Genome Biology

CrossMark

Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons

Qingxin Song[1], Tianzhen Zhang[2], David M. Stelly[3] and Z. Jeffrey Chen[1,2]*

Abstract

Background: Polyploidy is a pervasive evolutionary feature of all flowering plants and some animals, leading to genetic and epigenetic changes that affect gene expression and morphology. DNA methylation changes can produce meiotically stable epialleles, which are transmissible through selection and breeding. However, the relationship between DNA methylation and polyploid plant domestication remains elusive.

Results: We report comprehensive epigenomic and functional analyses, including ~12 million differentially methylated cytosines in domesticated allotetraploid cottons and their tetraploid and diploid relatives. Methylated genes evolve faster than unmethylated genes; DNA methylation changes between homoeologous loci are associated with homoeolog-expression bias in the allotetraploids. Significantly, methylation changes induced in the interspecific hybrids are largely maintained in the allotetraploids. Among 519 differentially methylated genes identified between wild and cultivated cottons, some contribute to domestication traits, including flowering time and seed dormancy. *CONSTANS* (*CO*) and *CO-LIKE* (*COL*) genes regulate photoperiodicity in *Arabidopsis*. *COL2* is an epiallele in allotetraploid cottons. *COL2A* is hypermethylated and silenced, while *COL2D* is repressed in wild cottons but highly expressed due to methylation loss in all domesticated cottons tested. Inhibiting DNA methylation activates *COL2* expression, and repressing *COL2* in cultivated cotton delays flowering.

Conclusions: We uncover epigenomic signatures of domestication traits during cotton evolution. Demethylation of *COL2* increases its expression, inducing photoperiodic flowering, which could have contributed to the suitability of cotton for cultivation worldwide. These resources should facilitate epigenetic engineering, breeding, and improvement of polyploid crops.

Keywords: DNA methylation, Epigenomics, Cotton, Photoperiod, Hybrid, Polyploidy, Biotechnology

Background

Polyploidy or whole genome duplication (WGD) is a pervasive evolutionary feature of some animals and all flowering plants [1, 2], leading to genetic and epigenetic changes that affect gene expression and morphology [3–5]. Estimates indicate that two rounds of ancestral WGD occurred before the divergence of extant seed plants and angiosperms, giving rise to the diversification of genes and pathways important to seed and flower development and eventually the dominance of angiosperms on the earth [6, 7]. Many important crops including wheat, cotton, and canola are allopolyploids, which usually arise via fusion of $2n$ gametes between species or by interspecific hybridization followed by genome doubling [3, 8]. Genomic interactions in the polyploids can induce genetic and epigenetic changes including DNA methylation [1, 3]. DNA methylation changes can produce meiotically stable epialleles [9, 10] which are

* Correspondence: zjchen@austin.utexas.edu
[1]Department of Molecular Biosciences, Institute for Cellular and Molecular Biology, and Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX 78712, USA
[2]State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China
Full list of author information is available at the end of the article

Song *et al. Genome Biology* (2017) 18:99

Page 2 of 14

transmissible through natural selection and breeding. For example, stable DNA methylation in promoters can be inherited as epialleles, which confer symmetric flower development in *Linaria vulgaris* [11] and quantitative trait loci of colorless non-ripening and vitamin E content in tomato [12, 13]. In plants, DNA methylation occurs in CG, CHG, and CHH (H = A, T, or C) contexts through distinct pathways [14]. In *Arabidopsis*, maintenance methylation of CG and CHG is regulated by *METHYLTRANSFERASE1* (*MET1*) and *CHROMOMETHYLASE3* (*CMT3*), respectively [15–17]. De novo CHH methylation is established by RNA-directed DNA methylation (RdDM) and CMT2-mediated pathways [18–20]. DNA methylation is essential for maintaining animal and plant development. Methylation defects are embryonic lethal in animals [21], induce additional epigenetic changes during self-pollination in *Arabidopsis* [22], and lead to lethality in rice [23]. DNA methylation is also responsible for seed development [24] and adaptation to environments [25]. Furthermore, DNA methylation changes are associated with expression of homoeologous genes in resynthesized and natural *Arabidopsis* allotetraploids [26–28], natural *Spartina* allopolyploids [29], and paleopolyploid beans [30]. However, epigenomic resources in polyploids are very limited, and the functional role of epialleles in morphological evolution and crop domestication remains largely unknown.

Cotton is the largest source of renewable textile fiber and an excellent model for studying polyploid evolution and crop domestication [31, 32]. Allotetraploid cotton was formed approximately 1– 1.5 million years ago (MYA) [33] by interspecific hybridization between two diploid species, one having the A genome like in *Gossypium arboreum* (Ga, A2) and *Gossypium herbaceum* (A1), and the other resembling the D5 genome found in extant species *Gossypium raimondii* (Gr); divergence of A-genome and D-genome ancestors is estimated at ~6 MYA (Fig. 1a). The allotetraploid diverged into five or more species [32, 34]. Two of them, *Gossypium hirsutum* (Gh, Upland cotton) and *Gossypium barbadense* (Gb, Pima cotton), were independently domesticated for higher fiber yield and wider geographical distribution; these characteristics were accompanied by extraordinary morphological changes including loss of photoperiod sensitivity, reduction in seed dormancy, and conversion from tree-like wild species to an annual crop [31, 33, 35].

