

DR QIAN-HAO ZHU (Orcid ID : 0000-0002-6505-7417)

Article type : Original Article

Expansion of *MIR482/2118* by a Class II transposable element in cotton

Enhui Shen^{1,2†}, Tianzi Chen^{3†}, Xintian Zhu¹, Longjiang Fan¹, Jie Sun⁴, Danny J. Llewellyn⁵, Iain Wilson⁵ and Qian-Hao Zhu^{5,*}

¹Institute of Crop Sciences and Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China.

²New Rural Development Institute, Zhejiang University, Hangzhou 310058, China.

³Provincial Key Laboratory of Agrobiolgy, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China.

⁴Key Laboratory of Oasis Eco-agriculture, College of Agriculture, Shihezi University, Shihezi, Xinjiang 832000, China.

⁵CSIRO Agriculture and Food, Black Mountain Laboratories, GPO Box 1700, Canberra, ACT 2601, Australia.

[†]Enhui Shen and Tianzi Chen contributed equally to this work.

*Correspondence: Qian-Hao Zhu, +61 2 62464903; qianhao.zhu@csiro.au

Enhui Shen: ttxsenhui@gmail.com

Tianzi Chen: actzi0503@gmail.com

Xintian Zhu: zhuxt6040@163.com

Longjiang Fan: fanlj@zju.edu.cn

Jie Sun: sunjie@shzu.edu.cn

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/TPJ.14885](#)

This article is protected by copyright. All rights reserved

Danny J. Llewellyn: danny.llewellyn@csiro.au

Iain Wilson: iain.wilson@csiro.au

Qian-Hao Zhu: qianhao.zhu@csiro.au

Running title: TE-capture drove expansion of *MIR482/2118*

KEYWORDS: miR482/2118, *NBS-LRR*, miRNA evolution, transposable element, gene duplication, cotton, *Gossypium* spp.

SUMMARY

Some plant miRNA families contain multiple members generating identical or highly similar mature miRNA variants. Mechanisms underlying expansion of miRNA families remain elusive, although tandem and/or segmental duplications have been proposed. In this study in two tetraploid cottons *Gossypium hirsutum* and *G. barbadense*, and their extant diploid progenitors *G. arboreum* and *G. raimondii*, we investigated the gain and loss of members of the miR482/2118 superfamily, which modulates the expression of nucleotide-binding site leucine-rich repeat (*NBS-LRR*) disease resistance genes. We found significant expansion of *MIR482/2118d* in *G. raimondii*, *G. hirsutum* and *G. barbadense*, but not in *G. arboreum*. Several newly expanded *MIR482/2118d* loci have mutated to produce different miR482/2118 variants with altered target gene specificity. Based on detailed analysis of sequences flanking these *MIR482/2118* loci, we found that this expansion of *MIR482/2118d* and its derivatives has resulted from an initial capture of a *MIR482/2118d* by a Class II DNA transposable element (TE) in *G. raimondii* prior to the tetraploidisation event followed by transposition to new genomic locations in *G. raimondii*, *G. hirsutum* and *G. barbadense*. The “GosTE” involved in capture and proliferation of *MIR482/2118d* and its derivatives belongs to the *PIF/Harbinger* superfamily generating a 3-bp target site duplication upon insertion at new locations. All orthologous *MIR482/2118* loci in the two diploids were retained in the two tetraploids, but mutation(s) in miR482/2118 were observed across all four species as well as in different cultivars of both *G. hirsutum* and *G. barbadense*, suggesting a dynamic co-evolution of miR482/2118 and its *NBS-LRR* targets. Our results provide fresh insights into the mechanisms contributing to *MIRNA* proliferation and enrich our knowledge on TEs.

INTRODUCTION

miRNAs are 21-24 nucleotide-long small non-coding RNAs generated from *MIRNA* genes that, like protein-coding genes, are transcribed by RNA polymerase II (Voinnet, 2009). miRNAs are key regulators of gene expression and play important roles in plant development and stress responses (Sunkar *et al.*, 2012; D’Ario *et al.*, 2017; Li *et al.*, 2017). Most plant *MIRNAs* are found in intergenic regions, but intronic *MIRNAs*, including a unique type termed *MIRTRON*, have been reported (Zhu *et al.*, 2008). *De novo MIRNAs* can arise from inverted duplication of target gene fragments (Allen *et al.*, 2004; Fahlgren *et al.*, 2007), spontaneous fold-back of self-complementary sequences (de Felippes *et al.*, 2008) or from miniature inverted-repeat transposable elements (MITEs) (Piriyapongsa and Jordan, 2008). Some plant *MIRNA* families, especially the highly conserved ones, contain multiple members producing identical or slightly

different mature miRNA sequences. Members of the same *MIRNA* family, such as *MIR156*, can be dispersed across different chromosomes (Wang *et al.*, 2007), or in some cases, such as *MIR395*, can be clustered, with different cluster sizes and intergenic distances (Guddeti *et al.*, 2005). These dispersed and/or clustered *MIRNAs* are thought to have been generated from tandem and/or segmental duplications followed by diversification through genetic drift (Maher *et al.*, 2006; Wang *et al.*, 2007).

Transposable elements (TEs) make up a large proportion of plant genomes. They play a vital role in shaping genome structure and genetic diversity through replication and transposition (Bennetzen and Wang, 2014). TEs are classified into two classes. Class I TEs or retrotransposons that transpose using a 'copy-and-paste' strategy by generating RNA intermediates, and Class II TEs or DNA transposons that transpose using the 'cut-and-paste' mechanism via DNA intermediates (Wicker *et al.*, 2007). Based on studies on gene duplication mediated by Class I TEs, it is thought that duplication of gene fragments by TEs is widespread in eukaryotic genomes. By contrast, relatively few studies have shown that Class II TEs are involved in capture and reshuffling of genomic and gene fragments (Stefan and Jiang, 2018). One such study in rice showed that fragments from more than 1000 genes had been captured in *Mutator*-like TEs (Jiang *et al.*, 2004). Lack of studies on gene duplication mediated by DNA transposons is largely due to the difficulties associated with confident identification of such duplication events. Genes duplicated by retrotransposons, or retrocopies, usually have key features in common that can be used in their identification and characterisation. By contrast, DNA transposon super-families do not share common features and due to genetic drift the intact structure of DNA transposons are usually recognisable only for relatively young duplicates. Therefore, unambiguous identification of gene duplications caused by DNA transposons relies on the presence of their intact structures, including their characteristic terminal inverted repeat (TIR) and target site duplications (TSDs) (Stefan and Jiang, 2018).

The genus *Gossypium* contains more than 50 species, including two cultivated diploids *G. herbaceum* (Ghe, A₁ genome) and *Gossypium arboreum* (Gar, A₂ genome), and two cultivated allotetraploids *G. hirsutum* (Ghr, AD₁ genome) and *G. barbadense* (Gba, AD₂ genome). The two cultivated allotetraploids originated in the New World from hybridisation between an A-genome progenitor closely resembling Gar or Ghe, with a D-genome progenitor closely resembling *G. raimondii* (Gra, D₅ genome) approximately 1-2 million years ago (MYA) (Wendel, 1989; Wendel and Cronn, 2003). It is estimated that Gra and the ancestor of the extant A-genome species diverged approximately 6.8 MYA (Cronn *et al.*, 2002). Ghr and Gba were independently domesticated, and produce high yield and superior quality fibre, respectively. The genomes of Gar, Gra, Ghr and Gba have been sequenced, providing insights into the origin and evolution of the allotetraploids (Paterson *et al.*, 2012; Zhang *et al.*, 2015; Du *et al.*, 2018; Wang *et al.*, 2019). These genome sequences also provide resources and an opportunity for investigations on the evolution of gene families, including *MIRNA* families.

MIR482/2118 is a unique *MIRNA* that produces two mature miRNA isoforms, miR482 and miR2118 (both 22-nt) with 20-nt overlapping sequence, from a common precursor RNA. It is an ancient superfamily that emerged first in the Gymnosperms (de Vries *et al.*, 2015; Xia *et al.*, 2015). In dicots and gymnosperms, miR482/2118 members co-ordinately regulate many nucleotide-binding site leucine-rich repeat (*NBS-LRR*) disease resistance genes by targeting their conserved P-loop domains (Zhai *et al.*, 2011; Li *et al.*, 2012; Shivaprasad *et al.*, 2012; Zhu *et al.*, 2013; Xia *et al.*, 2015; Canto-Pastor *et al.*, 2019) and are thought to be continually co-evolving with their resistance gene targets (Gonzalez *et al.*, 2015; Zhang *et al.*, 2016). miR482/2118 post-transcriptionally represses the transcript levels of *NBS-LRR* genes, but upon infection of viral, bacterial and fungal pathogens miR482/2118 levels are themselves repressed to relieve the suppression of their target disease resistance genes (Shivaprasad *et al.*, 2012; Zhu *et al.*, 2013; Canto-Pastor *et al.*, 2019), suggesting a critical role for the miR482/2118-*NBS-LRR* regulatory module in the balance of disease resistance and fitness of plants. Like most other plants, cotton contains hundreds of *NBS-LRR* genes with approximately 12% of them being targets of miR482/2118 (Zhu *et al.*, 2013). To better understand the impact of polyploidisation on the evolution of miR482/2118 and to potentially improve this crop's productivity by manipulating the miR482/2118-*NBS-LRR* regulatory network, a prerequisite is to have a comprehensive inventory of all the *MIR482/2118* and their target *NBS-LRR* genes in cotton.

In this study, we did a genome-wide investigation of the *MIR482/2118* family in the four cotton species (Gar, Gra, Ghr and Gba), and found that, compared to Gar, the other three cotton species had experienced significant expansion of *MIR482/2118d*. We did detailed analyses of sequences flanking the expanded *MIR482/2118d* and its derivatives and found that capture of a *MIR482/2118d* by a Class II DNA transposon followed by its transposition, *i.e.* gene duplication by TEs, is the underlying mechanism responsible for expansion of *MIR482/2118d*. The TEs involved in proliferation of *MIR482/2118d* belong to the *PIF/Harbinger* superfamily based on the similarity in their terminal inverted repeats (TIRs), sub-terminal sequences and target site duplication (TSD). Our results provide an example of *MIRNA* gene amplification by DNA transposons.

RESULTS

***MIR482/2118d* experienced significant expansion in Gra and the two tetraploids**

To understand the impact of polyploidisation on the evolution of the *MIR482/2118* family and their *NBS-LRR* target genes, we performed genome-wide analyses of miR482/2118 in two diploid (Gar and Gra) and two tetraploid (Ghr and Gba) cotton species. We identified 9 and 19 *MIR482/2118* loci in Gar and Gra, respectively, and 36 (16 and 20 from the At and Dt sub-genome, respectively) and 41 (16 and 25 from the At and Dt sub-genome, respectively) in Ghr and Gba, respectively (Figure 1, Table S1). Synteny analysis

based on annotated genes surrounding each *MIR482/2118* locus was performed to investigate their evolutionary relationships. Orthologs of each of the 9 Gar *MIR482/2118* loci (*MIR482/2118c, d, e, f, g, h, i, k, l*) were found in the At sub-genomes of Ghr and Gba, although ghr-miR482/2118c_A04 had at some point mutated to ghr-miR482/2118a_A04 (Figures 1, 2, S1a). Based on its surrounding genes and pre-miRNA, Gar-miR482/2118i was likely to have been generated by a single point mutation from an ancestral Gar locus orthologous to *MIR482/2118b* which is present in the tetraploids on Ghr-A13 (on scaffold3700 that was assigned to A13) and Gba-A13 (Figure S1, Table S2).

