

DNA methylation and gene expression dynamics during cotton ovule and fiber development

Qingxin Song¹, Xueying Guan^{1,2}, Z. Jeffrey Chen^{1§}

5 ¹ Department of Molecular Biosciences and Institute for Cellular and Molecular Biology and
Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin,
Texas 78712, USA

² State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural
10 University, Nanjing 210095, China

[§] Corresponding author

Email addresses: zichen@austin.utexas.edu

15 Running title: DNA methylation dynamics during fiber development

Kew Words: DNA methylation, cotton, fiber, polyploidy, gene expression, small RNAs,
epigenetics, transposons

Abstract

Cotton is the largest source of renewable textile fiber and a successful model of transgenic applications in crop production. However, improving cotton production using fiber-related transgenes is somewhat difficult. This is probably related to unique epigenetic and gene expression changes during fiber development. Here we show that inhibiting DNA methylation impairs fiber development. Genome-wide methylcytosine-, mRNA-, and small RNA-sequencing analyses reveal minor changes in CG and CHG methylation and distinct changes in CHH methylation among different tissues. In ovules CHH hypermethylation is associated with small RNA-directed DNA methylation (RdDM) and expression changes of nearby genes in euchromatin. Remarkably, ovule-derived fiber cells not only maintain euchromatic CHH methylation, but also generate additional heterochromatic CHH hypermethylation independent of RdDM, which represses transposable elements (TEs) and nearby genes including fiber-related genes. Furthermore, DNA methylation contributes to the expression bias of homoeologous genes in ovules and fibers. This spatiotemporal DNA methylation in promoters could act as a double-lock feedback mechanism to regulate TE and gene expression, which could be translated into genomic and biotechnological improvement of agronomic traits.

The most widely-cultivated cotton (*Gossypium hirsutum* L., AADD) is an allotetraploid species, which originated 1-2 million years ago from interspecific hybridization between A-genome species, resembling *Gossypium herbaceum* or *Gossypium arboreum*, and D-genome species, resembling *Gossypium raimondii* (Wendel 1989). The intergenomic interaction in
5 allotetraploid cottons induces longer fiber and higher yield, coincident with expression bias of fiber-related homoeologous genes (Flagel et al. 2008; Pang et al. 2009), which provides the basis of selection and domestication for agronomic traits in cotton and many other polyploid crops (Wendel 1989; Chen et al. 2007). Each cotton fiber is a singular cell derived from an epidermal cell of the ovule, undergoing rapid cell elongation and cellulose biosynthesis, and
10 ~100,000 fiber cells develop semi-synchronously in each ovule (seed) and can reach six centimeters in length (Kim and Triplett 2001; Guan et al. 2014). In early stages of fiber development, rapid cell growth is associated with a dramatic increase of DNA content by endoreduplication (Van't Hof 1999; Wu et al. 2006) and dynamic changes in gene expression and small RNAs (Pang et al. 2009; Guan et al. 2014; Yoo and Wendel 2014). Interestingly, DNA
15 methylation changes are related to seasonal variation of fiber development in cotton (Jin et al. 2013) and is also shown to change among different tissues including fibers based on the methylation-sensitive high-performance liquid chromatography (HPLC) assay (Osabe et al. 2014). Moreover, over-expressing fiber-related transgenes often leads to the unexpected outcome of fiber phenotypes (Walford et al. 2011). These data indicate a potential role for DNA
20 methylation in gene expression and phenotypic traits such as cotton fiber, which could be selected by breeding.

DNA methylation is heritable epigenetic information in most eukaryotes and mediates many epigenetic phenomena, including imprinting and transposon silencing (Bird 1992; Richards 1997; Law and Jacobsen 2010; Haag and Pikaard 2011; Schultz et al. 2012). In plants, DNA is methylated in CG, CHG and CHH (H=A, T, or C) sites through distinct pathways. In *Arabidopsis*, CG methylation is maintained by METHYLTRANSFERASE1 (MET1) (Kankel et al. 2003), a homolog of mammalian DNMT1 (Li et al. 1992). Plant-specific CHROMOMETHYLASE3 (CMT3) is primarily responsible for CHG methylation, which is coupled with H3K9 dimethylation (Lindroth et al. 2001; Cao and Jacobsen 2002). CHH methylation is established *de novo* by DOMAINS REARRANGED METHYLTRANSFERASE1 and 2 (DRM1 and DRM2) (Cao et al. 2003) through the RNA-directed DNA methylation (RdDM) pathway (Wassenegger et al. 1994), involving 24-nt small interfering RNAs (siRNAs) (Law and Jacobsen 2010; Haag and Pikaard 2011). Recent studies found that independent of RdDM, CHH methylation could also be established by CMT2 (Zemach et al. 2013; Stroud et al. 2014), through histone H1 and DECREASE-IN-DNA-METHYLATION1 (DDM1) activities (Jeddeloh et al. 1999). The methylome data indicate that CMT2 and RdDM pathways preferentially function in heterochromatic and euchromatic regions, respectively (Zemach et al. 2013; Stroud et al. 2014).

Although genome-wide DNA methylation has been examined in *Arabidopsis* (Zhang et al. 2006), soybean (Song et al. 2013), maize (Gent et al. 2013), and other plants and animals (Zemach et al. 2010), the biological roles of RdDM and CMT2-dependent methylation pathways remain elusive. Moreover, DNA methylation patterns and changes during cotton fiber development are unknown. In this study, we tested the effects of DNA methylation on cotton ovule and fiber development. Using methylcytosine-sequencing (MethylC-seq), RNA-seq, and

small RNA-seq analyses, we examined CG, CHG, and CHH methylation patterns in fibers, ovules, and leaves and analyzed differentially methylated regions (DMRs) between the ovule and leaf (OL) and between the fiber and ovule (FO). The methylation patterns in the gene body and 5' and 3' flanking sequences were comparatively analyzed with TE densities and expression levels of nearby genes as well as small RNA loci. The results provide unique roles of CG, CHG, and CHH methylation in cotton ovule and fiber development and biased expression of homoeologous genes.

RESULTS

10 DNA methylation dynamics in cotton leaves, ovules, and fibers

DNA methylation affects growth and development in plants and animals (Bird 1992; Richards 1997). Methylation levels change among different cotton tissues (Osabe et al. 2014). To investigate the biological role of DNA methylation in fiber development, we harvested ovules at -3- 0 days post anthesis (DPA) from the allotetraploid cotton *Gossypium hirsutum* L. acc. TM-1 and cultured them *in vitro* with or without treatment of the DNA methyltransferase inhibitor, 5-aza-deoxycytidine (5-aza-dC) (Haaf et al. 1993). Fiber length and ovule size were suppressed in the initiation and elongation stages by 5-aza-dC treatment (Fig. 1A-C), providing pharmacological evidence for the requirement of DNA methylation during fiber and ovule development.

20 This result led to the investigation of genome-wide DNA methylation changes during ovule and fiber development. We performed whole-genome bisulfite sequencing in leaves, ovules at 0 DPA, and fibers at 14 DPA with two biological replicates and ovules at 14 DPA with

one biological replicate. The genomes of *G. arboreum* (Li et al. 2014) and *G. raimondii* (Paterson et al. 2012) were modified using single nucleotide polymorphism (SNP) generating from genome sequencing reads of *G. hirsutum*, and then used as references for mapping sequence reads and data analysis. The bisulfite-conversion rates were over 99% (Supplemental Table S1). Approximately 80% of cytosines were covered by at least one uniquely mapped read, and the mean coverage of cytosines was 7.5-fold or higher for all tissues. Methylation levels between biological replicates were highly correlated (Pearson $r > 0.8$), indicating reproducibility of the data (Fig. 1D; Supplemental Fig. S1A). Overall, CG and CHG methylation levels were similar among fibers, ovules and leaves but were lower in the D-subgenome than in the A-subgenome for all tissues (Fig. 1D). However, CHH methylation levels were much higher in fibers (~14%) than in ovules (~8.1%) and leaves (~7.6%) and were slightly higher in the D-subgenome than in the A-genome for all tissues (Fig. 1D). This high methylation level in fibers was not related to the developmental stage because the methylation level in ovules was slightly lower at 14 DPA than at 0 DPA (Supplemental Fig. S1B). To study functions of small RNAs in DNA methylation changes in cotton tissues, we generated small RNA-seq data using leaves, ovules at 0 DPA and fibers at 14 DPA. The high CHH methylation level in fibers correlated with the small RNA-seq data, in which 24-nt siRNAs were higher in fibers and ovules than in leaves (Fig. 1E) and consistent with the published small RNA-seq data (Pang et al. 2009).