Here we generated single-base resolution methylomes of domesticated allotetraploid cottons and their wild tetraploid and diploid relatives. More than 12 million differentially methylated cytosines across all species were comparatively analyzed, which revealed different rates of DNA methylation and sequence changes and distinct methylation distributions in transposable elements (TEs) and genes. Integrating the data of methylomes with transcriptomes, we discovered more than 500 putative epialleles that may contribute to morphological and physiological changes during evolution and domestication of polyploid cottons. Genomic and functional analyses of an epiallele confirmed its role in photoperiodic flowering, contributing to wider geographical distribution of cotton. We propose that epigenomic resources can be used to improve crop production by epigenetic engineering and breeding.

Results

Different rates between DNA methylation changes and sequence evolution

To uncover DNA methylation changes during cotton evolution and domestication, we generated single-base resolution methylomes from diploid *G. arboreum* (A2), diploid *G. raimondii* (D5), their interspecific hybrid (A2D5), wild allotetraploid *G. hirsutum* (wGh), wild allotetraploid *G. barbadense* (wGb), allotetraploid *G. tomentosum* (Gt), allotetraploid *G. mustelinum* (Gm), allotetraploid *G. darwinii* (Gd), cultivated allotetraploid *G. hirsutum* (cGh), and cultivated allotetraploid *G. barbadense* (cGb) (Fig. 1a; Additional file 1: Table S1). To exclude the effect of nucleotide variation across species (especially between C and T) on DNA methylation analysis, we identified 352,667,453 conserved cytosines (~48% of the total cytosines of the genome) between all species and present in two biological replicates for further analysis (Additional file 2: Figure S1). Among them, 12,045,718 (~3.4% of) differentially methylated cytosines (DmCs) were found across all species; there were more DmCs between diploid cottons and tetraploid cottons (diploid vs. tetraploid) than for other comparisons (diploid vs. diploid cottons, wild tetraploid vs. wild tetraploid, and wild vs. cultivated cottons) (Fig. 1b).

Methylation divergent levels at CG and non-CG sites, respectively, that were conserved among all species (Additional file 2: Figure S1) were used to generate neighbor-joining phylogenetic trees. Phylogenetic trees with CG and non-CG sites recapitulated the known evolutionary relationships of cotton species [32], including sister taxa relationships between *G. hirsutum* and *G. tomentosum* and between *G. barbadense* and *G. darwinii* (Fig. 1c; Additional file 2: Figure S2). This suggests concerted evolution between DNA sequence and methylation changes. Gene-body methylated genes occur largely in CG sites [36] and evolve slowly [37]. To test the relationship between methylation and sequence evolution in genic regions, we divided orthologous genes into CG body-methylated ($P_{CG} < 0.05$) and CG body-unmethylated genes ($P_{CG} > 0.95$) using a binomial test with body-methylation levels [37]. To reduce the effect of non-CG methylation on CG methylation analysis, CHG or CHH body-methylated orthologous genes ($P_{CHG} < 0.05$ or $P_{CHH} < 0.05$) were removed. Among CG body-methylated genes, the percentage of CG methylation
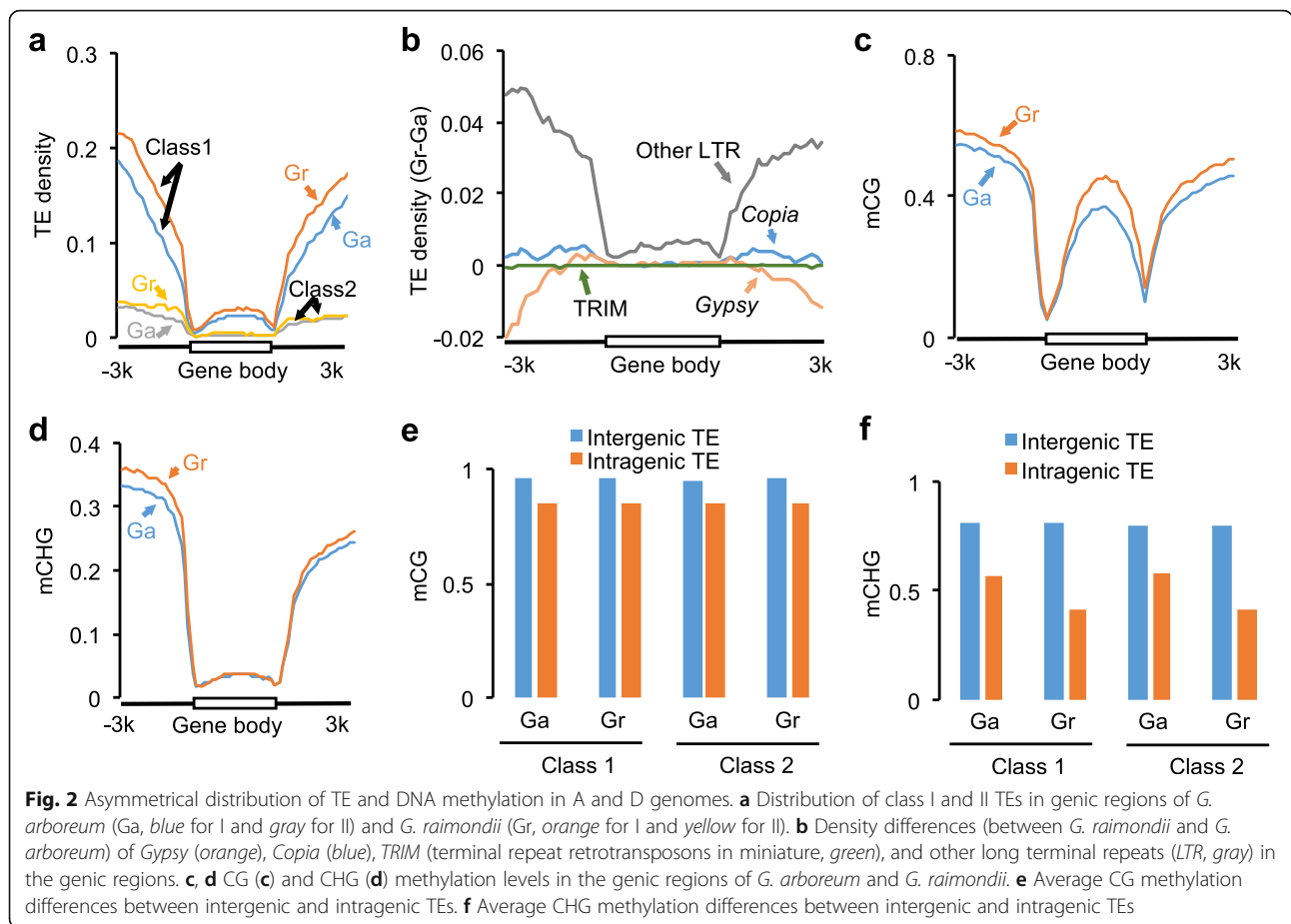
Song *et al. Genome Biology* (2017) 18:99