Each of the Gar *MIR482/2118* loci has an ortholog in Gra (Figure 2, Table S2), suggesting they were present in the common ancestors of Gar and Gra and represent the ancestral diploid state. In addition to these 9 orthologous *MIR482/2118* loci, Gra has 4 additional *MIR482/2118d* loci and 6 *MIR482/2118* loci producing three different mature miRNA sequences (miR482/2118j, o, p) that were not present in Gar. The precursors of miR482/2118j, o, p are almost identical to that of miR482/2118d (Figure S1c), indicating that these *MIR482/2118* loci are closely related to each other. The discrepancy of these *MIR482/2118* loci between Gra and Gar could be a result of gains in Gra or losses in Gar after the divergence of Gar and Gra from their common ancestor. In view of the presence of orthologs of all 9 Gar *MIR482/2118* loci in Gra and the close relationship amongst the 10 *MIR482/2118* loci present in Gra but absent in Gar (Figure S1c, Table S2), it is unlikely that all orthologs of the 10 *gra-MIR482/2118d, j, o, p* loci had been lost in Gar. We thus favour the conclusion that these 10 *gra-MIR482/2118d, j, o, p* loci were newly generated in Gra (see next section for more details of the proposed mechanism for this expansion).

According to synteny analysis, the Dt sub-genomes of Ghr and Gba have inherited orthologs of the 9 ancestral *MIR482/2118* loci from Gra, but not *MIR482/2118j, o, p*. It is possible that the tetraploids lost these orthologs *MIR482/2118j, o, p* after the polyploidisation event but it is most likely that these *MIR482/2118j, o, p* loci were newly evolved in Gra at some time after the polyploidisation event (Figures 2, S2, Table S2). Compared to Gar and Gra, many more new *MIR482/2118d* loci were found in Ghr (6 and 5 on At and Dt sub-genome, respectively, 1 on a scaffold without chromosome location information) and Gba (5 and 9 on At and Dt sub-genome, respectively) (Table S2), indicating that *MIR482/2118d* has experienced significant expansion in the two tetraploids and that all *MIR482/2118* loci found on the At subgenome (except the ones orthologous to *gar-MIR482/2118d_Ch05*) were newly generated after tetraploidisation.

TE-capture of a *MIR482/2118d* in Gra contributed to its expansion

Significant expansion of *MIR482/2118d* was observed in Gra and the two tetraploids, but between the two diploids, expansion was found only in Gra, but not in Gar, we therefore reasoned that the expansion must have occurred in Gra after the divergence of Gar and Gra (also refer to the reasons mentioned above),

but before the tetraploidisation event, and carried on in the two tetraploids after tetraploidisation giving each a unique complement of both shared and different *MIR482/2118d* members. To uncover the underlying mechanism for the expansion, we extracted 10-kb sequences flanking each of the *gra-MIR482/2118d*, *j*, *o*, *p* (11 in total) loci and identified the longest homologous sequence segment between every two loci based on pair-wise comparisons. The longest homology (~5700 bp) was found between *gra-MIR482/2118d_Ch08* and *gra-MIR482/2118p_Ch013* (Figure 3a). *Gra-MIR482/2118d_Ch09* shared the shortest homology (~220 bp) with other sequences. Sequences homologous to the ~220-bp around *gra-MIR482/2118d_Ch09* were found in each of the other 10 sequences, whose sequence homology could be extended a further ~300-bp upstream (Figure S3). At the 3' ends, ~220-bp homologous sequences were observed in 9 of the 10 sequences (except *gra-MIR482/2118d_Ch013*) (Figure S3), despite their variable sequence length and internal sequence homology. Interestingly, 15-bp imperfect TIRs were found at the 5' and 3' ends of those 9 sequences, although the pair of TIRs of *gra-MIR482/2118p_Ch013* were more divergent than the others due to numerous mutations in the 3' TIR (Figure 3b). Flanking each pair of TIRs we found a 3-bp target site duplication (TSD, usually TTA or TAA) (Figure 3), indicating that these *miR482/2118*-containing sequences are Class II DNA TEs (Feschotte and Pritham, 2007). We then checked for the presence of TIRs and TSD in the sequences flanking *miR482/2118d* and its derivatives in *Ghr* and *Gba*. TIRs identical or similar to those found in *Gra* as well as the signature 3-bp TSD were also found in the vast majority of the sequences containing *miR482/2118d* and its derivatives in these two other cotton species (Tables S3). These sequences were defined as TEs-with-*miR482/2118*. The TIRs and sub-terminal sequences of these TEs-with-*miR482/2118* showed high similarity to those of the *PIF/Harbinger* superfamily TEs that usually also have a TSD of TTA or TAA (Feschotte and Pritham, 2007; Fattash *et al.*, 2013). These results indicate that expansion of *MIR482/2118d* and its derivatives was likely to be the result of transposition of one or more of these TEs to multiple new genomic locations.

The *Gar* genome contains only a single *MIR482/2118d* locus (*gar-MIR482/2118d_Ch05*), whose ortholog in *Gra* is *gra-MIR482/2118d_Ch09* that is not associated with a pair of 5' and 3' TIRs (Figures 3a, S4a). *Gar-miR482/2118d_Ch05* and *gra-miR482/2118d_Ch09* are located ~210-bp downstream of *gar-miR482/2118e_Ch05* and *gra-miR482/2118e_Ch09*, respectively, *i.e.* *MIR482/2118d* and *MIR482/2118e* are arranged in a tandem configuration in both *Gar* and *Gra*, which is conserved in *Theobroma cacao* (Figure S5), with which cotton shared a common ancestor ~60 MYA (Paterson *et al.*, 2012). Thus, *gra-MIR482/2118d/e_Ch09* is likely to be the ancestral locus from which other *gra-MIR482/2118d* loci were derived after its capture (a ~600 bp segment containing *MIR482/2118d* but not *MIR482/2118e*) by a TE on another chromosome through the mechanism of double-strand break (DSB) repair (see a model in Discussion). We estimated that the capture happened around 1.4 MYA, not long before the tetraploidisation event that was proposed to have occurred 1-2 MYA (Wendel, 1989; Wendel

and Cronn, 2003). The TE involved in the capture was likely to have been on chromosome 13, which generated the TE containing *gra-MIR482/2118d_Chr13* or *gra-MIR482/2118p_Chr13* (*MIR482/2118p* was mutated from *MIR482/2118d* after the capture event), as other *MIR482/2118d* loci were estimated to be less diverged from *gra-MIR482/2118d_Chr09* than *gra-MIR482/2118p_Chr13* and *gra-MIR482/2118d_Chr13*. The homologous sequence (~30 bp) at the 3' end of the donor (*gra-MIR482/2118d_Chr09*) and the potential DSB site on the chromosome 13 TE (acceptor) might have played a role in the capture of *gra-MIR482/2118d_Chr09* (Figure S4b). Transposition of the *gra-MIR482/2118d*-containing TE generated other *gra-MIR482/2118d* and *gra-MIR482/2118j, o, p* due to point mutations in the mature miR482/2118 sequences.

Based on comparison of the TEs-with-miR482/2118 in Gra, Ghr and Gba (Tables S2, S3), we found: 1) orthologs of all four of the newly evolved *gra-MIR482/2118d* were found in Ghr and Gba, implying their expansion before the tetraploidisation event; 2) *ghr-* and *gba-MIR482d_D13b* are the orthologs of *gra-MIR482/2118p*, implying mutation of *gra-MIR482/2118p* from *MIR482/2118d* in Gra after the tetraploidisation event; 3) at least three *MIR482/2118d* loci (*MIR482/2118d_A06*, *_A11* and *_D03*) have orthologs in Ghr and Gba but are not found in Gra and Gar, suggesting that they might have proliferated after the tetraploidisation event but before divergence of Ghr and Gba, or that the proliferation had occurred in Gra before the tetraploidisation event but the corresponding ortholog was lost in Gra (possible only for *MIR482/2118d_D03*); 4) most *ghr-* and *gba-MIR482/2118d*-containing TEs, including all those found on the At sub-genome, were new to Ghr and Gba, suggesting a continuous expansion of *MIR482/2118d* after tetraploidisation.

The proliferated TEs containing *MIR482/2118d* are only a small portion of a larger GosTE family

We designated the TEs containing *MIR482/2118d* or its derivatives as the “GosTE” family. They could be classified into 23 subfamilies (GosTE1 to GosTE23) based on their combined 5' and 3' TIR identities (Tables 1, S3). We used two criteria, fully matching 5' and 3' TIRs to those found in any of the GosTE1 to GosTE23 TEs and a total length between the TIRs of <10-kb, to identify potential TEs-without-miR482/2118. The candidate TEs were separated into two groups (Group I: GosTE1 to GosTE23; Group II, GosTE24 to GosTE46) depending on whether or not they have the pair of 5' and 3' TIRs identical to those of GosTE1 to GosTE23 (see Experimental procedures for details). Of the Group I subfamilies (*i.e.* GosTE1 to GosTE23), 9 have members without miR482/2118 identified in Gra, Ghr or Gba but none in Gar, and 7 of them have >10 GosTE members with the highest number observed in GosTE11. In both Ghr and Gba, the majority of the Group I TE members were found in the Dt sub-genome. For the subfamilies with <10 members, no TE member was identified in the At subgenome (Tables 1, S3). Of the 23 Group II subfamilies (GosTE24 to GosTE46), 3, 13, 16 and 14 were identified in Gar, Gra, Ghr and Gba, respectively. Gar has fewer subfamilies and total numbers of GosTEs than the other three cottons. Of the 51

Group II GosTE members found in Ghr (26) and Gba (25), only 3 (~6%) were present in the At sub-genome, consistent with the results of fewer At subgenome TEs in Group I subfamilies (GosTE1 to GosTE23) (Table 1). These observations largely agreed with the distribution of the 5' and 3' TIRs in the A and D genomes as well as in the At and Dt sub-genomes (Table S4).

The number of TE members in each subfamily within a genome/subgenome varied significantly, from 0 to 237 (GosTE11 in Gra) (Table 1). In total, 7 subfamilies (~15%) contain 10 or more TE members and all of them (GosTE7, 8, 9, 10, 11, 15 and 23) are from Group I, while ~37% (17/46) of subfamilies contained only a single member (Table 1). On average, TEs-without-miR482/2118 are much shorter (1166 bp) than TEs-with-miR482/2118 (3221 bp). The majority TEs-without-miR482/2118 have a length of either ~620 bp or ~3000 bp, with the first being the most predominant, largely because, of the 7 subfamilies containing 10 or more TE members, 5 subfamilies contain the majority of members having a size of ~620 bp (Figure 4, Table S3). The shorter TEs (~620 bp) are typical miniature inverted-repeat transposable elements (MITEs), whose proliferation, as reported in other species (Lu *et al.*, 2012; Fattash *et al.*, 2013), is (sub-)family dependent. In support of their MITE identity, the internal sequences of GosTE11 members are all highly similar. These sequences are also highly related to the similar sized members from other GosTE subfamilies (Figure S6), suggesting that all these TEs could have a common origin. The GosTE11 Subfamily, which contains the highest number of TE members, has 0, 3 and 6 TEs-with-miR482/2118 in Gra, Ghr and Gba. The sizes of the GosTE11 members with-miR482/2118 in Ghr and Gba are 3300 bp and ~2844 bp, respectively (Table S3), which is much longer than that of the typical TEs of the same Subfamily (~620 bp). Because an ortholog of these TEs-with-miR482/2118 was not found in Gra, they have most likely been generated by sequence recombination between TE member(s) of GosTE11 and TEs-with-miR482/2118 of other GosTE subfamilies after the tetraploidisation event. As with the TEs-with-miR482/2118, the majority of TEs-without-miR482/2118 in all GosTE subfamilies also had a pair of TSDs, either TTA or TAA (Table S3).