On each chromosome, DNA methylation was more abundant in transposable element (TE)-rich than gene-rich regions in all tissues (Fig. 1F; Supplemental Fig. S1C). However, 24-nt siRNAs, which induce CHH methylation through RdDM (Haag and Pikaard 2011; Law et al. 2013), were highly enriched in gene-rich regions. These 24-nt siRNAs were derived from TEs but not

from coding sequences in gene-rich regions (Supplemental Fig. S1D). Because each fiber cell is expanded from an epidermal cell of the ovule, further comparison was made between the ovule and leaf (OL) and/or between the fiber and ovule (FO), which represented the transitional stages between ovule and fiber development. OL CHH hypermethylation correlated with gene-rich regions, showing the same trend as the siRNAs (Fig. 1F). On the contrary, FO CHH hypermethylation was enriched in TE- and repeat-rich regions, characteristic of heterochromatin. The data indicate that OL CHH hypermethylation preferentially occurs in euchromatic regions, whereas FO CHH hypermethylation predominates in heterochromatic regions (Fig. 1F; Supplemental Fig. S1E).

While CG and CHG methylation levels in genes and TEs were similar among different tissues (Supplemental Fig. S2), CHH methylation differences in genes were primarily distributed in 5' upstream and 3' downstream sequences, especially between the ovule and leaf (Fig. 1G). Among all TEs, mean CHH methylation levels were higher in the fiber but indistinguishable between the ovule and leaf (Fig. 1H). The higher CHH methylation levels in fibers and ovules were correlated with the closer distances of TEs to the gene (Fig. 1I). When TEs were more than 4-6-kb away from the gene, CHH methylation levels became similar between the ovule and leaf. Consistently, the CHH methylation levels mirrored the siRNA distribution patterns, which decreased as they were further away from the gene, indicating the role of the RdDM pathway in CHH methylation of TEs near the gene (Fig. 1J).

20

Differentially methylated regions

As CG and CHG methylation levels were similar among different tissues examined, further analysis was focused on CHH methylation changes among these tissues. We predict that differentially methylated regions (DMRs) between tissues could play a role in biological function. To test this, we identified 28,108 CHH-hypermethylated DMRs between the ovule and leaf (OL CHH-hyper DMRs) and 94,449 CHH-hypermethylated DMRs between the fiber and ovule (FO CHH-hyper DMRs) (Supplemental Table S2). A subset of these DMRs was validated by bisulfite-sequencing individually cloned genomic fragments (Supplemental Fig. S3), confirming the results of these DMRs from genome-wide analysis. Compared to genome-wide distributions, both OL and FO CHH-hyper DMRs were under-represented in the gene-body (especially exon) (Fig. 2A). However, OL CHH-hyper DMRs predominated in intergenic regions (73%) relative to the genome-wide average (30%), whereas FO CHH-hyper DMRs were enriched in TEs (69% vs. 60%). Among TEs, OL CHH-hyper DMRs were more abundant in TEs <1-kb (68%) than the genome-wide average (30%), while FO CHH-hyper DMR distribution was similar to the genome-wide distribution (Fig. 2B). Methylation levels of most OL CHH-hyper DMRs were higher in fibers than in ovules, indicating that OL CHH-hyper DMRs continued to be hypermethylated in elongating fibers after they elongated from ovule epidermal cells (Fig. 2C). Interestingly, among the OL CHH-hyper DMRs, siRNA expression levels were significantly higher in ovules than in leaves ($P < 1e^{-100}$, Wilcoxon rank-sum test), and this difference was not obvious between the fiber and ovule in FO CHH-hyper DMRs (Fig. 2D). The data suggest that siRNAs induce CHH hypermethylation in the ovule, whereas the CHH methylation increase in the fiber was independent of the RdDM pathway. In addition to CHH methylation, CG and CHG methylation

levels also increased in both FO CHH-hyper DMRs and OL CHH-hyper DMRs, and CHG methylation increased more than CG methylation (Fig. 2E).

Effects of OL CHH-hyper DMR on gene expression

5 To test the role of methylation changes in gene expression, we compared RNA-seq data with DMRs between the ovule and leaf and/or between the fiber and ovule. Down-regulated genes in the ovule relative to the leaf were significantly enriched in the genes overlapping with OL CHH-hyper DMRs in the gene body (Fig. 3A; Supplemental Table S3), which is consistent with the notion that CHH methylation in the gene body correlates negatively with gene expression
10 (Schmitz et al. 2013). However, up-regulated genes in ovules were significantly enriched in the genes overlapped with OL CHH-hyper DMRs in the flanking sequences (1 kb), especially upstream sequences (Fig. 3A; Supplemental Table S3), indicating a positive correlation of CHH methylation in flanking sequence with gene expression. For the genes that were preferentially
15 expressed in the ovule (ovule-preferred genes), the CHH methylation level in the flanking sequences was higher in the ovule than in the leaf, whereas CHH methylation in flanking sequences of the leaf-preferred genes was similar between the ovule and leaf (Fig. 3B). To further test the relationship between CHH methylation and gene expression, we divided genes into quartiles based on their expression levels in each tissue (from low to high) and compared them with DNA methylation differences. CHH and CHG methylation in the gene body were
20 negatively associated with gene expression, and moderately transcribed genes showed the highest CG methylation level in the gene body (Fig. 3C; Supplemental Fig. S4), which are consistent with previous findings (Zilberman et al. 2007; Schmitz et al. 2013). Moreover, CHH

methylation in the 5' upstream coincided with siRNA abundance, which was positively associated with gene expression (Fig. 3C,D). The siRNA levels in the promoter regions of ovule-preferred genes were significantly higher in ovules than in leaves ($P < 1e^{-4}$, Wilcoxon rank-sum test) (Fig. 3E). These results indicate a unique role for the RdDM pathway in OL CHH

5 hypermethylation and up-regulation of genes in ovules. This positive correlation between CHH methylation and gene expression was unexpected but also reported in other plants including maize (Gent et al. 2013) and soybean (Song et al. 2013).

The higher methylation levels of most OL CHH-hyper DMRs in fibers than in ovules may suggest a role for increased methylation in fibers in gene expression. To test this, we extracted

10 18,107 OL CHH-hyper DMRs that showed higher methylation levels (cut-off value >0.05) in fibers than in ovules to examine effects of these DMRs on gene expression in fibers. In contrast to the up-regulation of genes in ovules, down-regulated genes in fibers relative to ovules were significantly enriched in the genes overlapping with these OL CHH-hyper DMRs in the upstream or downstream sequences, suggesting that further increasing methylation levels of some OL

15 CHH-hyper DMRs in fibers represses nearby genes (Fig. 4A). For example, expression of *GhMYB25L*, a key factor for fiber development (Walford et al. 2011), was induced at the fiber initiation stage (0 DPA) but repressed in the elongation stage (Fig. 4B). Methylation levels in the flanking sequences of *GhMyb25L* increased in ovules at 0 DPA and continued to increase in fibers at 14 DPA. These data indicate that CHH hypermethylation in fibers regulates *GhMYB25L*

20 expression during fiber development (Fig. 4B).