Page 3 of 14
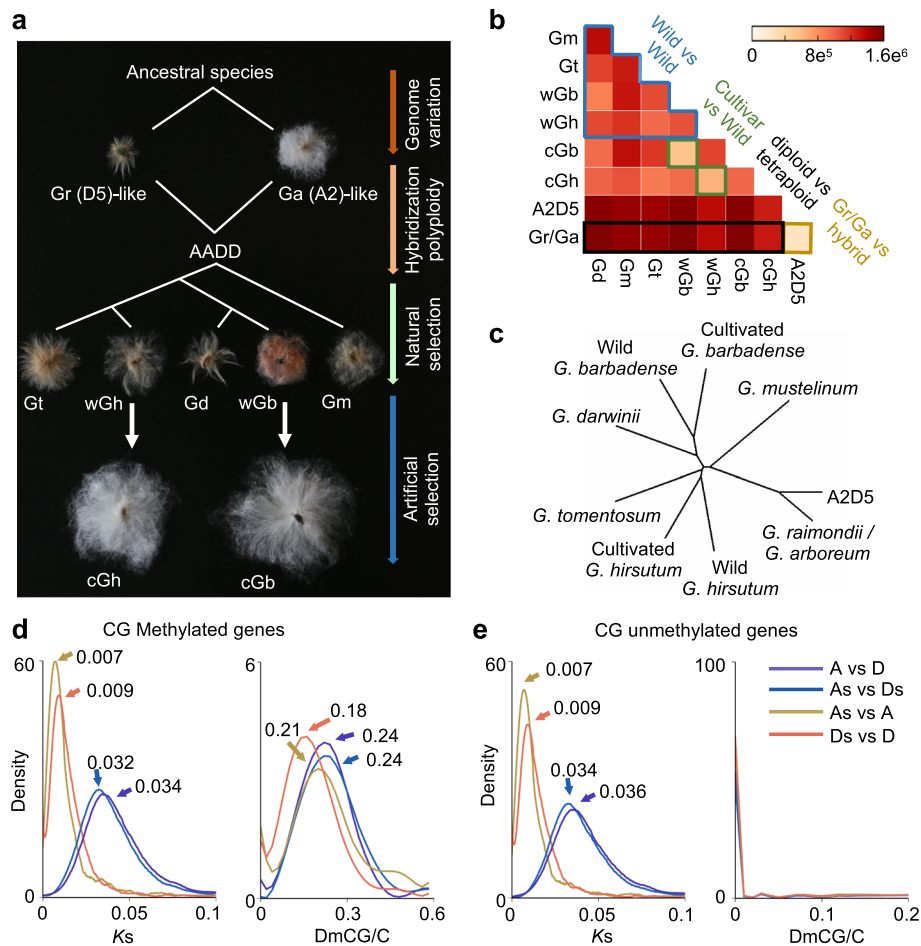
Fig. 1 Evolution of DNA methylation and genome sequence during polyploidization in cotton. **a** Allotetraploid cotton (*AADD*) was formed between A-genome species like *G. arboreum* (*Ga*) and D-genome species like *G. raimondii* (*Gr*), giving rise to five allotetraploid species: wild *G. hirsutum* (*wGh*), wild *G. barbadense* (*wGb*), *G. tomentosum* (*Gt*), *G. darwinii* (*Gd*), and *G. mustelinum* (*Gm*). Wild Gh and Gb are domesticated into cultivated *G. hirsutum* (*cGh*) and *G. barbadense* (*cGb*), respectively. **b** The number of differentially methylated cytosines in CG context (*DmCG*) in each pairwise comparison between different cotton species as shown in **a**. A2D5 is an interspecific hybrid between Ga (A2) and Gr (D5). *Blue, green, black*, and *yellow* brackets indicate comparisons of wild vs. wild, cultivated vs. wild allotetraploids, diploid vs. allotetraploid, and diploid parents vs. interspecific hybrid, respectively. **c** Phylogenetic tree was reconstructed based on genome-wide mCG divergent levels among cotton species. **d**, **e** Distribution of synonymous substitution values (*Ks*) (*left*) and gene-body DmCG percentages (*right*) of 6781 methylated orthologous genes (**d**) and 4063 unmethylated orthologous genes (**e**). *As*, *Ds* A subgenome and D subgenome in cultivated *G. hirsutum*, A *G. arboreum*, D *G. raimondii*. Peak values are indicated by *arrows*. The rate of methylation changes in each gene pair was estimated as the number of DmCG divided by the total number of CG in the gene body