Identification of a large number of GosTEs and a subfamily with hundreds of members suggest that some of these GosTEs were active in the history of cotton evolution and/or are still active, *i.e.* some members of GosTE could be autonomous and capable of encoding transposase. Using the approaches described in the Experimental procedures, we identified 12 candidate autonomous GosTEs (Gra-GosTE2-1, Gra-GosTE7-2, Gra-GosTE33-4, Gra-GosTE11-181, Gra-GosTE11-25, Ghr-GosTE7-7, Ghr-GosTE29-2, Ghr-GosTE35-1, Ghr-GosTE35-2, Ghr-GosTE40-1, Gba-GosTE11-148, Gba-GosTE35-3) with 5, 5 and 2 in Gra, Ghr and Gba, respectively (Figure S7, Table S5). Among these candidate autonomous GosTEs, three pairs (Gra-GosTE33-4 and Ghr-GosTE7-7; Gra-GosTE11-181 and Gba-GosTE11-148; Ghr-GosTE35-2 and Gba-GosTE35-3) are orthologs, but Gra-GosTE33-4 and Ghr-GosTE7-7 were assigned to different subfamilies due to a mismatch in their 5' TIRs (Table S3). Four more candidate autonomous

GosTEs (Gra_Chr11a, Gra_Chr08, Gba_D05 and Gba_D10), which were not identified as GosTE members due to mutation(s) in TIRs (Gba_D10 was shown in Figure S8 as an example), were identified based on ortholog searches which brought the total number of candidate autonomous GosTEs to 16. Orthologs of several of the 16 candidate autonomous GosTEs, such as Gra-GosTE28-1 and Gra_Chr11b, were found in the corresponding cotton species but were not considered as a candidate autonomous GosTEs due to lack of an intact DDE domain or absence of the MYB-domain that is present in all the candidate autonomous GosTEs. Gra-GosTE7-2, Gra-GosTE11-25 and Ghr-GosTE29-2 could be newly generated in the corresponding cotton species as their orthologs were not found in other cotton species (Table S5).

Mutations in miR482/2118 were found not only at the species-level but also at the cultivar-level

As mentioned above, single nucleotide changes were observed in the mature miR482/2118 sequences that turned miR482/2118c_A04 into miR482/2118a_A04 in Ghr and miR482/2118b into miR482/2118i in Gar. Synteny in the surrounding genes between Gra and the two tetraploids as well as the highly similar precursor sequences indicate that *MIR482/2118p* was also mutated from one of the *MIR482/2118d* loci on chromosome 13 in Gra (Figures 1, S1a). Two unique miR482/2118 members (miR482/2118m, n) were found in Gba and were likely also derived from point mutations of duplicated miR482/2118d (Figures 1, S1c). Furthermore, a single nucleotide insertion in gba-miR482/2118l_D12 has probably made this homoeolog dysfunctional (Figure S1e). In addition to these mutations found in the mature miR482/2118 sequences, mutations were also found in other parts of pre-miR482/2118. For example, the sequence changes found in the pre-miR482/2118g-A03 in Gba would prevent proper processing of miR482/2118g and thus this locus may also not now be functional (Figure S1f).

We examined expression of different miR482/2118 members across different tissues and pathogen challenges in available cotton small RNA datasets (Figure 5, Table S6) and found considerable variation in expression levels of different miR482/2118 members across the species with many, particularly those that are more recently derived, being very lowly expressed or not detected. In theory, each *MIR482/2118* locus would produce two isoforms, miR482 and miR2118, with a 20-bp overlapping sequence (Figure 1). Indeed, some *MIR482/2118* loci generate both miR482 and miR2118 at relatively high levels in all four species. However, for most loci, miR482 and miR2118 tend to be differentially expressed, such as *MIR482/2118c*, *e*, *g*, *l* (Figure 5), potentially suggesting a diverged role for the miR482 and miR2118 produced from the same locus.

For the species-specific miR482/2118 members, such as ghr-miR482/2118a_A04, gar-miR482/2118i, gra-miR482/2118p, gba-miR482/2118m_A05, gba-miR482/2118n_D11 and gba-miR482/2118q_D05, we would not expect to detect their expression in the species in which they have not been identified in genomic sequences. However, we detected very low levels of miR482a, miR2118b,

miR482/2118i, miR2118n and miR2118q in cotton species that do not apparently contain these loci (boxed in Figure 5). These are therefore likely to be artefacts resulting from sequencing errors except for miR2118q in Ghr as it was detected in all samples from the Ghr cultivar Siokra 1-4 but not in other Ghr accessions, including TM-1 and MCU-5 (Table S6). To make sure detection of miR2118q in Siokra 1-4 was not due to sequencing errors, we searched for miR2118q in the genomes of TM-1, MCU-5 and Siokra 1-4, and only found a full match in Siokra 1-4 at a locus generating miR482/2118e (chromosome D05) but did not detect it in the genome sequences from TM-1 and MCU-5. Similarly, gba-miR482/2118e_D05 has mutated to gba-miR482/2118q_D05 in accessions Pima S-7 and 3-79 but it was not found in cultivar Hai7124 (Figures 1, S1d). Together, these results indicate that there are both species-level differences and cultivar-level differences in the presence and expression of some miR482/2118 members in cotton, probably a result of the dynamic interaction with their target genes, particularly *NBS-LRR* genes that can also vary in sequence between cultivars.

Point mutations in miR482/2118 diversified the regulatory capacity of the *MIR482/2118* family

Most predicted targets (50-77%) of miR482/2118 in the cotton species examined are *NBS-LRR* genes related to disease resistance (Table S7). Mutations in the mature sequences of miR482 and/or miR2118 would potentially change their interactions with *NBS-LRRs*, and consequently disease responses. Based on an analysis of predicted target genes, we found simultaneous gain and loss, and net gain or net loss for the targets of the newly evolved miR482/2118 members (Figure 6). The effect of loss of targets in many cases would be minimal because 1) most new miR482/2118 members (miR482/2118j, m, n, o, p) found in Gra, Ghr and/or Gba were mutated forms of miR482/2118d that are already generated from multiple loci; and 2) as both Ghr and Gba are tetraploids, loss of a functional *MIR482/2118* from one subgenome would not affect production of the miRNA from its homoeologous locus. The most significant effect of mutations in the mature sequences of miR482/2118 would thus be in having more genes recruited to regulation by the *MIR482/2118* family, as more than half of the newly evolved miR482/2118 members are predicted to have acquired new targets and 10 of the 28 gained targets are *NBS-LRR* genes (Figure 6). For example, all the 4 gained targets due to the mutation from gar-miR2118b to gar-miR2118i are *NBS-LRRs*. Mutation from ghr-miR482c to ghr-miR482a would have lost 6 *NBS-LRRs*, but gained 3 new ones (Figure 6, Table S7). These results suggest that gain of targets might be a driving force for retention and expression of the mutated miR482/2118 members. For the newly evolved miR482/2118j, m, n, o, p, gain of target(s) was observed, but their expression was not detected in any of the small RNA datasets used in this study (Figures 5, 6). This is probably due to their extremely low expression levels, and detection of their presence requires a much deeper sequencing depth than the datasets used in this study. Alternatively, it suggests the regulatory relationships between these new miR482/2118 members and their targets have not been established and are yet to be fixed by positive selection. Gar-miR482/2118i and ghr-miR482a, on the

other hand, were similar or higher in expression to some of the more ancestral miR482/2118 members (Figure 5), suggesting they are now fully functional and integrated with their gained regulatory targets. For example, we could demonstrate that *Gh_A12G0912*, a homolog of *Arabidopsis TAO1* (target of *AVRB operation1*), a *TIR-NBS-LRR* disease resistance gene induced by the *Pseudomonas syringae* effector AvrB, was a miR482a-specific target and is cleaved by ghr-miR482a to produce 21-nt phased small RNAs (Figure S9).

DISCUSSION

Several duplication mechanisms, including whole genome duplication, tandem duplication, segmental duplication and TE-mediated duplication have been proposed to explain the presence of multigene families in higher plant genomes (Freeling, 2009). Regarding the origins of multiple members of certain *MIRNA* families, tandem gene duplications and segmental duplications have been proposed (Maher *et al.*, 2006; Wang *et al.*, 2007). In this study, based on analyses of syntenic relationship of *MIR482/2118d* and its derivatives among cotton species as well as of sequences flanking the duplicated *MIR482/2118d* and its derivatives, we showed that TE-mediated duplication is responsible for the significant expansion of *MIR482/2118d* in the two cultivated tetraploid cotton species, Ghr and Gba, and their diploid progenitor Gra. The initial capture of *MIR482/2118d* by a Class II TE (GosTE) occurred ~1.4 MYA in Gra, likely shortly before the tetraploidisation event that is thought to have occurred 1-2 MYA (Wendel, 1989; Wendel and Cronn, 2003). The *MIR482/2118d* member on Gra chromosome 9 and a GosTE on Gra chromosome 13 appear to be the donor and the acceptor of the capture event, respectively. We propose that the capture was driven by DSB repair, which occurs frequently in TEs due to their higher than normal recombination frequency (Wicker *et al.*, 2010). Transposition of the newly generated TE-with-miR482/2118d and/or the transposed copies of the TEs-with-miR482/2118d contributed to expansion of *MIR482/2118d* in Gra, and subsequently in Ghr and Gba after tetraploidisation (before and/or after divergence of Ghr and Gba), spreading new *MIR482/2118d* to both At and Dt sub-genomes (Figure 7). DSB repair, mainly via non-homologous end joining (NHEJ) that relies on microhomology (<10 bp) between the flanking sequences of the acceptor sites and the donor sites, has been proposed to be the main mechanism responsible for capture and movement of genomic and gene fragments (Wicker *et al.*, 2010; Chen *et al.*, 2013). We observed an ~30-bp homologous fragment at the 3' junction of the acceptor (Gra chromosome 13) and the donor (Gra chromosome 9) regions (Figure S5b), it might have played a role in the DSB-repair-triggered capture of *MIR482/2118d*.

Our knowledge of gene duplication mediated by TEs has predominantly been based on studies on retrotransposons, and only a few studies have reported gene movement mediated by DNA transposons (Jiang *et al.*, 2004; Wicker *et al.*, 2010; Chen *et al.*, 2013; Ferguson *et al.*, 2013; Morata *et al.*, 2018). Even in these reported studies, most duplicated loci had not retained the intact structure of the TEs, including the

TIRs and TSD. For example, Wicker and colleagues identified 10 and 20 TE-mediated gene duplication events in *Brachypodium* and rice, respectively, and of these 30 events, the TSD was found only in four (Wicker *et al.*, 2010). Furthermore, none of the duplicated genes reported so far have been *MIRNA* genes. Here, we provided sound evidence for the involvement of DNA transposons in the expansion of *MIR482/2118d*. In total, there are 10, 18 and 22 newly evolved *MIR482/2118d* and its derivatives in *Gra*, *Ghr* and *Gba*, respectively (Table S2). Of these 50 *MIR482/2118* loci generated by transposition of GosTEs, 41 had both intact TIRs and TSDs identified (Table S3). The TEs-with-miR482/2118 together with their related TEs-without-miR482/2118 form a large GosTE family with 46 subfamilies and 981 TE members (Table 1), of which the number of TEs-with-miR482/2118 is only a small portion (~5%) and only 16 are candidate autonomous elements (Table S5). We believe that the total number of GosTE members is underestimated based on the observation that some orthologs of the candidate autonomous elements were not identified as a GosTE member due to the use of the criterion of TIRs identical to those found in GosTE1 to GosTE46.