The functions of FO CHH hypermethylation

To our surprise, FO CHH-hyper DMRs were not significantly correlated with expression changes in fiber (Fig. 4C; Supplemental Table S4). This was partly because FO CHH-hyper DMRs tended to be excluded from the gene body (Fig. 2A) or from flanking sequences (1-kb) of the gene (Fig. 4D). Instead, FO CHH-hyper DMRs were enriched in TEs (Fig. 2A), and FO CHH hypermethylation was positively correlated with the TE density (Supplemental Fig. S1E). We predict that OL CHH hypermethylation mediates gene expression, while FO CHH hypermethylation represses TEs in heterochromatin. Indeed, RNA-seq data showed that more TEs were expressed in fibers than in ovules or leaves (Fig. 4E). These data indicate that CHH hypermethylation in fibers serve as a mechanism for repressing potential TE activation resulting from rapid fiber cell growth. As CHH methylation in heterochromatin is catalyzed by CMT2 through a DDM1-mediated process (Zemach et al. 2013; Stroud et al. 2014), we examined expression of *GhCMT2*, *GhDDM1* and other genes related to DNA methylation in leaves, ovules and fibers. Compared with leaves, *GhCMT2*, *GhCMT3* and *GhMET1* were up-regulated in ovules and fibers, and the genes involved in the RdDM pathway, including *DRM1*, *DMR2* and *RDR2* with an exception of *AGO4*, were up-regulated in ovules but not in fibers (Fig. 4D). However, cotton *Histone H1*, which is shown to impede the access of DNA methyltransferases to the heterochromatin (Wierzbicki and Jerzmanowski 2005; Zemach et al. 2013), was repressed only in the fiber but not in the ovule (Fig. 4F). The data suggest that up-regulation of *GhCMT2* and repression of cotton *HISTONE H1* may induce CHH hypermethylation in TEs and heterochromatin in fibers.

Influence of DNA methylation to expression of homoeologous genes

Methylation in the gene body and promoter regions has different effects on expression of homoeologous genes. In the gene body, CG methylation levels of most homoeologous genes were relatively equal (Fig. 5A) and did not correlate with homoeologous gene expression levels (Fig. 5B). However, we identified 533 pairs of A- and D-homoeologs that were differentially methylated at CHG and CHH sites, although non-CG methylation levels of most homoeologs were relatively low (Fig. 5A). The methylation levels of CHG or CHH methylation in the gene body were significantly anti-correlated with expression levels of A- and D-homoeologous alleles (Fig. 5B,C; Supplemental Table S5), indicating repression function of CHG and CHH methylation in expression variation of these homoeologous genes.

In the promoter regions, although more TEs exist in the A-subgenome (68%) than in the D-subgenome (61%), more genes in the D-subgenome than in the A-subgenome had higher methylation levels in CG, CHG and CHH sites (Fig. 5D). This is consistent with more TEs and siRNAs that were distributed in promoter regions of D-homoeologs than those of A-homoeologs ($P < 1e^{-80}$, Wilcoxon rank-sum test) (Fig. 5E). Homoeologous genes with higher CG or CHG methylation in promoter regions tended to be expressed at lower levels (Fig. 5F). For example, among 10 members of the *GhMYB25* family (homologs of *AtMYB25* or *AtMYB106*) (Machado et al. 2009; Walford et al. 2011), the D-homoeologs were methylated at higher levels in CG or CHG sites of promoter regions than the A-homoeologs and were also expressed at lower levels than the A-homoeologs (Fig. 5G). These data suggest a role for DNA methylation in the expression variation of homoeologous genes, which may contribute to fiber selection and improvement in the allotetraploid cotton.

DISCUSSION

We found little variation of CG and CHG methylation among cotton tissues, which is consistent with the data in maize, showing more variable CG and CHG methylation levels among genotypes than among tissues (Eichten et al. 2013). Although CHH methylation is low and cannot be effectively detected by microarrays (Eichten et al. 2013), it links to biological functions. In *Arabidopsis* intraspecific hybrids, CHH methylation changes are associated with the parent-of-origin effect on circadian clock gene expression and biomass heterosis (Ng et al. 2014). CHH methylation is mainly produced by 24-nt siRNAs through the RdDM pathway (Law and Jacobsen 2010; Haag and Pikaard 2011). In cotton, 24-nt siRNAs were enriched in gene-rich regions, which is similar to that in maize (Gent et al. 2013) and sugar beet (Dohm et al. 2013) but different from that in *Arabidopsis* (Ha et al. 2009) and soybean (Song et al. 2013). Enrichment of siRNAs in euchromatic regions is inconsistent with the overall repeat abundance in the genome because the repeat amount is higher in soybean (~59%) (Schmutz et al. 2010) than in sugar beet (~42%) (Dohm et al. 2013), but much lower than in maize (~85%) (Schnable et al. 2009). It is the TE property and location not the TE density that determines location of siRNA abundance and CHH methylation. These 24-nt siRNAs are derived from short TEs close to the genes, inducing CHH methylation flanking the genes, which is known as “CHH methylation island” in maize (Gent et al. 2013). However, in spite of similar siRNA distribution patterns between maize and cotton, CHH methylation patterns are different. In maize, siRNAs in gene-rich regions are associated with higher CHH methylation levels; but in cotton, CHH methylation levels were negatively correlated with the gene density and siRNA abundance (Fig. 1F). Furthermore, the percentage of methylcytosines in the CHH context is higher in cotton (~7.7%

in leaves) than in maize (~5% in ears and shoots), which is in contrast to a higher TE density in maize than in cotton. These results indicate that in addition to CHH methylation that is produced by the RdDM pathway in cotton as in maize, cotton has another active pathway to generate CHH methylation.

5 This CHH hypermethylation of many TEs in fibers is likely mediated by CMT2 and DDM1 (Zemach et al. 2013). The genes overlapping with FO CHH-hyper DMRs were not enriched in the differentially expressed genes, which is consistent with the finding that non-CG methylation generated by CMT2 does not regulate protein-coding genes (Stroud et al. 2014). The CHH hypermethylation in cotton fibers is probably promoted by both induction of *GhCMT2* and
10 repression of *HISTONE H1*. In ovules, *GhCMT2* is also induced but *HISTONE H1* is not up-regulated. As a result, CHH methylation was not enhanced in the heterochromatin of ovules, suggesting that chromatin changes by histone H1 is required for promoting CHH hypermethylation in heterochromatin between fibers and ovules. Fiber undergoes endoreduplication in early stages and rapid cell elongation and cellulose synthesis in later
15 stages (Van't Hof 1999; Wu et al. 2006). This dramatic change in the cellular environment (rapid elongation without cell division) of fibers could “elongate” or “loosen” chromatin structures in cell nucleus, leading to activation of TEs. Indeed, more TEs are active in fibers than in ovules and leaves (Fig. 2), and more siRNAs were generated in fibers than other tissues (Pang et al. 2009). The CHH hypermethylation induced in the fibers could function as a feedback
20 mechanism to repress TEs. Consistent with the requirement of DNA methylation for cotton fiber development, inhibiting DNA methylation by aza-dC not only reduces fiber cell initials but also slows down fiber cell elongation (Fig. 1). A confirmatory experiment is to generate the

transgenic plants that suppress CHH methylation and directly test the CHH methylation effect on fiber development.