changes (peaks at 0.18–0.24) was substantially higher than the substitution rate of coding sequence (*Ks* value peaks at 0.007–0.034) (Fig. 1d), suggesting that the methylation change rate is faster than the neutral sequence substitution rate. In the CG body-unmethylated genes, although the sequence variation remained at a similar level, the methylation peak disappeared (Fig. 1e).

DNA methylation divergence between progenitor-like diploid species

TEs are often associated with DNA methylation and genome complexity [14, 38, 39]. In diploid species, the *G. arboreum* genome is twofold larger and contains more TEs than the *G. raimondii* genome, probably because

of TE expansion in the centromeric and peri-centromeric regions [40, 41] (Additional file 2: Figure S3a). However, in genic regions there were more DNA TEs and especially retrotransposons in *G. raimondii* than in *G. arboreum* (Fig. 2a). For retrotransposons, *G. raimondii* had more *Copia* and unclassified long terminal repeats (LTRs) in flanking sequences of the gene body than *G. arboreum* (Fig. 2b). Terminal repeat retrotransposons in miniature (TRIMs), which are enriched near genes [42, 43], showed similar distribution patterns between *G. raimondii* and *G. arboreum* (Fig. 2b). Because of high CG methylation levels in the TEs, *G. raimondii* homoeologs were generally more methylated than *G. arboreum* homoeologs (Fig. 2c; Additional file 2: Figure S3b).

Song et al. Genome Biology (2017) 18:99

Page 4 of 14

Fig. 2 Asymmetrical distribution of TE and DNA methylation in A and D genomes. **a** Distribution of class I and II TEs in genic regions of *G. arboreum* (Ga, *blue* for I and *gray* for II) and *G. raimondii* (Gr, *orange* for I and *yellow* for II). **b** Density differences (between *G. raimondii* and *G. arboreum*) of *Gypsy* (*orange*), *Copia* (*blue*), *TRIM* (terminal repeat retrotransposons in miniature, *green*), and other long terminal repeats (*LTR*, *gray*) in the genic regions. **c**, **d** CG (**c**) and CHG (**d**) methylation levels in the genic regions of *G. arboreum* and *G. raimondii*. **e** Average CG methylation differences between intergenic and intragenic TEs. **f** Average CHG methylation differences between intergenic and intragenic TEs
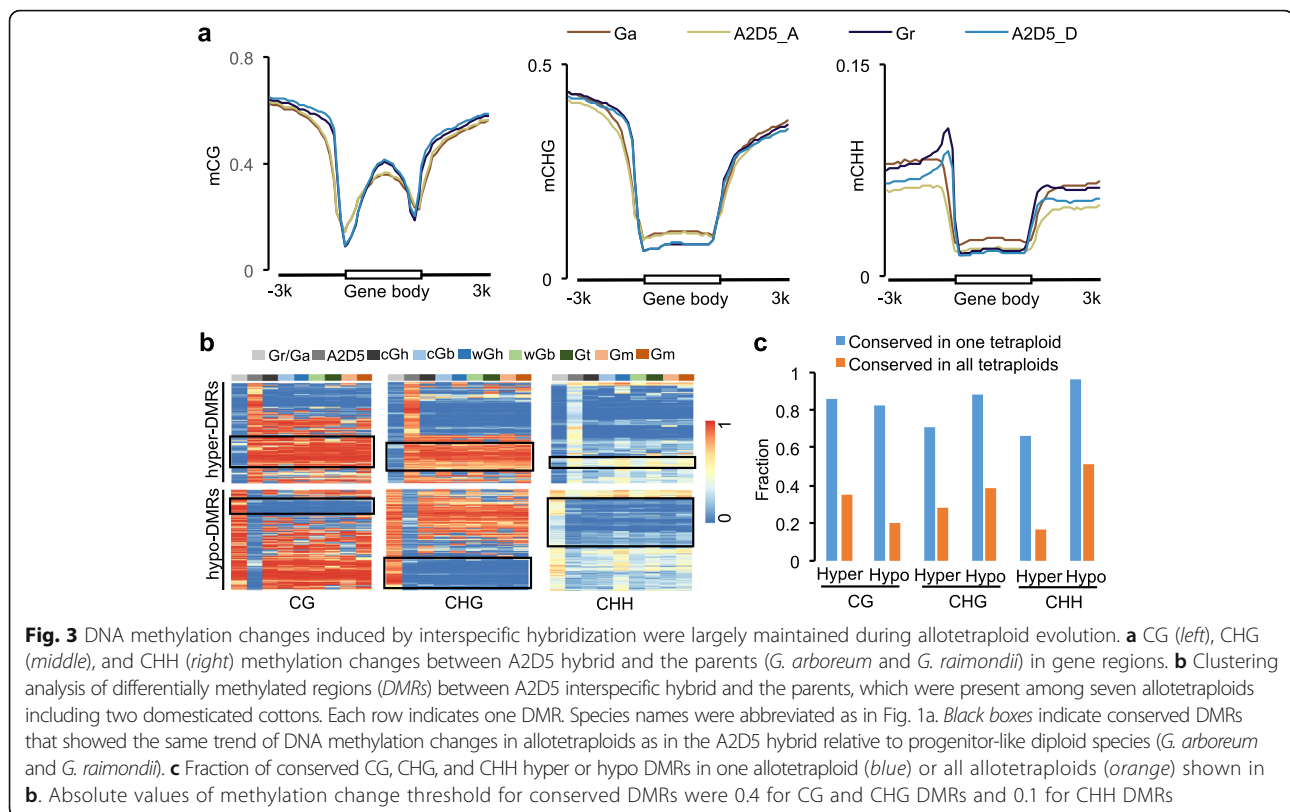
Although TEs were also associated with high CHG methylation levels (Additional file 2: Figure S3b), CHG methylation levels in the gene body were similar between *G. raimondii* and *G. arboreum* (Fig. 2d). CHG methylation correlates positively with the repressive histone mark H3K9 methylation and negatively with gene expression [19, 39]. We speculate that TEs inserted in the gene body (intragenic TEs) could gradually lose CHG methylation during evolution to prevent silencing. Consistent with the hypothesis, CHG methylation levels were lower in the intragenic TEs than in the intergenic TEs; reduction in CHG methylation (27–49%) is higher than decrease in CG methylation (~12%) (Fig. 2e, f). CHG methylation levels of intragenic TEs were decreased to lower levels in *G. raimondii* than in *G. arboreum* (Fig. 2e, f). As a result, CHG methylation levels in the gene body were similar between them, although *G. raimondii* has more TEs in the gene body than *G. arboreum* (Fig. 2b, d).