The GosTE family belongs to the *PIF/Harbinger* superfamily with 3-bp TSD (Table S3) that has been reported in several plant species, but no member of this superfamily has been previously shown to be involved in *MIRNA* proliferation (Jiang *et al.*, 2003; Zhang *et al.*, 2001, 2004; Grzebelus *et al.*, 2007; Markova *et al.*, 2015; Hsu *et al.*, 2019). GosTEs seem to be more active in the D and Dt (sub-)genomes than in the A and At (sub-)genomes (Table 1). For instance, *Gar* (A genome) contains only 5 TEs-without-miR482/2118 from 3 different subfamilies (~6.5% of the 46 subfamilies). Importantly, the TIRs of the GosTE subfamilies (GosTE7, 8, 9, 10, 11, 15 and 23) with more than 10 members (including both TEs-with-miR482/2118 and TEs-without-miR482/2118) had 0-3 hits in the A genome (Table S4), supporting a lack of expansion of these GosTE elements in the A genome. Consistently, in both *Ghr* and *Gba*, these TIRs had fewer hits in the At subgenome than in the Dt subgenome, with only 3% (37/289, *Ghr*) to 19% (75/386, *Gba*) of GosTEs being found in the At subgenome (Tables 1, S4). Further supporting this observation, no candidate autonomous GosTE element was identified in the A and At (sub)genomes. These results suggest that despite transposition of GosTEs from the Dt subgenome to the At subgenome after unification of the A and D genomes in a single nucleus, most transpositions of GosTEs still occurred within the subgenome where the transposase was encoded.

Some GosTE subfamily contains only a single member whereas seven subfamilies contain more than 10 members, especially GosTE11 that contains hundreds of members in *Gar*, *Ghr* and *Gba* (Table 1), suggesting different transposition activity of each GosTE subfamily that could be associated with its TIR sequence composition and the homology between the 5' TIR and 3' TIR. The active nature of GosTE11 suggests that its 5' and 3' TIRs are favourable to the transposase driven transposition of GosTEs. In addition, the 5' and 3' TIRs of GosTE11 are the reverse complement to each other, *i.e.* are a pair of perfect

TIRs (Table S4). This might contribute to its high transposition activity. Subfamily GosTE8 with 43 members also has a pair of perfect 5' and 3' TIRs that has only a single base pair difference compared to those of GosTE11. Of the 7 GosTE subfamilies with more than 10 members, 5 had most of their members with a size of ~620-bp, a typical length of MITEs, consistent with previous findings that MITE-like elements show a stronger transposition activity than autonomous elements (Jiang et al., 2003).

Several observations suggest that the GosTE family has been evolving rapidly by accumulation of mutation(s) in the TIRs and indels in the internal regions. In Gra, the 10 TEs-with-miR482/2118d, j, o, p were derived from transpositions of the initial TE involved in the capture of *MIR482/2118d*, it would be expected that they would have the same TIRs as the initial TE, *i.e.* Gra-GosTE3-1 (with *MIR482/2118d_Chr13*) or Gra-GosTE2-1 (with *gra-MIR482/2118p_Chr13*). The 3' end of Gra-GosTE3-1 could not be decisively determined and the 5' TIR of Gra-GosTE3-1 is identical to that of Gra-GosTE6-1 but differs from all other 8 TEs-with-miR482/2118 in Gra. Both the 5' and 3' TIRs of Gra-GosTE2-1 are also unique among the 10 TEs-with-miR482/2118 in Gra. The other 3 miR482/2118d-containing TEs (Gra-GosTE1-1, Gra-GosTE4-1 and Gra-GosTE5-1) all have unique 5' and 3' TIRs as well. The 4 miR482/2118j-containing TEs have the same 5' and 3' TIRs but they are different from the others. The mutation(s) observed in the TIRs have had to have occurred after the initial TE capture event. They might have compromised or even enhanced TE mobility as most of these subfamilies have only a single member but some, such as Gra-GosTE6 and Gra-GosTE7, have multiple members. The length of Gra-GosTE2-1 is 5737 bp, slightly shorter than Gra-GosTE1-1 but much longer than most of the others (Table S3), suggesting that insertions and deletions have occurred in the initial TE involved in the capture event. Both Gra-GosTE6 and Gra-GosTE7 subfamilies contain elements with or without miR482/2118, implying the involvement of sequence recombination/rearrangements between different elements presumably due to the presence of homologous sequences within the different elements.

NBS-LRRs are characterised by their fast and dynamic evolution to combat challenges from diverse and fast evolving pathogens. To balance disease resistance and fitness costs, expression of *NBS-LRRs* must be under strict scrutiny in the absence of pathogens but respond rapidly in their presence. The ancient *MIR482/2118* family could be the key player for achieving this balance as 1) miR482/2118 members have been shown to be negative regulators of *NBS-LRRs*, but their expression is repressed upon pathogen attack (Li *et al.*, 2012; Shivaprasad *et al.*, 2012; Zhu *et al.*, 2013), and 2) evolution of miR482/2118 has been shown to be driven by diversification of the sequences of the P-loops of *NBS-LRRs* across many plant genomes (Zhao *et al.*, 2015; Zhang *et al.*, 2016). Here, we found that the cotton *MIR482/2118* family has proliferated and experienced multiple point mutations during the last 1-2 MYA after the polyploidisation event that produced the two main cultivated species, and that this variation continues today even at the cultivar-level. The mutations significantly altered the regulatory capacity of miR482/2118 with both gain

and loss of *NBS-LRR* targets (Figure 6). Our results suggest that miR482/2118 has been conserved in cotton before and after tetraploidisation but continued to diversify in both diploid and tetraploid species through mutation of the mature miR482/2118 sequences resulting in loss or acquisition of *NBS-LRR* targets. Our results also imply that the interaction landscape between miR482/2118 and *NBS-LRRs* has been dynamic and rapidly shaped over relatively short evolutionary times. The challenge for the future will be how we can use this knowledge to breed better disease resistant crops without yield and quality penalties.

Gene duplications as a result of whole genome and segmental duplication have been an important evolutionary route for gene neo- and sub-functionalisation (Panchy *et al.*, 2016; Cerbin and Jiang, 2018; Liang and Schnable, 2018). This is also true for *MIR482/2118* as 5 (miR482/2118j, m, n, o, p) of the 7 (miR482/2118a, j, m, n, o, p, q) newly evolved mature miR482/2118 variants identified in this study were spontaneous mutations from proliferated miR482/2118d, and their newly acquired targets included not only *NBS-LRRs*, but also other genes with diverse functions.

EXPERIMENTAL PROCEDURES

Cotton genomes and small RNA datasets

The genome sequences and gene annotations of Gar (Du *et al.*, 2018), Gra (Paterson *et al.*, 2012), Ghr (Zhang *et al.*, 2015) and Gba (Hu *et al.*, 2019) were downloaded from COTTONGEN (<https://www.cottongen.org/>). All known miR482/2118 mature sequences were downloaded from miRBase (<http://www.mirbase.org/>, release 22) or collected from the literature (Zhu *et al.*, 2013; Wang *et al.*, 2016). To analyse the expression levels of individual miR482 and miR2118, we downloaded 34 available small RNA datasets from the National Center for Biotechnology Information (NCBI), including 5 for Gar, 5 for Gra, 3 for Gba, and 21 for Ghr each from different tissues or pathogen challenges (Table S6). We also used 24 small RNA datasets generated from roots of Ghr cultivars MCU-5 and Siokra 1-4 infected with *Verticillium dahliae* and mock controls.

Prediction of miR482 and miR2118, and their targets

We applied two prediction strategies in order to identify a comprehensive set of *MIR482/2118* genes in the four cotton species. First, the assembled known plant miR482/2118 sequences were mapped to each genome with a maximum of 2-nt mismatches by PatMaN (Prüfer *et al.*, 2008), the flanking sequences of the matches were then captured using custom perl scripts and subjected to secondary hairpin structure prediction using Mireap (<http://sourceforge.net/projects/mireap/>) with the following parameters: the hairpin structure has a free energy lower than -18 kcal/mol, the space between miRNA and miRNA* is less than 300-nt, the number of matched nucleotides between miRNA and miRNA* are more than 16, and the bulges between miRNA and miRNA* contain fewer than 4 nucleotides (Shen *et al.*, 2015). Second, the miR482/2118 members identified in each cotton species using Mireap were further used in reciprocal

homology searching by local BLASTn to confirm the Mireap predictions and also to uncover any other members missed by Mireap prediction.

The targets of miR482 and miR2118 in the four cotton species were predicted using the online tool psRNATarget and the transcript files of each species (Dai *et al.*, 2017). The ‘Maximum expectation’ was set to 3.0, ‘Length for complementarity scoring (HSP size)’ was set to 22-nt and other parameters were as the defaults.

Genomic synteny and conservation of *MIR482/2118* loci

The synteny or collinearity of the *MIR482/2118* loci among the four cotton species was searched by MCScanX (Wang *et al.* 2012). The gff3 annotation files and coding sequences of the annotated genes were used for genomic synteny analysis. The synteny blocks with at least 10 genes were screened using the jcvf program downloaded from GitHub (<https://github.com/tanghaibao/jcvf>). Each *MIR482/2118* locus and its flanking 10 protein-coding loci were retrieved from the four cotton species and compared. A syntenic *MIR482/2118* pair was defined when the upstream or downstream of the *MIR482/2118* locus in the two compared cotton species belongs to the same synteny block. The jcvf program was also applied to visualize the overall and local syntenic relationship among the four cotton genomes. The Easyfig program was used in generating the image of Figure 3a (Sullivan *et al.*, 2011).

Identification of transposable elements (TEs) related to those containing miR482/2118

In Gra, Ghr and Gba, the sequences containing a miR482/2118 and a pair of 15-bp 5' and 3' terminal inverted repeats (TIRs) were designated TEs-with-miR482/2118 and assigned to subfamilies GosTE1 to GosTE23 based on their TIR identity. TEs with the same 5' and 3' TIRs were grouped into the same subfamily regardless of their species of origin (Table 1). To identify other TEs (TE-without-miR482/2118) related to TE-with-miR482/2118, we used all the 5' and 3' TIRs of the TE-with-miR482/2118 as queries to identify their exact matches in the four cotton genomes and retrieved the coordinates of those matches. The genomic sequences defined by a pair of flanking 5' TIR and 3' TIRs separated by <10 kb were considered as a candidate TEs because the size of the longest TE-with-miR482/2118 is 5795 bp (Gra-GosTE1-1 containing *gra-MIR482/2118d_Chr08*). The candidate TEs were then separated into two groups. Group I included those with a pair of 5' and 3' TIRs identical to those of the TEs-with-miR482/2118, and they were classed as members of subfamilies GosTE1 to GosTE23 but lacking miR482/2118. The remaining candidate TEs with a pair of 5' and 3' TIRs different from that of GosTE1 to GosTE23 were assigned to Group II (subfamilies GosTE24 to GosTE46). For the TEs-without-miR482/2118 having a pair of perfect 5' and 3' TIRs, their orientations were determined based on aligning them with TEs of known orientation using MEGA7 (Kumar *et al.*, 2016). TEs with a length <1 kb were considered as putative MITEs. The

superfamily of the (MI)TEs was determined based on blast against the P-MITE database (Chen *et al.*, 2014) using the TIRs and sub-terminal sequences of GosTEs as queries.