Notably, the highly expressed genes are correlated with higher CHH methylation levels in promoter regions close (1-kb or less) to the transcription start site (TSS), which is unexpected but has been documented in maize (Gent et al. 2013), and soybean (Song et al. 2013), and now in cotton (Fig. 3). One possibility is that TE activation enhances expression of nearby genes. Transcription initiates from TEs through Pol II or Pol IV and then spreads to nearby genes (Zheng et al. 2009; Haag and Pikaard 2011). While most TEs are transcriptionally silenced in plant genomes, some TEs could activate nearby genes, as reported in *Arabidopsis* (Wang and Warren 2010), rice (Naito et al. 2009), and wheat (Kashkush et al. 2003). If TE activation occurs prior to the transcription of nearby genes, TE expression should not be correlated with the distance between TEs and genes. However, more 24-nt siRNAs are present in the TEs closer to the genes (Fig. 1J), suggesting another possibility that gene expression induces activation of nearby TEs. This may contribute to positive correlation between CHH methylation in promoters and gene expression. The regions near the TSS are probably in open chromatin formation to allow active transcription of both genes by RNA polymerase II and short TEs by RNA polymerase IV, a homolog of Pol II (Herr et al. 2005; Onodera et al. 2005; Zheng et al. 2009). This leads to high abundance of small RNAs near the promoters and transcripts from corresponding genes, as observed in cotton, maize (Gent et al. 2013), and soybean (Song et al. 2013). The siRNAs can induce CHH methylation through the RdDM pathway. Consistent with this hypothesis, ovule up-regulated genes were enriched in those overlapping with OL CHH-hyper DMRs in flanking sequences. However, further CHH methylation of OL CHH-hyper DMRs in fibers represses

nearby genes. This suggests that CHH methylation in promoters may act as a feedback mechanism to regulate these genes during ovule and fiber development. The spatiotemporal role of DNA methylation in expression changes of ovule- and fiber-related genes could also explain why overexpressing these genes may result in the unexpected outcome of fiber traits (Walford et al. 2011).

Finally, non-CG methylation in the gene body is associated with the expression bias of homoeologous genes in the allotetraploid cotton, providing the unique evidence for epigenetic regulation of non-additive expression of homoeologous genes in polyploid species (Liu and Wendel 2003; Chen 2007). Although *G. arboreum* (A-genome progenitor) (Li et al. 2014) contains more TEs than *G. raimondii* (D-genome progenitor) (Paterson et al. 2012), methylation levels and TE densities in promoters of many ovule- and fiber-related genes are higher in the D-homoeologs than in the A-homoeologs. Consequently, A-homoeologs of ovule- and fiber-related genes are expressed at higher levels than their D-homoeologs. This could reflect an evolutionary difference between fiber-bearing A-genome progenitors and nearly fiberless D-genome progenitors. Together, these results will advance our understanding of the biological significance of CHH methylation in developmental regulation and intergenomic interactions, which could be translated into genomic and biotechnological improvement of in polyploid plants, including most important crops that provide us with food (wheat), fiber (cotton), and oil (canola).

20

METHODS

Plant material

Gossypium hirsutum acc. TM-1 grew in the greenhouse at The University of Texas at Austin.

Leaves and bolls (0 and 14 DPA) were harvested with three biological replications. Ovules and fibers were carefully dissected from cotton bolls at 14 DPA and immediately frozen in liquid nitrogen for RNA and DNA preparation.

5 **Small RNA-seq library construction**

Total RNA was isolated from cotton leaves, ovules at 0 DPA and fibers at 14 DPA using Plant RNA Reagent (Life Technologies). By polyacrylamide gel electrophoresis, RNAs corresponding to ~15 to 30-nt in length from 20 µg total RNA were excised and eluted from the gel using 0.3 M NaCl. Small RNAs were precipitated using ethanol and dissolved in 6 µl RNase free water. Small RNA-seq libraries with three biological replicates were constructed using NEBNext® Multiplex Small RNA Library Prep Set (NEB, Ipswich, Massachusetts) according to the manufacturer's instructions. Small RNA-seq libraries were single-end sequenced for 50 cycles.

mRNA-seq library construction

After DNase treatment, total RNA (~5 µg) was subjected to construct strand-specific mRNA-seq libraries with two biological replications using NEBNext® Ultra™ Directional RNA Library Prep Kit (NEB, Ipswich, Massachusetts) according to the manufacturer's instructions. mRNA-seq libraries were single-end sequenced for 150 cycles.

MethylC-seq library construction

Genomic DNA was isolated from cotton leaves, ovules at 0 DPA and 14 DPA, and fibers at 14 DPA using CTAB method (Mei et al. 2004). Total genomic DNA (~5 µg) was fragmented to 100-1000 bp using Bioruptor (Diagenode, Denville, New Jersey). End repair (NEBNext® End Repair Module) was performed on the DNA fragment by adding an 'A' base to the 3'end (NEBNext®

dA-Tailing Module), and the resulting DNA fragment was ligated to the methylated DNA adapter (NEXTflex™ DNA Barcodes, Bioo Scientific, Austin, Texas). The adapter-ligated DNA of 200–400 bp was purified using AMPure beads (Beckman Coulter, Brea, California), followed by sodium bisulfite conversion using MethylCode™ Bisulfite Conversion Kit (Life Technologies, Foster City, California). The bisulfite-converted DNA was amplified by 12 cycles of PCR using LongAmp® *Taq* DNA Polymerase (NEB, Ipswich, Massachusetts) and purified using AMPure beads (Beckman Coulter, Brea, California). The Paired-End sequencing of the MethylC-seq libraries was performed for 101 cycles.

qRT-PCR

10 After DNase treatment, total RNA (2 µg) was used to produce first-strand cDNA with the Omniscript RT Kit (Qiagen, Valencia, California). The cDNA was used as the template for qRT-PCR using FastStart Universal SYBR Green Master (Roche, Indianapolis, India). The reaction was run on the LightCycler® 96 System (Roche, Pleasanton, California). The relative expression level was quantified using internal control cotton *HISTONE H3*.

15 Reference genome generation

Genome sequences of *G. arboreum* (Li et al. 2014) and *G. raimondii* were respectively downloaded from <http://cgp.genomics.org.cn> and <http://phytozome.jgi.doe.gov> (Paterson et al. 2012), respectively. DNA sequencing data of *G. hirsutum* cv. Acala Maxxa (SRR617482) was downloaded from NCBI (Paterson et al. 2012). The reads were mapped to genome sequences of *G. arboreum* and *G. raimondii* using Bowtie2 with default parameter and only mismatches with high base quality score (≥ 30) were used for variant calling (Langmead and Salzberg 2012). Calling variants was performed by SAMtools (Li et al. 2009). A SNP was identified when the

identity was more than 90% and read depth was over 10. We substituted each SNP in genome sequence of *G. arboreum* and *G. raimondii* to construct *G. hirsutum* genome, which was used as reference genome for aligning all high-throughput sequencing data in this study.

Small RNA-seq data analysis

5 After adapter clipping, small RNA-seq reads (18-30 nt) were mapped to reference genome using Bowtie2 allowing no mismatches. Multi-mapped small RNA-seq reads were evenly weighted and assigned to all locations as previously described method (Lu et al. 2012).

mRNA-seq data analysis

Gene annotation and TE classification of *G. arboreum* and *G. raimondii* were downloaded from
10 <http://cgp.genomics.org.cn> and <http://phytozome.jgi.doe.gov>, respectively. mRNA-seq reads were mapped to reference genome using Tophat (Trapnell et al. 2009). Uniquely mapped reads were extracted and analyzed by Cufflinks to determine gene and TE abundance using annotated genes and TEs (Roberts et al. 2011). The up-regulated or down-regulated genes were identified through double-filtering methods using the fold-change (>2-fold) and ANOVA tests ($P < 0.01$).

15 Identification of methylated cytosines

We applied Bismarck software to align reads to reference genome with default parameters (Krueger and Andrews 2011). In brief, the first 75 bases of unmapped reads were extracted and realigned to the genome. Only reads mapped to the unique sites were retained for further analysis. Reads mapped to the same site were collapsed into a single consensus read to reduce
20 clonal bias. For each cytosine, the binomial distribution was used to identify whether this cytosine was methylated. The probability p in binomial distribution $B(n, p)$ was referred to bisulfite conversion failure rate. The number of trials (n) in the binomial distribution was the

read depth. Only the cytosines covered by at least three reads in all compared tissues were considered for further analysis. Cytosines with P-values below $1e^{-5}$ were identified as methylcytosines.

