRF1 (Gorai.005G098700) were all identified and negative correlated expression was observed. Together with the described asymmetric expression of miRNA genes, responsive expression changes of target genes indicate a possible functional divergence of conserved miRNA families after speciation in the same genus. Exceptions to this expected pattern also were observed here, for example, miR156/157:SBP factor (Gorai.002G181700), miR160:ARF17 (Gorai.010G046000), miR167:ARF6 (Gorai.006G008700), and miR170/171:GRAS factor (Gorai.010G048800) (fig. 5*b*). This absence of negative correlated expression between miRNA genes and their targets has also previously been reported in both plants and animals (Voinnet 2009; Nunez-Iglesias et al. 2010; Lopez-Gomollon et al. 2012; Zhang et al. 2012). Thus, the functional significance of the striking asymmetry in conserved miRNA expression between the two cotton species remains obscure.

There are many possible explanations for the absence of perfect negative correlation between miRNA expression and expression of presumptive targets. A partial list includes regulation at other levels, including mRNA stability, the myriad factors involved in transcriptional regulation, the possibility that miRNAs and their targets have spatially separated expression in different domains or cell types (Voinnet 2009), threshold effects between miRNA abundance and target regulation (Mukherji et al. 2011), and feedback effects, where binding of the target-encoded protein, as "trans" enhancing factors, to upstream regulatory regions of the controlling miRNAs results in positively correlated gene expression (Megraw et al. 2006; Wu et al. 2009). Collectively, these and other factors may be involved in the various expression patterns of target genes observed here for the two *Gossypium* species.

Conclusion

We have shown that genome-wide composition characterization and evolutionary comparison of miRNA genes provides new perspectives on miRNA evolution. The results demonstrate the temporal scale and scope of miRNA family conservation at several phylogenetic levels and establish different origins and evolutionary histories of conserved miRNAs. Additionally, we demonstrate a striking asymmetric differential expression of the conserved, shared miRNA families in the two cotton species that is inversely associated with distance to neighboring TEs and negatively correlated with the expression of their target genes in most cases. Additional phylogenetically informed, comparative analyses in other *Gossypium* species and related outgroups will improve our understanding of miRNA categorization, genesis, and subsequent evolutionary fate. These studies may be especially informative when combined with functional analysis, including, for example, the use of miRNA gene knock-down or enhancing mutants and/or target gene mutagenesis.

Supplementary Material

Supplementary tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Corrinne E. Grover, Kara Grupp, Jixian Zhai, and Guanjing Hu for assistance with laboratory or data analysis and Heidrun Gundlach for annotation of the transposable elements in *G. arboreum*. We gratefully acknowledge support from the National Science Foundation Plant Genome Program (#0817707) and the Partner University Fund.

Literature Cited

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

Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. 2011. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in *Arabidopsis*. Nucleic Acids Res. 39:6919–6931.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Axtell MJ. 2008. Evolution of microRNAs and their targets: are all microRNAs biologically relevant? Biochim Biophys Acta. 1779:725–734.

Axtell MJ. 2013. Classification and comparison of small RNAs from plants. Annu Rev Plant Biol. 64:137–159.

Axtell MJ, Bartel DP. 2005. Antiquity of microRNAs and their targets in land plants. Plant Cell 17:1658–1673.

Beauclair L, Yu A, Bouché N. 2010. microRNA-directed cleavage and translational repression of the copper chaperone for superoxide dismutase mRNA in *Arabidopsis*. Plant J. 62:454–462.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 57:289–300.

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

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. Cell 136:642–655.

Chen K, Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 8:93–103.

Cuperus JT, Fahlgren N, Carrington JC. 2011. Evolution and functional diversification of miRNA genes. Plant Cell 23:431–442.

Debernardi JM, Rodriguez RE, Mecchia MA, Mail JF. 2012. Functional specialization of the plant miR396 regulatory network through distinct microRNA-target interactions. PLoS Genet. 8:e1002419.

Fahlgren N, et al. 2010. MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. Plant Cell 22:1074–1089.

Flagel LE, Wendel JF, Udall JA. 2012. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics 13:302.

Grimson A, et al. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455:1193–1197.

Grover CE, Wendel JF. 2010. Recent insights into mechanisms of genome size change in plants. J Bot. 14:1–8.