Identification of candidate autonomous GosTEs

We used two approaches to identify GosTEs containing potential transposase. Firstly, all GosTEs were subjected to ORF prediction using TransDecoder v5.5.0. The minimum protein length of ORFs was set to 50 amino acids. All predicted ORFs were searched against the PFAM database using HMMER v3.2.1 (Mistry *et al.*, 2013) to identify GosTEs with the signature DDE domain found in TPases of the *PIF/Harbinger* elements. Secondly, the TPase of the *PIF* element from *Zea mays* (Zhang *et al.*, 2004) was used as a query to search against the whole-genome sequences of the four cotton species using tblastn with a cut-off E value of 1.0×10^{-5} . The upstream and downstream regions of the hits were checked for the presence of a pair of TIRs identical to those of GosTE1 to GosTE46. GosTEs containing a DDE domain based on ORF prediction and a TPase hit based on tblastn were selected for detailed analyses of the DDE domain and other predicted ORFs. We found that most of these GosTEs have an ORF encoding the MYB-domain that was present in Zm-PIF (Zhang *et al.*, 2004). We thus considered the GosTEs with an intact DDE domain, *i.e.* having the three conserved D, D and E residues, and the MYB-domain as candidate autonomous elements.

Estimation of the capture time of *MIR482/2118d*

Because the initial capture event happened in *G. raimondii* before the tetraploidisation event, we used the TEs-with-miR482/2118 from *G. raimondii* in the analysis. The 10 *gra-MIR482/2118d* and its derivatives (Figure 3a) retained a 221-bp sequence fragment highly similar to the donor *gra-MIR482/2118d/e_Chr09*. The 221-bp fragment from all these *gra-MIR482/2118* was extracted, aligned using MUSCLE v3.8.15 (Edgar 2004), and used to calculate the Kimura two-parameter distance (K) against the donor sequence, *i.e.* the fragment from *gra-MIR482/2118d/e_Chr09*, by distmat embedded in EMBOSS v6.6.0 (Rice *et al.*, 2000). The activity (capture time T) of each K was estimated by the formula: $T = K/(2 * r)$, where K and r refer to the Kimura two-parameter distance and a general substitution rate ($r = 1.3 \times 10^{-8}$; Clark *et al.*, 2005), respectively. K was calculated by $K = -\frac{1}{2} \ln [(1 - 2p - q)\sqrt{1 - 2q}]$, where p and q are the proportion of sites that show transitional and transversional differences, respectively.

Measurement of the expression level of miR482 and miR2118

The expression level of individual miR482/2118 in each small RNA dataset was quantified as reads per million (RPM) using the formula: (the number of miR482 or miR2118 reads/the total number of clean small RNA reads) $\times 10^6$. Only the fully matched reads detected at least twice were counted. Data shown in Figure 1b were averages of all small RNA datasets for the given species.

DATA AVAILABILITY STATEMENT

The small RNA datasets used in estimating the expression level of individual miR482/2118 isoforms can be found in NCBI (<https://www.ncbi.nlm.nih.gov/>) under the following accession numbers: SRR1029586 to SRR1029588, SRR959748, and SRR959750 (*G. arboueam*), SRR616255 to SRR616257, SRR959884, and SRR959924 (*G. raimondii*), SRR080703, GSM699076, and GSM699077 (*G. barbadense*), SRR080704, SRR1586238 to SRR1586243, SRR473422 to SRR473425, SRR959501, SRR1586235 to SRR1586237, SRR324838 to SRR324840, SRR959858, GSM699074, and GSM699075 (*G. hirsutum*). The expression level of individual miR482/2118 isoforms in MCU-5 and Sikra 1-4 was estimated using unpublished small RNA datasets that can be accessed upon request by contact the corresponding author.

ACKNOWLEDGEMENTS

We thank Yuman Yuan (CSIRO Agriculture and Food) for his excellent technical support and the two anonymous reviewers for their insightful comments and suggestions. This work was supported by Cotton Breeding Australia, a joint venture between Cotton Seed Distributors Ltd. Australia and CSIRO Agriculture and Food. Enhui Shen's work in Australia was supported by a Scholarship for PhD Student Studying Overseas from Zhejiang University. Tianzi Chen's work in Australia was supported by a Jiangsu Government Scholarship for Overseas Studies.

AUTHOR CONTRIBUTIONS

QHZ conceived the study; ES, TC, XZ, JS, FL and QHZ did experiments and/or analysed the data; QHZ and ES wrote the manuscript; DJL and IW revised the manuscript. All authors read and approved the final version.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Sequence alignments of pre-miR482/2118 with mutations in their mature miRNA or precursor sequences.

Figure S2. Syntenic relationship of individual *MIR482/2118* locus.

Figure S3 Alignment of the 5' and 3' ends of the sequences showing in Figure 3(a).

Figure S4. Phylogenetic tree of *MIR482/2118d* and its derivatives from Gra, Ghr and Gba, and the homologous sequences potentially involved in capture of *MIR482d* by a GosTE.

Figure S5. Comparison of the *MIR482d/e* locus in *Theobroma cacao*, *G. raimondii* and *G. arboreum*.

Figure S6. Alignment of randomly selected members of the subfamilies *GosTE9,10,11,25,34,39,43* that have a typical size of ~617 bp. The 5' and 3' TIRs are highlighted in red.

Figure S7. Alignment of the DDE and MYB domains of GosTEs.

Figure S8. Alignment of Ghr-GosTE40-1 with its orthologous sequences in Gra and Gba.

Figure S9. Phased siRNAs generated from *Gh_A12G0912*.

Table S1. Pre-miR482/2118 in the four cotton species.

Table S2. Syntenic and orthologous relationship of the *MIR482/2118* loci in the four cotton species.

Table S3. List of the members of GosTE1 to GosTE49 in the four cotton species.

Table S4. Number of the fully matches of the TIRs of GosTE found in the genomes of the four cotton species.

Table S5. GosTEs containing the DDE and/or MYB domains and candidate autonomous GosTEs.

Table S6. The expression levels of individual miR482 and miR2118 members in different tissues and pathogen challenges in the four cotton species.

Table S7. Predicted miR482/2118 targets in the four cotton species.

REFERENCES

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*. *Nat. Genet.* **36**, 1282-1290.

Bennetzen, J.L. and Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* **65**, 505-530.

Canto-Pastor, A., Santos, B., Valli, A.A., Summers, W., Schornack, S. and Baulcombe, D.C. (2019) Enhanced resistance to bacterial and oomycete pathogens by short tandem target mimic RNAs in tomato. *Proc. Natl. Acad. Sci. USA*, **116**, 2755-2760.

Cerbin, S. and Jiang, N. (2018) Duplication of host genes by transposable elements. *Curr. Opin. Genet. Dev.* **49**, 63-69.

- Chen, J., Hu, Q., Zhang, Y., Lu, C. and Kuang, H.** (2014) P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.* **42**, D1176-1181.
- Chen, J., Huang, Q., Gao, D. et al.** (2013) Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nat. Commun.* **4**, 1595.
- Clark, R. M., Tavaré, S. and Doebley, J.** (2005) Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Mol. Biol. Evol.* **22**, 2304–2312.
- Cronn, R.C., Small, R.L., Haselkorn, T. and Wendel, J.F.** (2002) Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *Am. J. Bot.* **89**, 707-725.
- Dai, X., Zhuang, Z. and Zhao, P.X.** (2018) psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* **46**, W49-W54.
- D'Ario, M., Griffiths-Jones, S. and Kim, M.** (2017) Small RNAs: big impact on plant development. *Trends Plant Sci.* **22**, 1056-1068.
- de Vries, S., Kloesges, T. and Rose, L.E.** (2015) Evolutionarily dynamic, but robust, targeting of resistance genes by the miR482/2118 gene family in the *Solanaceae*. *Genome Biol. Evol.* **7**, 3307-3321.
- Du, X., Huang, G., He, S. et al.** (2018) Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat. Genet.* **50**, 796-802.
- Edgar, R. C.** (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.
- Fahlgren, N., Howell, M.D., Kasschau, K.D. et al.** (2007) High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of *MIRNA* genes. *PLoS One*, **2**, e219.
- Fattash, I., Rooke, R., Wong, A., Hui, C., Luu, T., Bhardwaj, P. and Yang, G.** (2013) Miniature inverted-repeat transposable elements: discovery, distribution, and activity. *Genome*, **56**, 475-486.
- Felippes, F.F., Schneeberger, K., Dezulian, T., Huson, D.H. and Weigel, D.** (2008) Evolution of *Arabidopsis thaliana* microRNAs from random sequences. *RNA*, **14**, 2455-2459.
- Ferguson, A.A., Zhao, D. and Jiang, N.** (2013) Selective acquisition and retention of genomic sequences by Pack-Mutator-like elements based on guanine-cytosine content and the breadth of expression. *Plant Physiol.* **163**, 1419-1432.
- Feschotte, C. and Pritham, E.J.** (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* **41**, 331-368.

- Freeling, M.** (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60**, 433-453.
- Gonzalez, V.M., Muller, S., Baulcombe, D. and Puigdomenech, P.** (2015) Evolution of *NBS-LRR* gene copies among dicot plants and its regulation by members of the miR482/2118 superfamily of miRNAs. *Mol. Plant*, **8**, 329-331.
- Grzebelus, D., Lasota, S., Gambin, T., Kuchеров, G. and Gambin, A.** (2007) Diversity and structure of *PIF/Harbinger*-like elements in the genome of *Medicago truncatula*. *BMC Genomics* **8**:409.
- Guddeti, S., Zhang, D.C., Li, A.L., Leseberg, C.H., Kang, H., Li, X.G., Zhai, W.X., Johns, M.A. and Mao, L.** (2005) Molecular evolution of the rice miR395 gene family. *Cell Res.* **15**, 631-638.
- Hsu, C.C., Lai, P.H. and Chen, T.C. et al.** (2019) *PePIF1*, a *P*-lineage of *PIF-like* transposable element identified in protocorm-like bodies of *Phalaenopsis orchids*. *BMC Genomics* **20**:25.
- Hu, Y., Chen, J., Fang, L., et al.** (2019) *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat Genet.* **51**, 739-748.
- Jiang, N., Bao, Z., Zhang, X., et al.** (2003) An active DNA transposon family in rice. *Nature*, **421**:163–167.
- Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R.** (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569-573.
- Kumar, S., Stecher, G. and Tamura, K.** (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870-1874.
- Li, F., Pignatta, D., Bendix, C., Brunkard, J.O., Cohn, M.M., Tung, J., Sun, H., Kumar, P. and Baker, B.** (2012) MicroRNA regulation of plant innate immune receptors. *Proc. Natl. Acad. Sci. USA*, **109**, 1790-1795.
- Li, S., Castillo-Gonzalez, C., Yu, B. and Zhang, X.** (2017) The functions of plant small RNAs in development and in stress responses. *Plant J.* **90**, 654-670.
- Liang, ZK. and Schnable, J.C.** (2018) Functional divergence between subgenomes and gene pairs after whole genome duplications. *Mol. Plant*, **11**, 388-397.
- Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W. and Kuang, H.** (2012) Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol. Biol. Evol.* **29**, 1005-1017.
- Maher, C., Stein, L. and Ware, D.** (2006) Evolution of *Arabidopsis* microRNA families through duplication events. *Genome Res.* **16**, 510-519.