Differentially methylated regions

5 Differentially methylated regions (DMRs) for CHH methylation were identified using 100-bp sliding-windows. Mean methylation level was calculated for each window (Schultz et al. 2012). Windows containing at least eight cytosines in the CHH context covered by at least three reads were selected for identifying DMRs. ANOVA test using 2 biological replications ($P < 0.05$) and difference of methylation levels (cut-off value > 0.1 between two compared samples) were
10 used to determine CHH DMRs between two compared tissues. The cut-off value was set at 0.05 for the comparison between ovule and fiber methylation levels in OL CHH-hyper DMRs.

Homoeologous genes

Protein sequences of *G. arboreum* were aligned to protein sequences of *G. raimondii* using blastp with an E-value less than $1e^{-10}$. Alignment information was used by MCScanx to identify
15 homoeologous genes (Wang et al. 2012) (Score > 2000 and E-value $< 1e^{-10}$).

Z-score for enrichment of DMRs flanking genes

We first randomly extracted 1000 of 1-kb regions in the genome for 500 times. Percentage of DMR-overlapping regions in 1000 random regions was calculated at each time. Z-scores were calculated as $(x-\mu)/\sigma$ where x was percentage of DMR-flanking genes in all genes, μ and σ were
20 mean and standard deviation of percentages of DMR-overlapping regions in random loci.

Cotton ovule in vitro culture and 5-aza-dC treatment

Ovules were removed from flower buds at -3 or 0 DPA. The ovules were sterilized with 75% alcohol, washed with autoclaved water, and cultured in Beasley-Ting (BT) liquid media (20 ml) containing 5 μ M indole-3-acetic acid (IAA) and 0.5 μ M gibberellic acid³ (GA3) at 30°C for 4-14 days in dark (Beasley and Ting 1974; Kim and Triplett 2001). Treatment of 5-aza-2'-deoxycytidine (aza-dC, 10 mg/L) was applied in the BT liquid media.

Data access

All raw data have been submitted to NCBI GEO database under accession number GSE61774.

Acknowledgements

We thank the Genomic Sequencing and Analysis Facility at The University of Texas at Austin for sequencing methylomes, RNA, and small RNA libraries using HiSeq 2500. This work is supported by the grants from the National Science Foundation Plant Genome Research Program (IOS1025947) and the Cotton Incorporated (07-161).

Author contributions

QS and ZJC conceived and designed the project; QS performed the experiments; XG provided materials and reagents; QS and ZJC analyzed data and wrote the paper.

References

- Beasley, C.A. and Ting, I.P. 1974. The effects of plant growth substances on *in vitro* fiber development from unfertilized cotton ovules. *Amer J Bot* **61**: 188-194.
- Bird, A. 1992. The essentials of DNA methylation. *Cell* **70**(1): 5-8.
- Cao, X., Aufsatz, W., Zilberman, D., Mette, M.F., Huang, M.S., Matzke, M., and Jacobsen, S.E. 2003. Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Curr Biol* **13**(24): 2212-2217.
- Cao, X. and Jacobsen, S.E. 2002. Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr Biol* **12**(13): 1138-1144.

- Chen, Z.J. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol* **58**: 377-406.
- Chen, Z.J., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X., Stelly, D.M., Rabinowicz, P.D., Town, C.D. et al. 2007. Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol* **145**(4): 1303-1310.
- 5 Dohm, J.C., Minoche, A.E., Holtgrawe, D., Capella-Gutierrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sorensen, T.R., Stracke, R., Reinhardt, R. et al. 2013. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*.
- Eichten, S.R., Vaughn, M.W., Hermanson, P.J., and Springer, N.M. 2013. Variation in DNA Methylation Patterns is More Common among Maize Inbreds than among Tissues. *Plant Genome-U.S.* **6**(2).
- 10 Flagel, L., Udall, J., Nettleton, D., and Wendel, J. 2008. Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biol* **6**: 16.
- Gent, J.I., Ellis, N.A., Guo, L., Harkess, A.E., Yao, Y., Zhang, X., and Dawe, R.K. 2013. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res* **23**(4): 628-637.
- 15 Guan, X., Song, Q., and Chen, Z.J. 2014. Polyploidy and small RNA regulation of cotton fiber development. *Trends in Plant Science* **19**(8): 516-528.
- Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K.D., Chapman, E.J., Carrington, J.C., Chen, X., Wang, X.J., and Chen, Z.J. 2009. Small RNAs serve as a genetic buffer against genomic shock in *Arabidopsis* interspecific hybrids and allopolyploids. *Proc Natl Acad Sci USA* **106**(42): 17835-17840.
- 20 Haaf, T., Werner, P., and Schmid, M. 1993. 5-Azadeoxycytidine distinguishes between active and inactive X chromosome condensation. *Cytogenetics & Cell Genetics* **63**(3): 160-168.
- Haag, J.R. and Pikaard, C.S. 2011. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat Rev Mol Cell Biol* **12**(8): 483-492.
- 25 Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. 2005. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308**(5718): 118-120.
- Jeddeloh, J.A., Stokes, T.L., and Richards, E.J. 1999. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat Genet* **22**(1): 94-97.
- Jin, X., Pang, Y., Jia, F., Xiao, G., Li, Q., and Zhu, Y. 2013. A potential role for CHH DNA methylation in cotton fiber growth patterns. *PLoS ONE* **8**(4): e60547.
- 30 Kankel, M.W., Ramsey, D.E., Stokes, T.L., Flowers, S.K., Haag, J.R., Jeddeloh, J.A., Riddle, N.C., Verbsky, M.L., and Richards, E.J. 2003. *Arabidopsis* MET1 cytosine methyltransferase mutants. *Genetics* **163**(3): 1109-1122.
- Kashkush, K., Feldman, M., and Levy, A.A. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* **33**(1): 102-106.
- 35 Kim, H.J. and Triplett, B.A. 2001. Cotton fiber growth in planta and in vitro: Models for plant cell elongation and cell wall biogenesis. *Plant Physiol* **127**(4): 1361-1366.
- Krueger, F. and Andrews, S.R. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**(11): 1571-1572.
- 40 Langmead, B. and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4): 357-U354.
- Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M., Strahl, B.D., Patel, D.J., and Jacobsen, S.E. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**: 385-389.
- 45 Law, J.A. and Jacobsen, S.E. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**(3): 204-220.
- Li, E., Bestor, T.H., and Jaenisch, R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**(6): 915-926.

- Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., Zou, C. et al. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nature Genetics* **46**(6): 567-572.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Proc, G.P.D. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
- Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., and Jacobsen, S.E. 2001. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**(5524): 2077-2080.
- Liu, B. and Wendel, J.F. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol* **29**(3): 365-379.
- Lu, J., Zhang, C., Baulcombe, D.C., and Chen, Z.J. 2012. Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of Arabidopsis seeds. *Proc Natl Acad Sci USA* **109**(14): 5529-5534.
- Machado, A., Wu, Y., Yang, Y., Llewellyn, D.J., and Dennis, E.S. 2009. The MYB transcription factor GhMYB25 regulates early fibre and trichome development. *Plant J* **59**(1): 52-62.
- Mei, M., Syed, N.H., Gao, W., Thaxton, P.M., Smith, C.W., Stelly, D.M., and Chen, Z.J. 2004. Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*). *Theor Appl Genet* **108**(2): 280-291.
- Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**(7267): 1130-1134.
- Ng, D.W., Miller, M., Yu, H.H., Huang, T.Y., Kim, E.D., Lu, J., Xie, Q., McClung, C.R., and Chen, Z.J. 2014. A Role for CHH Methylation in the Parent-of-Origin Effect on Altered Circadian Rhythms and Biomass Heterosis in *Arabidopsis* Intraspecific Hybrids. *Plant Cell* **26**(6): 2430-2440.
- Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O., and Pikaard, C.S. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**(5): 613-622.
- Osabe, K., Clement, J.D., Bedon, F., Pettolino, F.A., Ziolkowski, L., Llewellyn, D.J., Finnegan, E.J., and Wilson, I.W. 2014. Genetic and DNA methylation changes in cotton (*Gossypium*) genotypes and tissues. *PLoS one* **9**(1): e86049.
- Pang, M., Woodward, A.W., Agarwal, V., Guan, X., Ha, M., Ramachandran, V., Chen, X., Triplett, B.A., Stelly, D.M., and Chen, Z.J. 2009. Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (*Gossypium hirsutum* L). *Genome Biol* **10**(11): R122.
- Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J. et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**(7429): 423-427.
- Richards, E.J. 1997. DNA methylation and plant development. *Trends Genet* **13**(8): 319-323.
- Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology* **12**(3).
- Schmitz, R.J., He, Y., Valdes-Lopez, O., Khan, S.M., Joshi, T., Urich, M.A., Nery, J.R., Diers, B., Xu, D., Stacey, G. et al. 2013. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* **23**(10): 1663-1674.
- Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J. et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**(7278): 178-183.