Hybrid-induced DNA methylation changes are conserved in polyploids

The DNA methylation difference between the species could induce changes in allopolyploids [26, 27, 44].

Surprisingly, compared with the parents, CG and CHG methylation levels in the gene body were similar, but CHH methylation levels were decreased in the interspecific hybrids (A2D5) that were formed between *G. raimondii* and *G. arboreum* (Fig. 3a). This is different from *Arabidopsis* intraspecific hybrids that display higher methylation levels in all contexts than the parents [45, 46]. We further examined whether hybridization-induced methylation changes could be maintained during evolution, compared to polyploidization of allotetraploid cottons that had diverged into five or more different species [34]. To test this, we identified 44,386 CG, 17,269 CHG, and 7461 CHH differentially methylated regions (DMRs) between the interspecific hybrid (A2D5) and parents (Additional file 3: Table S2). Among them, proportions of conserved CG, CHG, and CHH DMRs were 66–96% and 17–51% in at least one and all allotetraploid species tested, respectively, which were significantly higher than random events ($P < 1e^{-200}$, hypergeometric test) (Fig. 3b, c). Although the exact progenitors of ancient allotetraploids are unknown, the data suggest that a wide range of hybridization-induced DNA methylation changes is conserved during polyploid evolution.
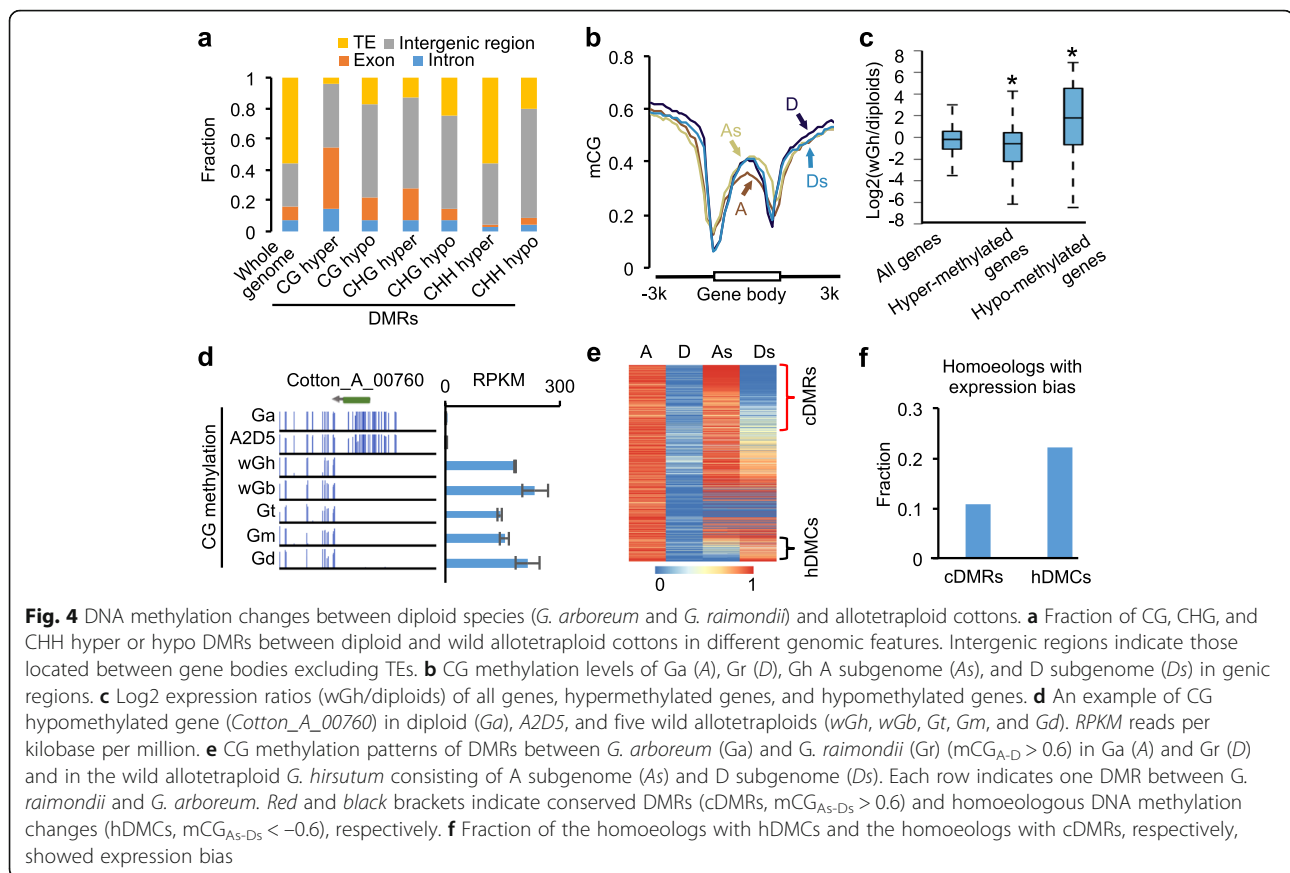
Song *et al. Genome Biology* (2017) 18:99