- Markova, D.N. and Mason-Gamer, R.J.** (2015) The role of vertical and horizontal transfer in the evolutionary dynamics of *PIF-like* transposable elements in *Triticeae*. *PLoS One* **10**: e0137648.
- Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. and Punta, M.** (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121.
- Morata, J., Marin, F., Payet, J. and Casacuberta, J.M.** (2018) Plant lineage-specific amplification of transcription factor binding motifs by miniature inverted-repeat transposable elements (MITEs). *Genome Biol. Evol.* **10**, 1210-1220.
- Panchy, N., Lehti-Shiu, M. and Shiu, S.H.** (2016) Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294-2316.
- Paterson, A.H., Wendel, J.F., Gundlach, H. 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.
- Prüfer, K., Stenzel, U., Dannemann, M., Green, R.E., Lachmann, M. and Kelso, J.** (2008) PatMan: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530-1531.
- Rice, P., Longden, L. and Bleasby, A.** (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277.
- Shen, E., Zou, J., Hubertus Behrens, F. et al.** (2015) Identification, evolution, and expression partitioning of miRNAs in allopolyploid *Brassica napus*. *J. Exp. Bot.* **66**, 7241-7253.
- Shivaprasad, P.V., Chen, H.M., Patel, K., Bond, D.M., Santos, B.A. and Baulcombe, D.C.** (2012) A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell*, **24**, 859-874.
- Sullivan, M.J., Petty, N.K. and Beatson, S.A.** (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, **27**, 1009–1010.
- Sunkar, R., Li, Y.F. and Jagadeeswaran, G.** (2012) Functions of microRNAs in plant stress responses. *Trends Plant Sci.* **17**, 196-203.
- Voinnet, O.** (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669-687.
- Wang, M., Tu, L., Yuan, D. et al.** (2019) Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat Genet.* **51**, 224-229.

Wang, Q., Liu, N., Yang, X., Tu, L. and Zhang, X. (2016) Small RNA-mediated responses to low- and high-temperature stresses in cotton. *Sci. Rep.* **6**, 35558.

Wang, S., Zhu, Q.H., Guo, X., Gui, Y., Bao, J., Helliwell, C. and Fan, L. (2007) Molecular evolution and selection of a gene encoding two tandem microRNAs in rice. *FEBS Lett.* **581**, 4789-4793.

Wang, Y., Tang, H., Debarry, J.D., et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wendel, J.F. (1989) New World tetraploid cottons contain Old World cytoplasm. *Proc. Natl. Acad. Sci. USA*, **86**, 4132-4136.

Wendel, J.F. and Cronn, R.C. (2003) Polyploidy and the evolutionary history of cotton. *Adv. Agron.* **78**, 139-186.

Wicker, T., Sabot, F., Hua-Van, A. et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973-982.

Wicker, T., Buchmann, J.P. and Keller, B. (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.* **20**, 1229-1237.

Xia, R., Xu, J., Arikiti, S. and Meyers, B.C. (2015) Extensive families of miRNAs and *PHAS* Loci in Norway Spruce demonstrate the origins of complex phasiRNA networks in seed plants. *Mol. Biol. Evol.* **32**, 2905-2918.

Zhai, J., Jeong, D.H., De Paoli, E. et al. (2011) MicroRNAs as master regulators of the plant *NB-LRR* defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev.* **25**, 2540-2553.

Zhang, T., Hu, Y., Jiang, W. et al. (2015) Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**, 531-537.

Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W.B. and Wessler, S.R. (2001) *P instability factor*: an active maize transposon system associated with the amplification of *Tourist-like* MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA*, **98**:12572-7.

Zhang, X., Jiang, N., Feschotte, C. and Wessler, S. R. (2004) *PIF*- and *Pong*-like transposable elements: distribution, evolution and relationship with *Tourist-like* miniature inverted-repeat transposable elements. *Genetics* **166**, 971-986.

Zhang, Y., Xia, R., Kuang, H. and Meyers, B.C. (2016) The diversification of plant *NBS-LRR* defense genes directs the evolution of microRNAs that target them. *Mol. Biol. Evol.* **33**, 2692-2705.

Zhao, M., Meyers, B. C., Cai, C., Xu, W. and Ma, J. (2015) Evolutionary patterns and coevolutionary consequences of *MIRNA* genes and microRNA targets triggered by multiple mechanisms of genomic duplications in soybean. *Plant Cell* **27**, 546-562.

Zhu, Q.H., Fan, L., Liu, Y., Xu, H., Llewellyn, D. and Wilson, I. (2013) miR482 regulation of *NBS-LRR* defense genes during fungal pathogen infection in cotton. *PLoS One*, **8**, e84390.

Zhu, Q.H., Spriggs, A., Matthew, L., Fan, L., Kennedy, G., Gubler, F. and Helliwell, C. (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res.* **18**, 1456-1465.

Figure legends

Figure 1. The mature miR482/2118 sequences and the number of *MIR482/2118* loci in the four cotton species.

^a The first 22-nt of each sequence is miR482 and the miR2118 sequences are underlined; nucleotides that are different from miR482c/2118c are highlighted in red. ^b gar-miR482i/2118i was evolved from gar-miR482b/2118b; ^c both gra-miR482o/miR2118o and gra-miR482p/miR2118p were evolved from gra-miR482d/2118d; ^d ghr-miR482a/2118a was evolved from ghr-miR482c/2118c; ^e both gba-miR482m/2118m and gba-miR482n/2118n were evolved from gba-miR482d/2118d; ^f gba-miR482q/2118q was evolved from gba-miR482e/2118e, existing in Gba accessions 3-79 (Wang *et al.*, 2019) and Pima S-7 but not in Hai7124 (Hu *et al.*, 2019); ^g the number in parentheses indicates the number of *MIR482/2118* loci on unassigned scaffolds; ^h the sequence of *gba-MIR482g/2118g* is found in the Gba genome, but the precursor would not be able to form a stem-loop structure required for generation of gba-miR482g/2118g due to mutations in its 5' arm (Figure S1f); ⁱ gba-miR482l/2118l has 1-bp insert in the position marked by an arrow (Figure S1e).

Figure 2. Syntenic relationship of *MIR482/2118* loci in the four cotton species. The right panel shows genome-wide overview of the syntenic relationship of *MIR482/2118* loci that are connected by lines highlighted in different colours. The multiple genome syntenic relationship was established based on macrosyntenic blocks. Blocks contain at least 10 one-to-one gene pairs are shown. The left panel (with dashed line border) shows the syntenic relationship of an orthologous *MIR482/2118* locus (*MIR482/2118a* and *MIR482/2118c* as an example) in the four cotton species. Orthologous genes are linked by grey lines. For simplicity, *MIR482/2118* was shortened as *MIR482*.

Figure 3. Expansion of the *MIR482/2118* family in *G. raimondii* caused by transposition of a transposable element that captured a *MIR482d/2118d*. For simplicity, *MIR482/2118* was shortened as *MIR482*.

(a) Sequence conservation amongst gra-*MIR482*-containing transposable elements (TEs). Each orange line represents a gra-*MIR482*-containing TE defined by a pair of red arrows representing 5' and 3' terminal inverted-repeats or TIRs (the 3' TIR was not found in *gra-MIR482d*_Chr13) or the donor (*gra-MIR482d/e*_Chr09) sequence, from which all other *gra-MIR482d*, *gra-MIR482o*, *gra-MIR482p* and *gra-MIR482j* were derived. Sequence similarities are color-coded and blank spaces represent deletions. The Easyfig program was used in generating the image (Sulliva *et al.*, 2011).

(b) Comparison of the TIRs and target site duplication (TSD) from the gra-*MIR482/2118*-containing TEs. TIR nucleotides that are different from those of *gra-MIR482d*_Chr08 are highlighted in red. TSDs are in bold type.

Figure 4. Size distribution of the transposable elements of the subfamilies GosTE1 to GosTE46. The subfamilies with 10 or more members are also shown individually. Each TE is indicated by a grey dot and the width of the bulge is proportional to the number of TEs within a specific size class. Different subfamilies are in a different colour.

Figure 5. The expression levels of individual miR482 and miR2118 members.

Data shown are averages of all small RNA datasets from each species. The numbers with a pink box indicate that the expression of the corresponding miR482 or miR2118 should not be detected due to absence of the miRNA (except ghr-miR482q/2118q) in the reference genomes. Detection of these miRNAs could be caused by sequencing errors. ^a ghr-miR482q/2118q was absent in TM-1, but present in the cultivar Siokra 1-4.

Figure 6. Changed targeting capacity caused by mutations in miR482/2118 for each of the four cotton species. Numerator and denominator represent the total number of targets and number of targets related to disease resistance, respectively. Targets on the left are lost when the miRNA sequence mutates while those on the right are predicted as newly evolved targets. Unchanged targets are shown in the intersection.

Figure 7. A working model for the capture, proliferation and spread of the *MIR482/2118d* members in cotton. A double-strand break (DSB) occurred in the TE located on ChrB. One 3' end of the DSB invaded downstream of *MIR482/2118d* on ChrA due to the presence of a short homologous sequence between ChrA and ChrB. The invaded 3' end extended based on sequence of ChrA and switched back to ChrB after copying the *MIR482/2118d* locus. Consequently, the ChrA sequence containing the *MIR482/2118d* locus was captured by the TE of ChrB as a result of DSB repair. The TE containing the *MIR482/2118d* locus could transpose to other genomic locations on the same chromosome (ChrB) or a different chromosome (e.g. ChrC). The new *MIR482/2118d*-containing TEs and the originally TE-with-captured-*MIR482/2118d* locus are all capable of further transposition to spread *MIR482/2118d* across the whole genome. The sequences of both TE and *MIR482/2118d* are able to diverge due to mutations and/or homologous recombination to give rise to variants of *MIR482/2118d* and/or different length of TEs. The *MIR482/2118* locus is indicated by a red rectangle. Each TE is defined by a pair of curly braces with the 5' and 3' TIRs shown as red triangles.

Table 1. Number of transposable element members in each GosTE subfamily

Table 1. Number of transposable element members in each GosTE subfamily

TE Sub-family [†]	Gar	Gra	Ghr		Gba		Total	TE Sub-family [‡]	Gar	Gra	Ghr		Gba		Total
			At	Dt	At	Dt					At	Dt	At	Dt	
GosTE1		1		1		1	3	GosTE24	2						2
GosTE2		1					1	GosTE25	2	1	1(1)		1	1	6(1)
GosTE3		1					1	GosTE26	1					1	2
GosTE4		1					1	GosTE27		1	1			1	3
GosTE5		1					1	GosTE28		1					1
GosTE6 [§]		2			1	1(1) [¶]	4(1)	GosTE29		2	1	2	1	1(1)	6(1)
GosTE7 [§]		16	1	15(1)	1	16	49(1)	GosTE30		3		3		2	8
GosTE8 [§]		7	4	11(1)	9	11	42(1)	GosTE31		1					1
GosTE9 [§]		3	1	4	2	3	13	GosTE32		2		2(1)		4	8(1)
GosTE10 [§]		2	1	5	1	5	14	GosTE33		5		1(1)			6
GosTE11 [§]		237	29	156(12)	53	224(3)	699(13)	GosTE34		1					1
GosTE12				1		1	2	GosTE35		3		2(1)		3	8(1)
GosTE13				1		1	2	GosTE36		3		1		2	6
GosTE14				1			1	GosTE37		1		1		1	3
GosTE15 [§]				5	1	6	12	GosTE38		1				1	2
GosTE16				1		1	2	GosTE39				1			1
GosTE17				(1)			(1)	GosTE40				1			1
GosTE18 [§]		1		(1)		1	2(1)	GosTE41				1		1	2
GosTE19						1	1	GosTE42				1			1
GosTE20						1	1	GosTE43				1		1	2
GosTE21					1		1	GosTE44						2	2
GosTE22						1	1	GosTE45				1			1
GosTE23 [§]		4		9	4	10	27	GosTE46				1		1	2
Total	0	277	36	210(16)	73	284(4)	880(20)		5	25	1	21(4)	2	22(1)	76(5)

[†] At least one member of each subfamily contains *MIR482/2118*.