- Schnable, P.S. Ware, D. Fulton, R.S. Stein, J.C. Wei, F. Pasternak, S. Liang, C. Zhang, J. Fulton, L. Graves, T.A. et al. 2009. The b73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956): 1112-1115.
- Schultz, M.D., Schmitz, R.J., and Ecker, J.R. 2012. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28**(12): 583-585.
- 5 Song, Q.X., Lu, X., Li, Q.T., Chen, H., Hu, X.Y., Ma, B., Zhang, W.K., Chen, S.Y., and Zhang, J.S. 2013. Genome-Wide Analysis of DNA Methylation in Soybean. *Molecular Plant* **6**(6): 1961-1974.
- Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J., and Jacobsen, S.E. 2014. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nat Struct Mol Biol* **21**(1): 64-72.
- 10 Trapnell, C., Pachter, L., and Salzberg, S.L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.
- Van't Hof, J. 1999. Increased nuclear DNA content in developing cotton fiber cells. *Am J Bot* **86**: 776-779.
- Walford, S.A., Wu, Y., Llewellyn, D.J., and Dennis, E.S. 2011. GhMYB25-like: a key factor in early cotton fibre development. *Plant J* **65**(5): 785-797.
- 15 Wang, Y.H. and Warren, J.T., Jr. 2010. Mutations in retrotransposon AtCOPIA4 compromises resistance to *Hyaloperonospora parasitica* in *Arabidopsis thaliana*. *Genet Mol Biol* **33**(1): 135-140.
- Wang, Y.P., Tang, H.B., DeBarry, J.D., Tan, X., Li, J.P., Wang, X.Y., Lee, T.H., Jin, H.Z., Marler, B., Guo, H. et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**(7).
- 20 Wassenegger, M., Heimes, S., Riedel, L., and Sanger, H.L. 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**(3): 567-576.
- Wendel, J.F. 1989. New World tetraploid cottons contain Old World cytoplasm. *Proc Natl Acad Sci USA* **86**(11): 4132-4136.
- 25 Wierzbicki, A.T. and Jerzmanowski, A. 2005. Suppression of histone H1 genes in Arabidopsis results in heritable developmental defects and stochastic changes in DNA methylation. *Genetics* **169**(2): 997-1008.
- Wu, Y., Machado, A.C., White, R.G., Llewellyn, D.J., and Dennis, E.S. 2006. Expression profiling identifies genes expressed early during lint fibre initiation in cotton. *Plant Cell Physiol* **47**(1): 107-127.
- 30 Yoo, M.J. and Wendel, J.F. 2014. Comparative evolutionary and developmental dynamics of the cotton (*Gossypium hirsutum*) fiber transcriptome. *PLoS Genet* **10**(1): e1004073.
- Zemach, A., Kim, M.Y., Hsieh, P.H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. 2013. The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. *Cell* **153**(1): 193-205.
- 35 Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**(5980): 916-919.
- Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. et al. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **126**(6): 1189-1201.
- 40 Zheng, B., Wang, Z., Li, S., Yu, B., Liu, J.Y., and Chen, X. 2009. Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. *Genes Dev* **23**(24): 2850-2860.
- Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39**(1): 61-69.
- 45

Figure legends

Figure 1. Effects of DNA methylation on ovule and fiber development and genome-wide distribution of DNA methylation in leaves, ovules, and fibers. (A) Scanning electron micrograph (SEM) showing shorter and fewer fibers in ovules treated by 5-aza-dC. Ovules at -3--4 DPA were cultured *in vitro* for 4 days with and without 5-aza-dC (10mg/L). (B) Shorter fibers in ovules after the aza-dC treatment for 14 days. (C) Quantitative analysis of fiber length in (B) (student's test, $p < 0.01$). (D) Percentage of methylated cytosine (mC) in leaves, ovules, and fibers. A and D indicate A- and D-subgenomes in the allotetraploid cotton, respectively. (E) Size distribution of small RNAs in different tissues. (F) Circle plots of fiber CG, CHG, CHH methylation, fiber small RNA density, gene density, TE density, and ovule-leaf and fiber-ovule CHH methylation among 12 A-homoeologous chromosomes (upper) and 13 D-homoeologous chromosomes (lower). The color bar indicates the same scale that was used from low (green) to high (red) densities. (G) Distribution of CHH methylation in genes. (H) Distribution of CHH methylation in TEs. (I) CHH methylation levels in TEs relative to the distance from the nearest gene. (J) siRNA levels (per bp per million reads) in TEs relative to the distance from the nearest gene.

Figure 2. Genomic distribution of differentially methylated regions (DMRs). (A) Percentage of DMRs in TE (brown), intron (dark yellow), exon (blue), and intergenic region (green). (B) Percentage of DMRs in TEs with different sizes. (C) Heatmap of CHH methylation levels in OL CHH-hyper DMRs in leaves, ovules and fibers. (D) Kernel density plot of 24-nt siRNA fold change

(ovule vs. leaf or fiber vs. ovule) in OL CHH-hyper DMRs and FO CHH-hyper DMRs. (E)

Percentage of CG, CHG and CHH methylation in OL CHH-hyper DMRs and FO CHH-hyper DMRs.

Figure 3. Relationship of ovule-leaf (OL) CHH-hyper DMRs and gene expression changes

5 between the ovule and leaf. (A) Fold-enrichment of differentially expressed genes (DEGs) in those overlapping with OL CHH-hyper DMRs in flanking sequences or gene body, relative to all genes. The hypergeometric test was used to infer statistical significance (*: $P < 0.05$; **: $P < 1e^{-10}$). (B) CHH methylation distribution in genes of ovule-preferred and leaf-preferred genes in the leaf and ovule. Leaf-preferred and ovule-preferred genes were identified by both fold-change of
10 gene expression (> 5 -fold) and ANOVA test (< 0.01). (C) Relationships between CHH (left), CHG (middle) and CG (right) methylation and expression levels of genes in ovules, which were divided into four quartiles (high, red; mid-high, green; mid-low, blue; and low, yellow). (D) The relationship between siRNA density and the genes in the four quartiles as in (C). (E) Box plot showing siRNA abundance in promoter regions of ovule-preferred and leaf-preferred genes.

15

Figure 4. Influence of fiber-ovule (FO) CHH-hyper DMRs to gene expression and TE activity. (A)

Fold-enrichment of DEGs between the fiber and ovule in the genes overlapping with OL CHH-hyper DMRs, which were further hypermethylated in fibers, relative to all genes. The hypergeometric test was used to infer statistical significance (*: $P < 0.05$; **: $P < 1e^{-10}$). (B) An
20 example of fiber CHH hypermethylation correlating with gene repression in the fiber. (C) fold-enrichment of DEGs between the fiber and ovule in the genes overlapping with FO CHH-hyper DMR in the gene body and flanking sequences, relative to all genes. (D) Enrichment (z-score) of

DMRs in flanking sequences (1 kb) of genes. The dash line indicates the statistical significant level ($P < 0.05$). (E) Number of expressed TEs in the leaf, ovule and fiber. (F) Relative expression levels of genes related to DNA methylation in the leaf, ovule and fiber. Cotton *HISTONE H3* was used as internal control for qRT-PCR analysis. The dash line indicates the expression level that is

5 normalized to the leaf.