Page 5 of 14

Fig. 3 DNA methylation changes induced by interspecific hybridization were largely maintained during allotetraploid evolution. **a** CG (*left*), CHG (*middle*), and CHH (*right*) methylation changes between A2D5 hybrid and the parents (*G. arboreum* and *G. raimondii*) in gene regions. **b** Clustering analysis of differentially methylated regions (*DMRs*) between A2D5 interspecific hybrid and the parents, which were present among seven allotetraploids including two domesticated cottons. Each row indicates one DMR. Species names were abbreviated as in Fig. 1a. *Black boxes* indicate conserved DMRs that showed the same trend of DNA methylation changes in allotetraploids as in the A2D5 hybrid relative to progenitor-like diploid species (*G. arboreum* and *G. raimondii*). **c** Fraction of conserved CG, CHG, and CHH hyper or hypo DMRs in one allotetraploid (*blue*) or all allotetraploids (*orange*) shown in **b**. Absolute values of methylation change threshold for conserved DMRs were 0.4 for CG and CHG DMRs and 0.1 for CHH DMRs

Role of DNA methylation in homoeologous expression bias

To test the effects of DNA methylation on polyploid diversification, we identified conserved sequences (2 kb or longer) between diploid and allotetraploid species and analyzed DMRs between them (see "Methods"). We found 100,246 CG, 109,424 CHG, and 252,042 CHH DMRs between allotetraploid and diploid species, ~30% of which were common to all five allotetraploid cottons (Additional file 2: Figure S4; Additional file 4: Table S3). The common DMRs in allotetraploid cottons except for CHH-hyper DMRs were enriched in genic and intergenic regions, but largely excluded from TEs (Fig. 4a).

For genes, CG-hyper DMRs were significantly enriched in the gene body, while CG-hypo DMRs were enriched in intergenic regions (Fig. 4a). In genic regions, A and D homoeologs of allotetraploid cottons showed higher CG methylation levels in the gene body ($P < 1e^{-100}$, Wilcoxon signed-rank test) but lower CG methylation levels in the promoters ($P < 1e^{-100}$, Wilcoxon signed-rank test) than diploid species (Fig. 4b). Among diploid species, CG methylation levels of the gene body were lower in A-genome *G. arboreum* than in D-genome *G. raimondii*. However, in allotetraploid cotton, CG methylation levels were higher in the A than in the D homoeologs (Fig. 4b), indicating that more CG methylation occurred in the A

homoeologs of allotetraploid cotton. This is consistent with higher H3K4me3 levels in the D than in the A homoeologs in allotetraploid cotton [47], as intragenic DNA methylation antagonizes H3K4 trimethylation in plants [48, 49] and mammalian cells [50]. Using an absolute value of at least 0.6 in CG methylation changes between wild *G. hirsutum* and diploid species, we identified 501 hypermethylated and 137 hypomethylated genes in wild *G. hirsutum* compared with diploid species (Additional file 5: Table S4). CG methylation changes in the gene body significantly induce gene expression changes (Fig. 4c, d; Additional file 2: Figure S5), indicating a role for CG methylation in altering expression of these homoeologs in allotetraploids.