[‡] All members of these subfamilies do not contain *MIR482/2118*.

[§] Subfamilies with TE-without-miR482/2118 also identified.

[¶] Numbers in parentheses indicate the number of TE on unassigned Scaffolds.

miR482/2118 member	Sequence ^a	Gar	Gra	Ghr-At	Ghr-Dt	Gba-At	Gba-Dt
miR482a/2118a	TCTTTCCTACTCCTCCCATACCAC			1 ^d			
miR482b/2118b	TCTTGCCTACTCCACCCATGCCAC		1	(1) ^g	1	1	1
miR482c/2118c	TCTTTCCTACTCCTCCCATTCCAC	1	1		1	1	1
miR482d/2118d	TCTTTCCAAATTCCTCCCATTCCAC	1	5	7	11(1) ^g	6	15
miR482e/2118e	TCTTGCCGATTCACCCATGCCTA	1	1	1	1	1	1
miR482f/2118f	TCTTTCCTATGCCCCCATTCCAC	1	1	1	1	1	1
miR482g/2118g	TCTTCCCAACACCTCCCATACCCT	1	1	1	1	1 ^h	1
miR482h/2118h	TCTTACCAACCCTCCCATACCCT	1	1	1	1	1	1
miR482i/2118i	TCTTGCCTACTCCGCCCATGCCAC	1 ^b					
miR482j/2118j	TCTTTCATAATTCCTCCCATTCCAC		4	(1) ^g		1	
miR482k/2118k	TCTTCCCAAACCTCCAATTCCAA	1	1	1	1	1	1
miR482l/2118l	TCTTCCCGAGGCCACCCATTCCAG	1	1	1	1	1	1 ⁱ
miR482m/2118m	TCTTTCCAAATTCCTCTCATTCAC					1 ^e	
miR482n/2118n	TCTTTTCAATTCCTCCCATTCCAC						1 ^e
miR482o/2118o	TCTTACCAATTCCTGCCATTCCAC		1 ^c				
miR482p/2118p	TCTTTCCAAATTCCTCCTATTCCAC		1 ^c				
miR482q/2118q	TCTTGCCGGTTCACCCATGCCTA						1 ^f