Figure 5. DNA methylation differences in homoeologous genes. (A) Pairwise plots of CG (left), CHG (middle), and CHH (right) methylation levels (Y-axis, D-homologs; X-axis, A-homoeologs) in the gene body for homoeologous genes. Blue and red dots indicate the expression level of an A-homoeolog that is lower and higher than that of the corresponding D-homoeolog, respectively.

10 (B) Number of homoeologous genes that were differentially methylated in the gene body and showed expression differences (A < D, blue; A > D, red). The levels of CG (>0.4), CHG (>0.4), and CHH (>0.1) methylation difference and fold-changes of expression (>2-fold) between A- and D-homoeologs were used as cut-off values. (C) An example of hypermethylation in the gene body

15 correlating with repression of the A-homoeolog. Colors indicate CG (green), CHG (blue), and CHH (red) methylation, and gene expression (black). (D) Pairwise plots of CG (left), CHG (middle), and CHH (right) methylation levels (Y-axis, D-homologs; X-axis, A-homoeologs) in promoter regions (2-kb) of homoeologous genes. Same colors were used as in (A). (E) Box plot showing more TEs and siRNAs in promoters of homoeologous genes derived from the D-

20 subgenome than those from A-subgenome. (F) Number of homoeologous genes that were differentially methylated in CG (left), CHG (middle), and CHH (right) sites in promoter regions and showed expression differences (A<D, blue; A>D, red). Same cut-off values were used as in

(B). (G) CG and CHG methylation levels in promoter regions of *GhMYB25* family genes were anti-correlated with expression bias of A- and D-homoeologs. FPKM: Fragments per kilobase of transcripts per million mapped reads. Brackets indicate pairs of homoeologs.