Polyploid formation often leads to intergenomic interactions including homoeologous exchanges [51], which are predicted to occur between methylated and unmethylated homoeologous loci. Among 5850 DMRs identified between the diploid species (A vs. D), 1165 DMRs (~20%) remained unchanged between A- and D-subgenome homoeologs (As vs. Ds) in the wild *G. hirsutum*, which designated conserved DMRs (cDMRs) (Fig. 4e). However, 339 DMRs (~6.8%) had opposite patterns between A and D homoeologs in the wild *G. hirsutum*, which designated homoeologous DNA methylation changes (hDMCs) (Fig. 4e). Interestingly, more homoeologs with hDMCs

Song et al. Genome Biology (2017) 18:99

Page 6 of 14

Fig. 4 DNA methylation changes between diploid species (*G. arboreum* and *G. raimondii*) and allotetraploid cottons. **a** Fraction of CG, CHG, and CHH hyper or hypo DMRs between diploid and wild allotetraploid cottons in different genomic features. Intergenic regions indicate those located between gene bodies excluding TEs. **b** CG methylation levels of Ga (*A*), Gr (*D*), Gh A subgenome (*As*), and D subgenome (*Ds*) in genic regions. **c** Log2 expression ratios (wGh/diploids) of all genes, hypermethylated genes, and hypomethylated genes. **d** An example of CG hypomethylated gene (*Cotton_A_00760*) in diploid (*Ga*), A2D5, and five wild allotetraploids (*wGh*, *wGb*, *Gt*, *Gm*, and *Gd*). RPKM reads per kilobase per million. **e** CG methylation patterns of DMRs between *G. arboreum* (Ga) and *G. raimondii* (Gr) (mCG$_{A-D}$ > 0.6) in Ga (*A*) and Gr (*D*) and in the wild allotetraploid *G. hirsutum* consisting of A subgenome (*As*) and D subgenome (*Ds*). Each row indicates one DMR between *G. raimondii* and *G. arboreum*. *Red* and *black* brackets indicate conserved DMRs (cDMRs, mCG$_{As-Ds}$ > 0.6) and homoeologous DNA methylation changes (hDMCs, mCG$_{As-Ds}$ < −0.6), respectively. **f** Fraction of the homoeologs with hDMCs and the homoeologs with cDMRs, respectively, showed expression bias

than the homoeologs with cDMRs displayed the expression bias (A/D ≠ As/Ds homoeologs) (Fig. 4f; Additional file 6: Table S5), indicating a role for hDMCs in homoeolog expression bias in the allotetraploids. We randomly selected three homoeolog pairs with hDMCs and expression bias, which were validated by bisulfite conversion and Sanger sequencing (Additional file 2: Figure S6), suggesting the reliability of these hDMCs.

During the evolution of allotetraploid cottons, some homoeolog pairs (A and D) lost one homoeolog [33], but the gene was present in both progenitor-like diploid species (*G. raimondii* and *G. arboreum*). To infer the role of DNA methylation in gene loss, we identified all homoeolog pairs with one copy lost in allotetraploid cottons but the corresponding orthologs present in both diploid species, and compared DNA methylation levels of these genes. Consistent with the previous finding of correlating DNA methylation with gene loss in paleopolyploid beans [30], the genes lost in the allotetraploids but present in diploid species showed significantly higher non-CG methylation levels in the gene body than their corresponding homoeologs that were retained in the allotetraploids ($P < 1e^{-20}$, Wilcoxon signed-rank test) (Additional file 2: Figure S7). This suggests that non-CG

methylation in the gene body could cause gene silencing and eventually gene loss.

DNA methylation contributes to cotton domestication

Cultivated *G. hirsutum* and *G. barbadense* possessed 7850 and 6148 CG DMRs, respectively, relative to their wild relatives (Additional file 7: Table S6). However, most DMRs (>85%) were specific to *G. hirsutum* or *G. barbadense*, and only ~14% were common to both species (Fig. 5a), consistent with the notion that *G. hirsutum* and *G. barbadense* were independently domesticated [35]. The shared DMRs were associated with 519 genes, which could generate putative epialleles (Additional file 8: Table S7). Gene Ontology (GO) analysis showed that these genes were enriched in several important biological processes, including metabolic processes, stress response, and regulation of seed dormancy (Fig. 5b). For example, cotton homologs of the *Arabidopsis* genes encoding TARGET OF RAPAMYCIN (TOR) and its binding partner RAPTOR were both demethylated in cultivated cottons but not in wild cottons (Additional file 8: Table S7). TOR, encoding a kinase, is a master regulator to adjust cell growth and development in response to nutritional signaling in plants and animals [52]. In *Arabidopsis*, root and shoot growth,