Figure 1

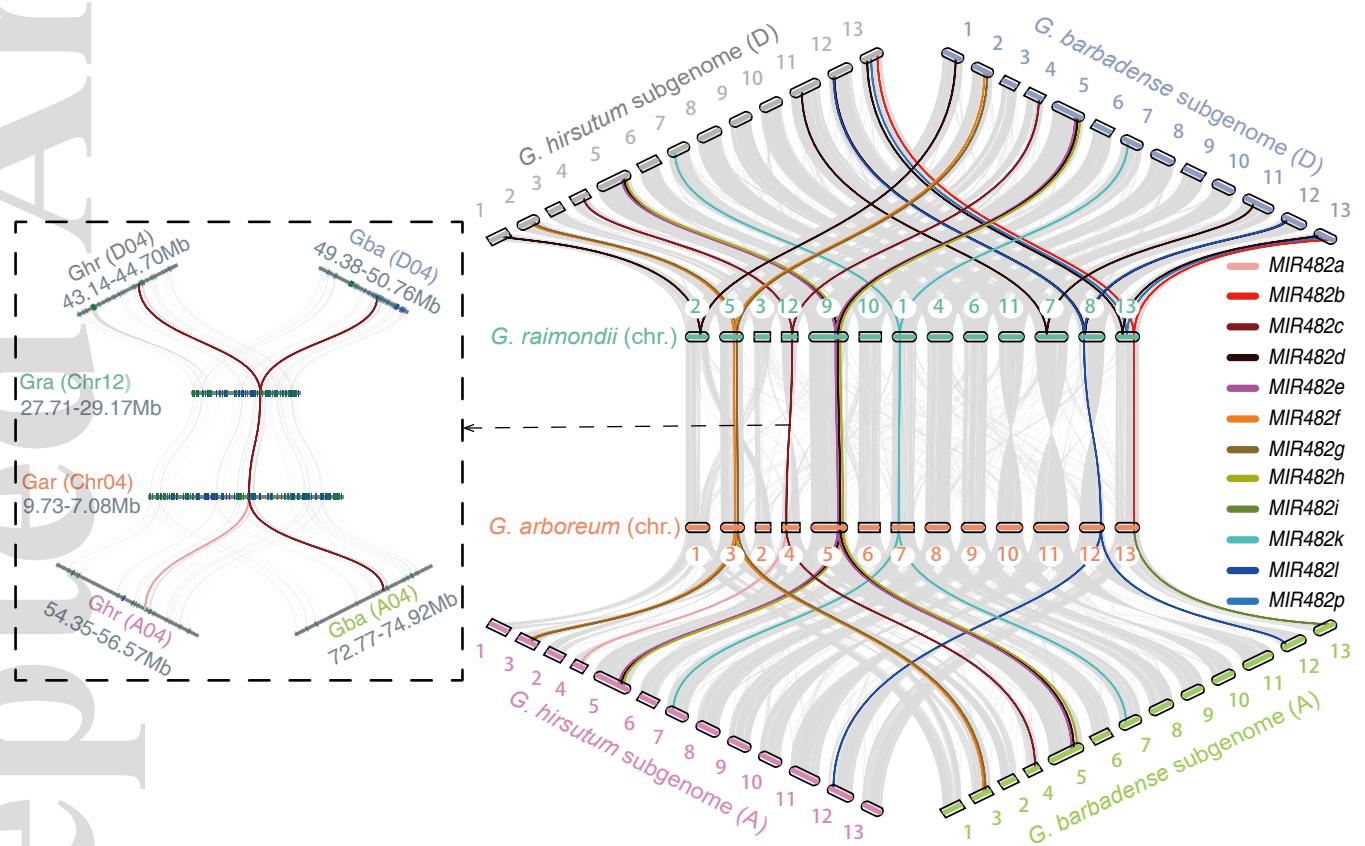


Figure 2

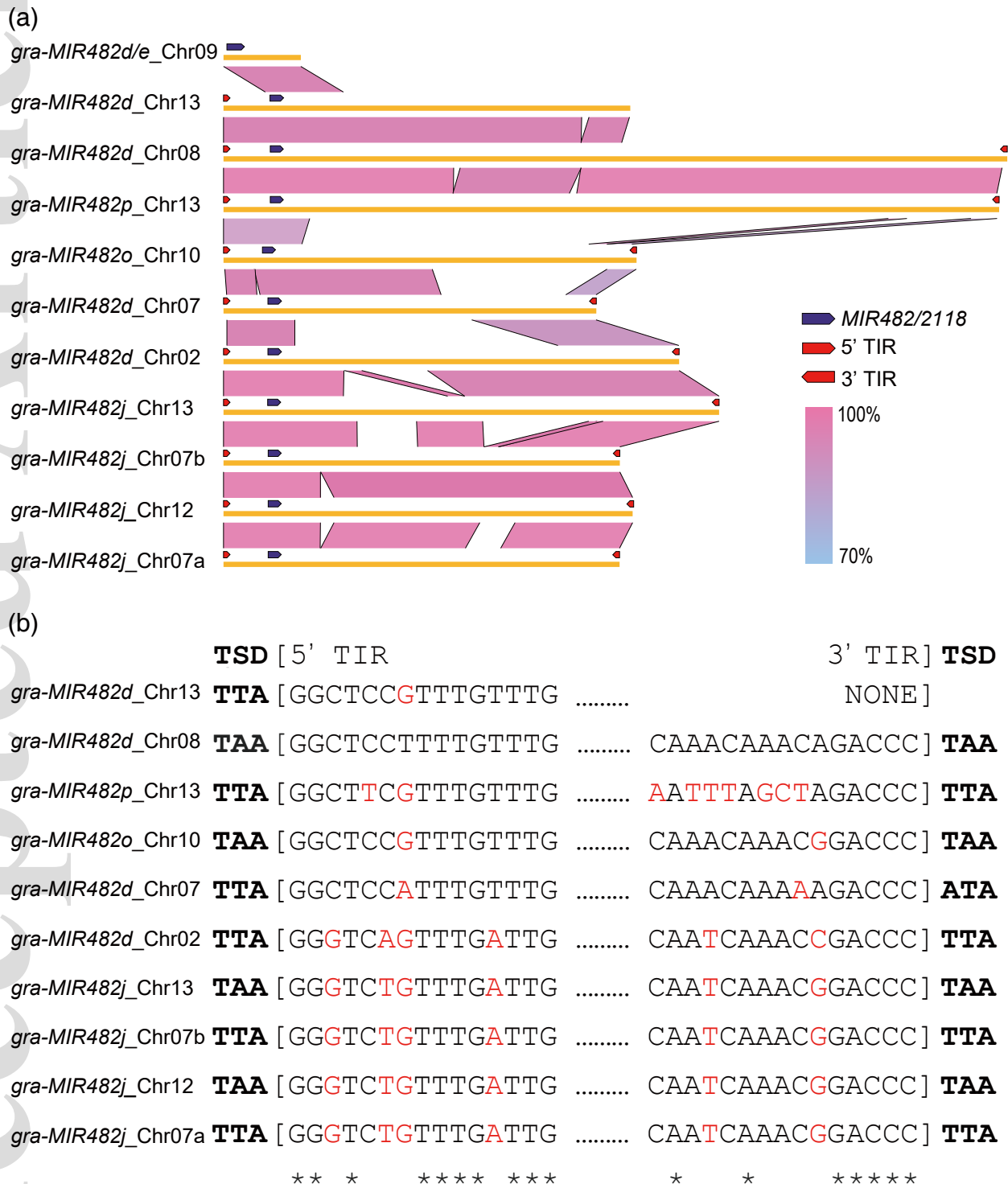


Figure 3

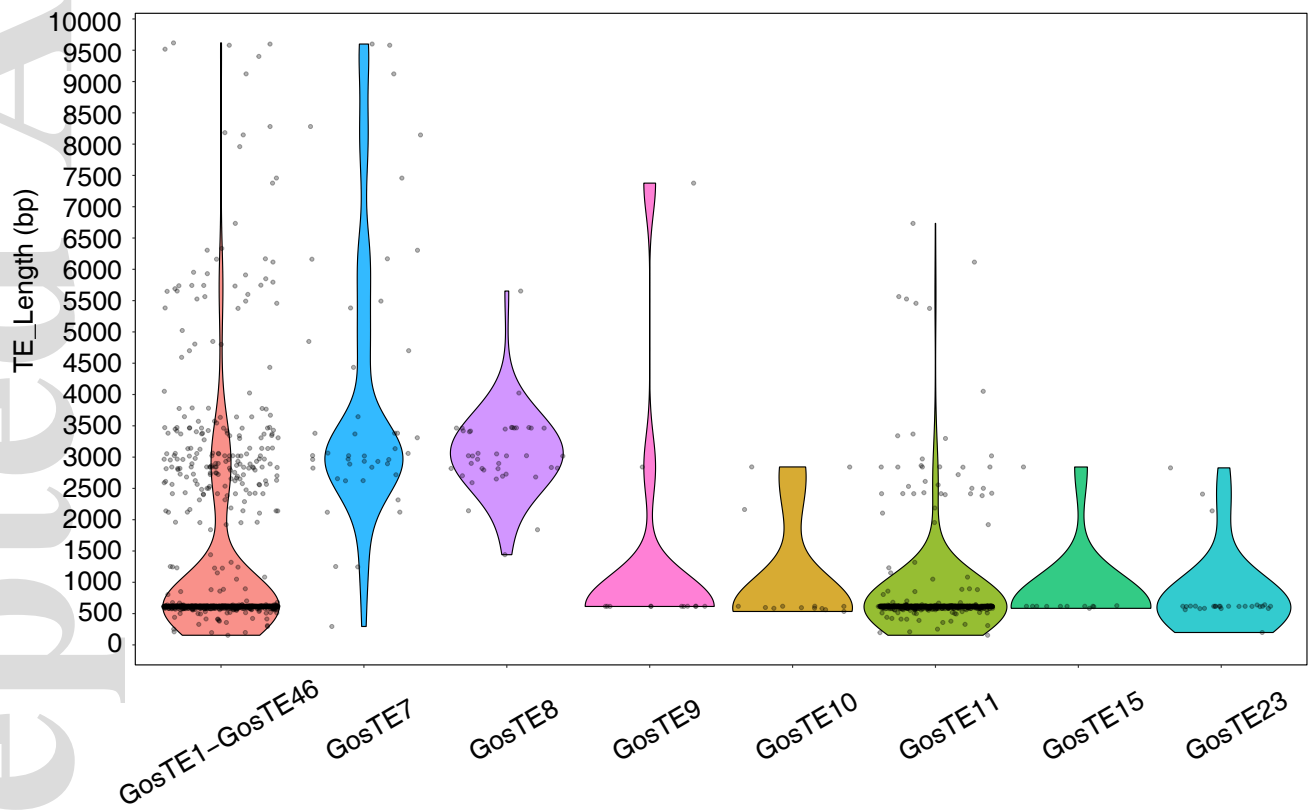


Figure 4

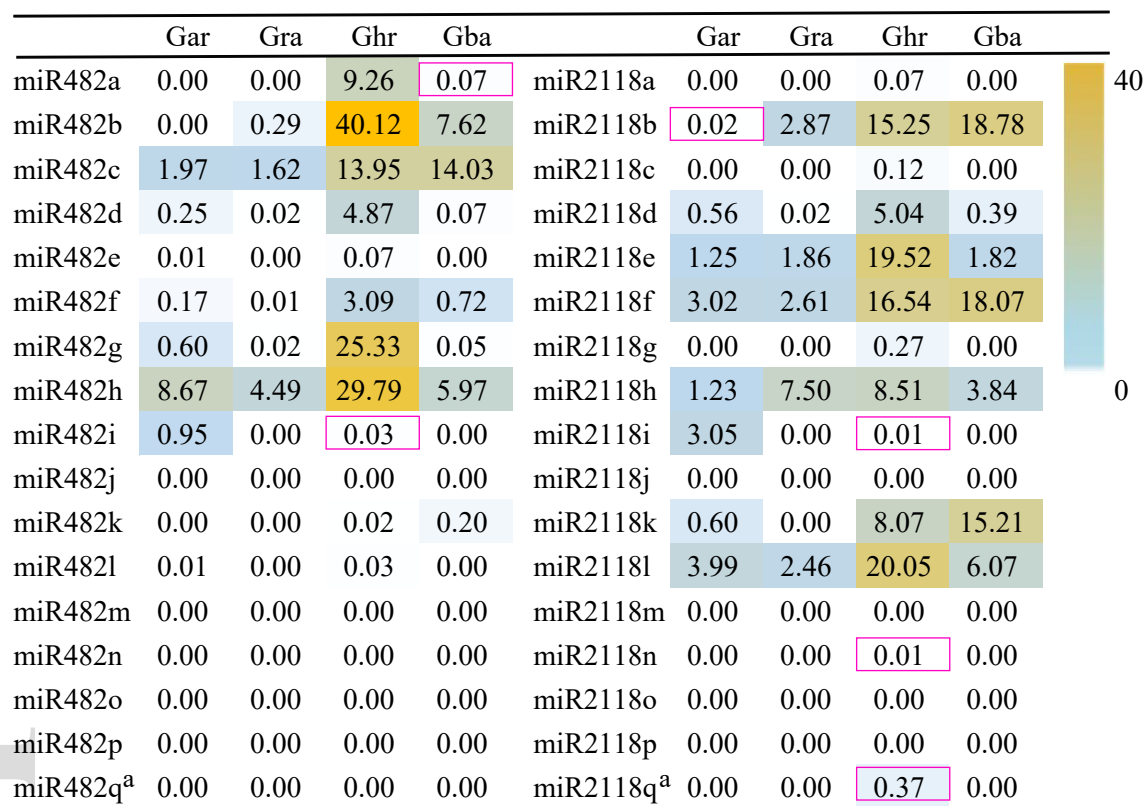


Figure 5

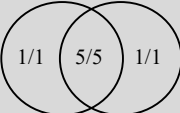
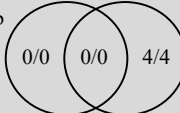
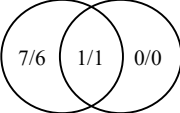
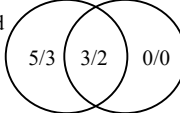
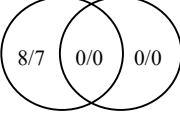
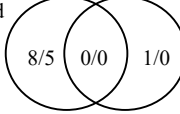
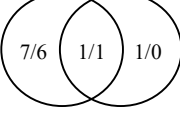
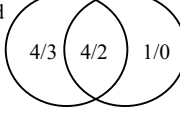


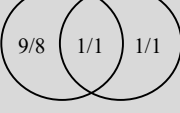
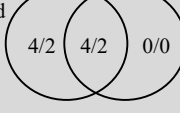
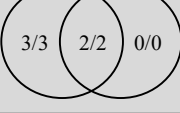
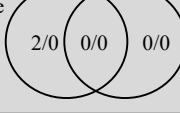
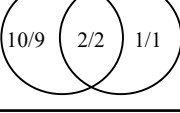
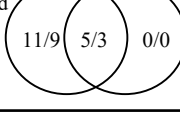
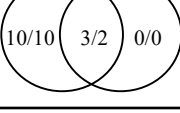
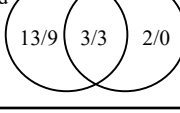
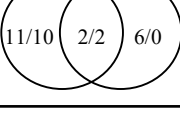
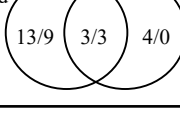
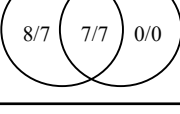
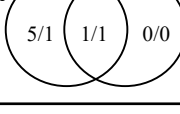
Number of targets of the original (left) and the newly evolved (right) miRNA	Loss or gain of targets due to the mutation	Number of targets of the original (left) and the newly evolved (right) miRNA	Loss or gain of targets due to the mutation
gar-miR482b  gar-miR482i	Loss: 1/1 Gain: 1/1	gar-miR2118b  gar-miR2118i	Loss: 0/0 Gain: 4/4
gra-miR482d  gra-miR482j	Loss: 7/6 Gain: 0/0	gra-miR2118d  gra-miR2118j	Loss: 5/3 Gain: 0/0
gra-miR482d  gra-miR482o	Loss: 8/7 Gain: 0/0	gra-miR2118d  gra-miR2118o	Loss: 8/5 Gain: 1/0
gra-miR482d  gra-miR482p	Loss: 7/6 Gain: 1/0	gra-miR2118d  gra-miR2118p	Loss: 4/3 Gain: 1/0
ghr-miR482c  ghr-miR482a	Loss: 14/6 Gain: 6/3	ghr-miR2118c  ghr-miR2118a	Loss: 4/2 Gain: 0/0
ghr-miR482d  ghr-miR482j	Loss: 9/8 Gain: 1/1	ghr-miR2118d  ghr-miR2118j	Loss: 4/2 Gain: 0/0
ghr-miR482e  ghr-miR482q	Loss: 3/3 Gain: 0/0	ghr-miR2118e  ghr-miR2118q	Loss: 2/0 Gain: 0/0
gba-miR482d  gba-miR482j	Loss: 10/9 Gain: 1/1	gba-miR2118d  gba-miR2118j	Loss: 11/9 Gain: 0/0
gba-miR482d  gba-miR482m	Loss: 10/10 Gain: 0/0	gba-miR2118d  gba-miR2118m	Loss: 13/9 Gain: 2/0
gba-miR482d  gba-miR482n	Loss: 11/10 Gain: 6/0	gba-miR2118d  gba-miR2118n	Loss: 13/9 Gain: 4/0
gba-miR482e  gba-miR482q	Loss: 8/7 Gain: 0/0	gba-miR2118e  gba-miR2118q	Loss: 5/1 Gain: 0/0

Figure 6

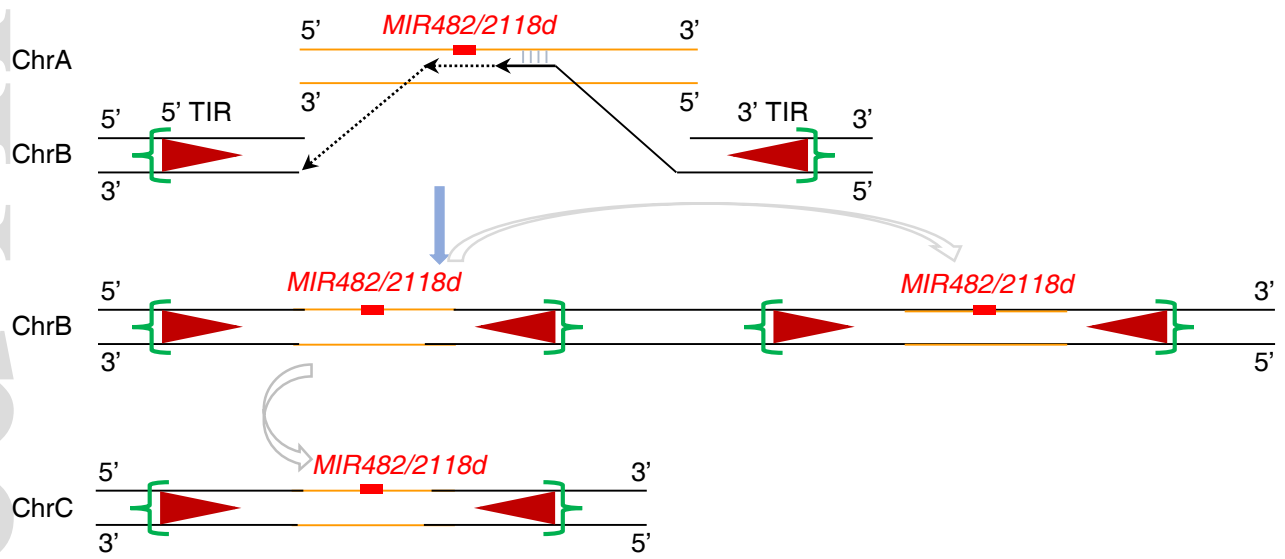


Figure 7