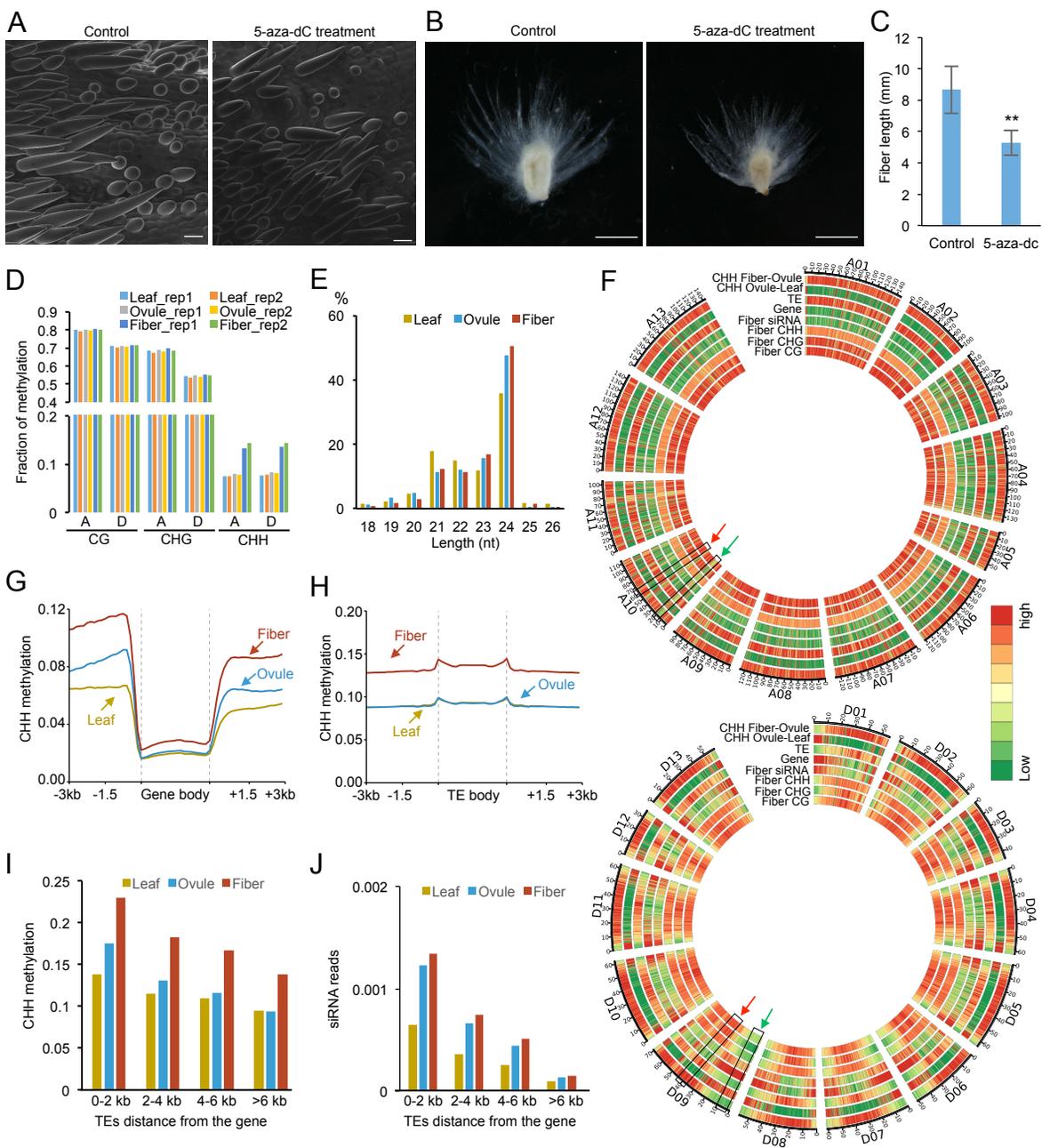


Figure 1

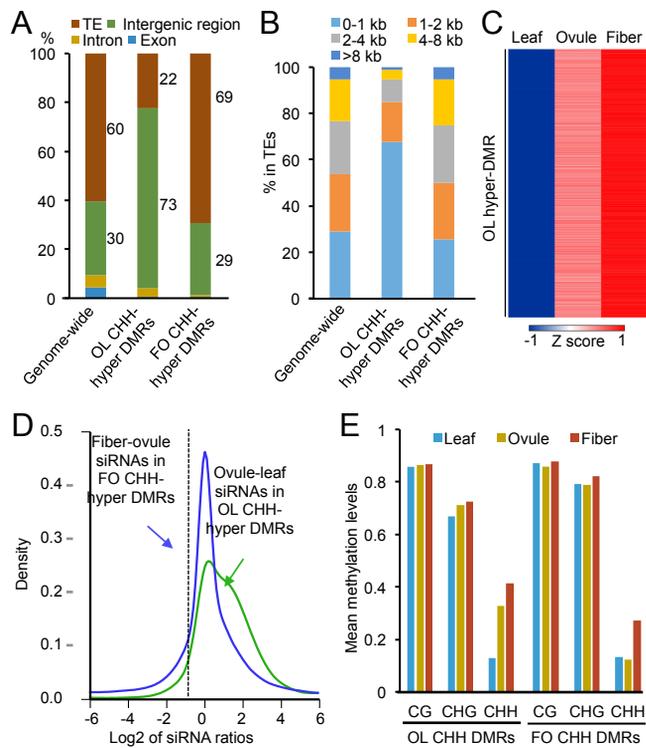


Figure 2

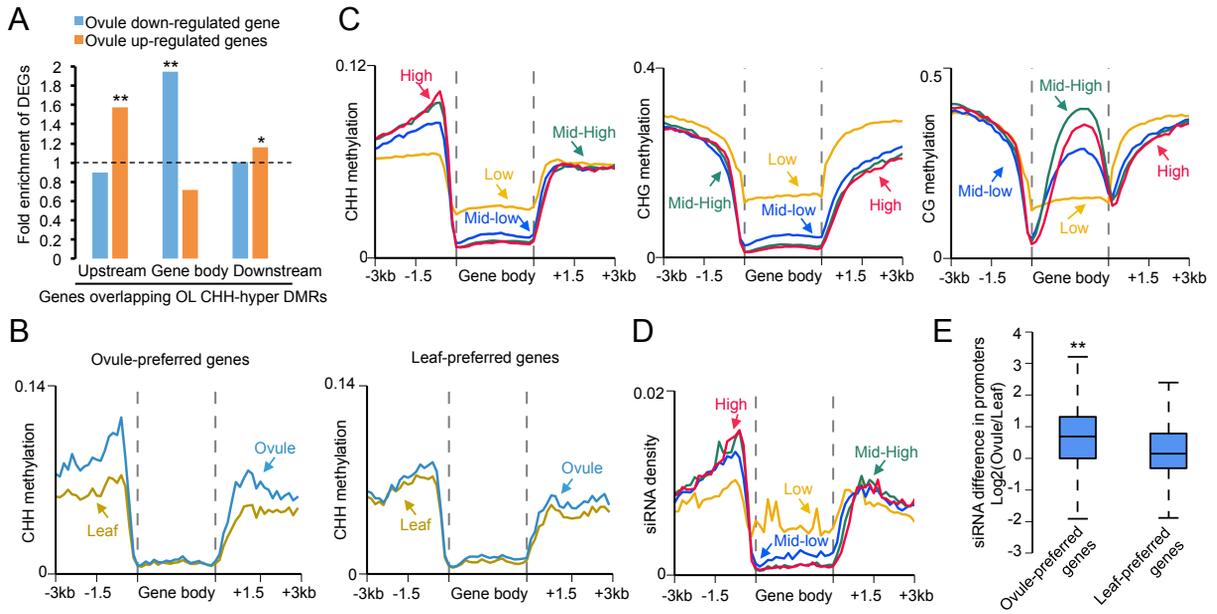


Figure 3

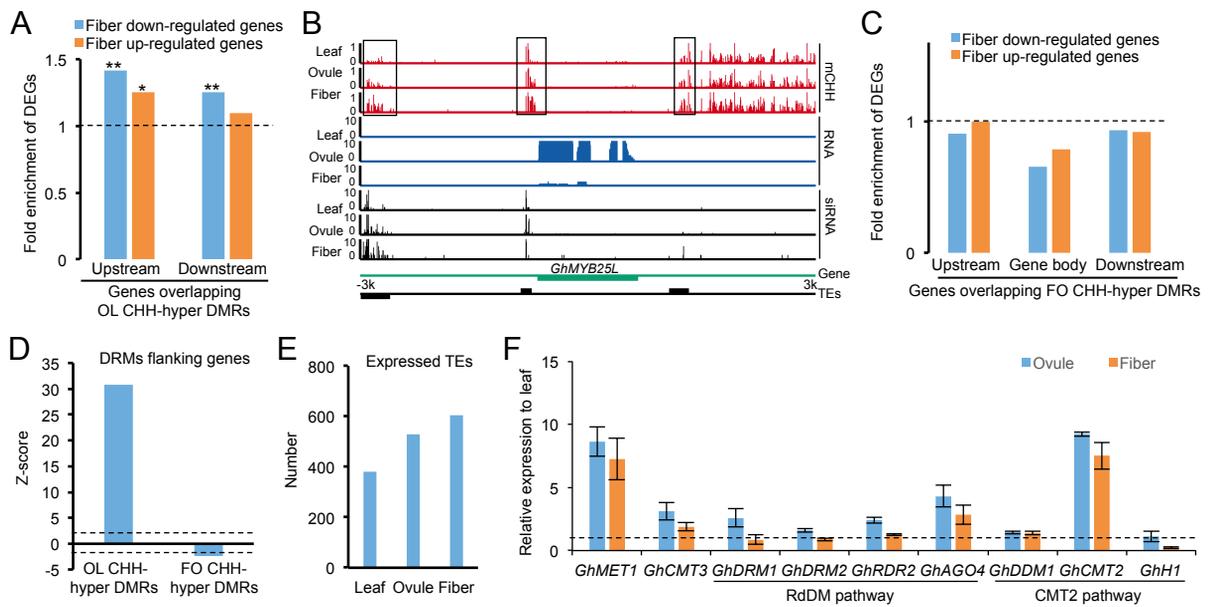


Figure 4

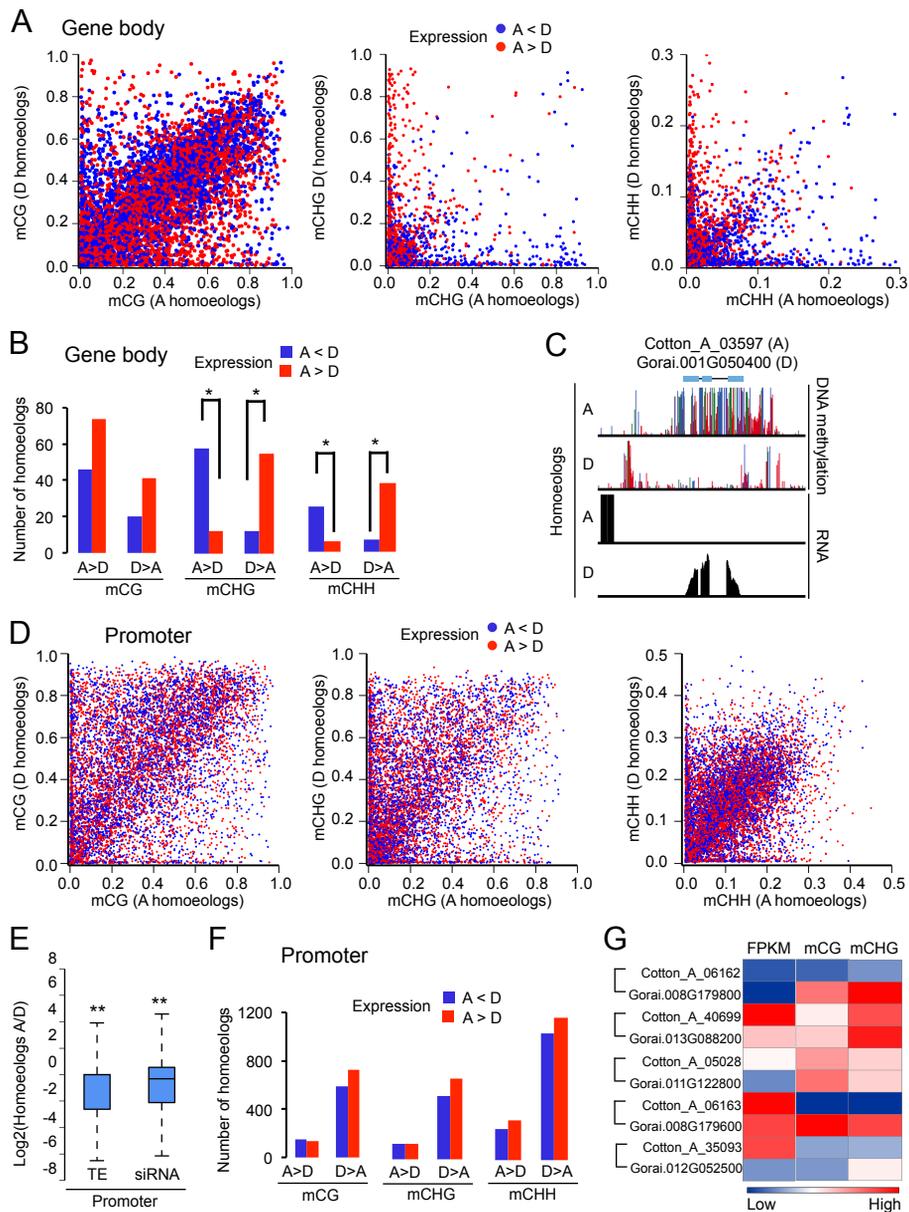


Figure 5