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 16: 1252–1261.
Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–1428.
Hu G, et al. 2011. Genomically biased accumulation of seed storage proteins in allopolyploid cotton. Genetics 189:1103–1115.
Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell. 14:787–799.
Kashkush K, Feldman M, Levy AA. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet. 33:102–106.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 39: D152–D157.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Lee C-T, Risom T, Strauss WM. 2007. Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. DNA Cell Biol. 26:209–218.
Lopez-Gomollon S, Mohorianu I, Szittya G, Moulton V, Dalmay T. 2012. Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions. Planta 236:1875–1887.
Lu C, Meyer BC, Green PJ. 2007. Construction of small RNA cDNA libraries for deep sequencing. Methods 43:110–117.
Maher C, Stein L, Ware D. 2006. Evolution of *Arabidopsis* microRNA families through duplication events. Genome Res. 16:510–519.
Mallory AC, Vaucheret H. 2006. Functions of microRNAs and related small RNAs in plants. Nat Genet. 38:S31–S36.
Megraw M, et al. 2006. MicroRNA promoter element discovery in *Arabidopsis*. RNA 12:1612–1619.
Meyers BC, et al. 2008. Criteria for annotation of plant microRNAs. Plant Cell 20:3186–3190.
Mukherji S, et al. 2011. MicroRNAs can generate thresholds in target gene expression. Nat Genet. 43:854–859.
Nunez-Iglesias J, Liu C-C, Morgan TE, Finch CE, Zhou XJ. 2010. Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. PLoS One 5:e8898.
Paradis E. 2010. Pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Paterson AH, et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 492: 423–427.
Rubio-Somoza I, Cuperus JT, Weigel D, Carrington JC. 2009. Regulation and functional specialization of small RNA-target nodes during plant development. Curr Opin Plant Biol. 12:622–627.
Senchina DS, et al. 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol Biol Evol. 20:633–643.
Sunkar R, Zhu J-K. 2004. Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. Plant Cell 16:2001–2019.
Voinnet O. 2009. Origin, biogenesis, and activity of plant microRNAs. Cell 136:669–687.
Wang X, Weigel D, Smith LM. 2013. Transposon variants and their effects on gene expression in *Arabidopsis*. PLoS Genet. 9:e1003255.
Wang Y-P, Li K-B. 2009. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics 10:218.
Wei M, et al. 2013. Comparative expression profiling of miRNA during anther development in genetic male sterile and wild type cotton. BMC Plant Biol. 13:66.
Wendel JF, Brubaker C, Alvarez I, Cronn R, Stewart JM. 2009. Evolution and natural history of the cotton genus. In: Paterson AH, editor. Genetics and genomics of cotton. New York: Springer. p. 3–22.
Wendel JF, Brubaker CL, Seelanan T. 2010. The origin and evolution of *Gossypium*. In: Stewart JM, Oosterhuis DM, Heitholt JJ, Mauney JR, editors. Physiology of cotton. Dordrecht (The Netherlands): Springer. p. 1–18.
Wendel JF, Flagel LE, Adams KL. 2012. Jeans, genes, and genomes: cotton as a model for studying polyploidy. In: Soltis PS, Soltis DE, editors. Polyploidy and genome evolution. Berlin (Germany): Springer Berlin Heidelberg. p. 181–207.
Wilcoxon F. 1947. Probability tables for individual comparisons by ranking methods. Biometrics 3:119–122.
Wu G, et al. 2009. The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. Cell 138:750–759.
Xie Z, Khanna K, Ruan S. 2010. Expression of microRNAs and its regulation in plants. Semin Cell Dev Biol. 21:790–797.
Yang X, Li L. 2011. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics 27:2614–2615.
Yoo M-J, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110: 171–180.
Zhai J, Arikita S, Simon SA, Kingham BF, Meyers BC. 2013. Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. Methods 13:S1046–2023.
Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA. 2006. Conservation and divergence of plant microRNA genes. Plant J. 46:243–259.
Zhang J-Z, et al. 2012. Identification of miRNAs and their target genes using deep sequencing and degradome analysis in trifoliate orange [*Poncirus trifoliata* L. Raf]. Mol Biotechnol. 51:44–57.
Zhang X, Shiu S, Cal A, Borevitz JO. 2008. Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. PLoS Genet. 4:e1000032.
Zhu C, Ding Y, Liu H. 2011. MiR398 and plant stress responses. Physiol Plant. 143:1–9.

Associate editor: Judith Mank

Downloaded from http://gbe.oxfordjournals.org/ at Iowa State University on October 8, 2015

Genome Biology and Evolution SMBE