Song *et al. Genome Biology* (2017) 18:99

Page 7 of 14
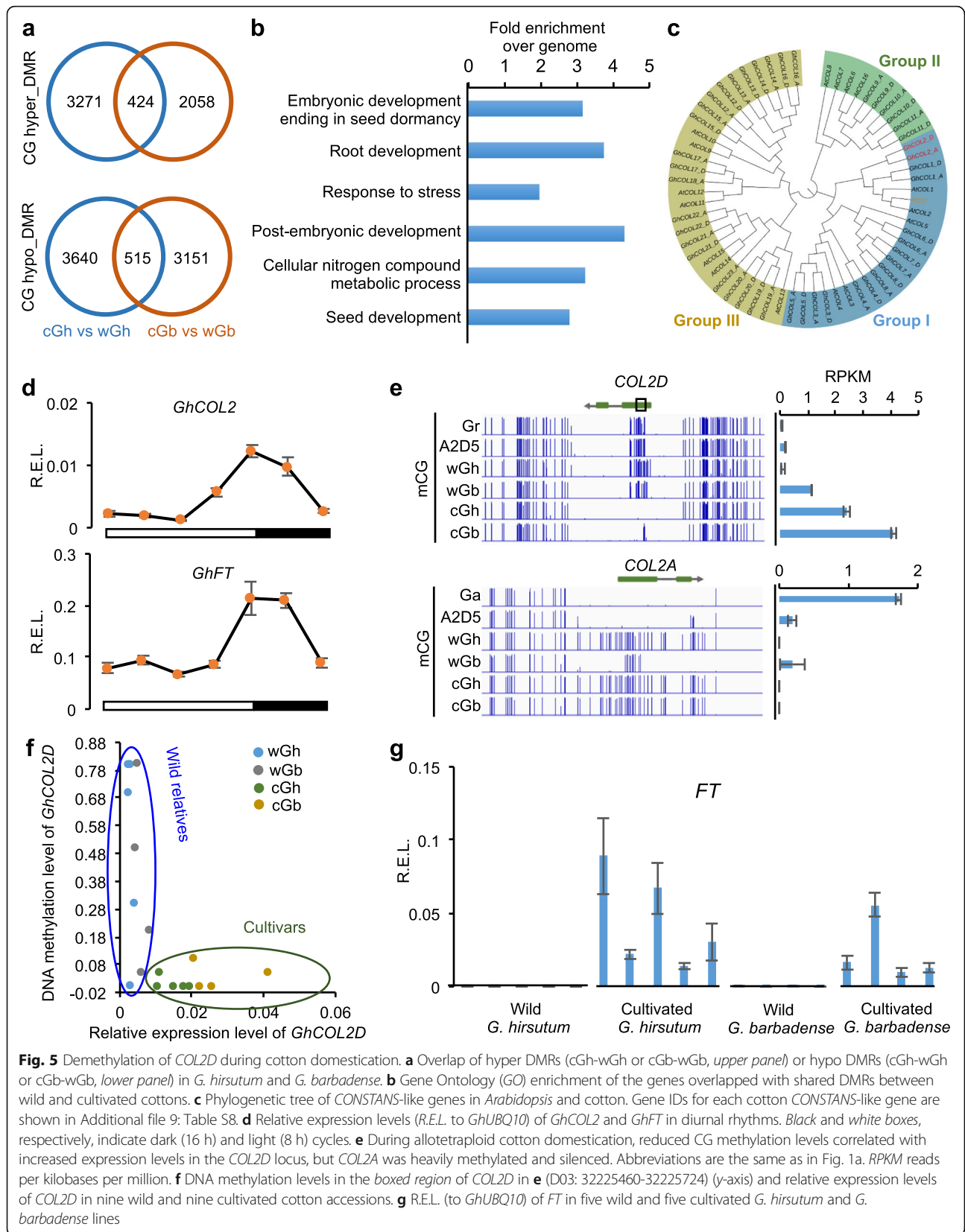
Fig. 5 Demethylation of *COL2D* during cotton domestication. **a** Overlap of hyper DMRs (cGh-wGh or cGb-wGb, *upper panel*) or hypo DMRs (cGh-wGh or cGb-wGb, *lower panel*) in *G. hirsutum* and *G. barbadense*. **b** Gene Ontology (*GO*) enrichment of the genes overlapped with shared DMRs between wild and cultivated cottons. **c** Phylogenetic tree of *CONSTANS*-like genes in *Arabidopsis* and cotton. Gene IDs for each cotton *CONSTANS*-like gene are shown in Additional file 9: Table S8. **d** Relative expression levels (*R.E.L.* to *GhUBQ10*) of *GhCOL2* and *GhFT* in diurnal rhythms. *Black* and *white boxes*, respectively, indicate dark (16 h) and light (8 h) cycles. **e** During allotetraploid cotton domestication, reduced CG methylation levels correlated with increased expression levels in the *COL2D* locus, but *COL2A* was heavily methylated and silenced. Abbreviations are the same as in Fig. 1a. *RPKM* reads per kilobases per million. **f** DNA methylation levels in the *boxed region* of *COL2D* in **e** (D03: 32225460-32225724) (*y*-axis) and relative expression levels of *COL2D* in nine wild and nine cultivated cotton accessions. **g** R.E.L. (to *GhUBQ10*) of *FT* in five wild and five cultivated *G. hirsutum* and *G. barbadense* lines

Song *et al. Genome Biology* (2017) 18:99

Page 8 of 14

seed yield, and stress tolerance are positively correlated with *TOR* expression levels [53, 54]. Disruption of *TOR* or *RAPTOR* leads to abnormal vegetative growth and seed abortion. In contrast, plants with moderate increase of *TOR* expression show enhanced root and shoot growth, greater resistance to osmotic stress, and higher seed production. These epialleles identified in tetraploid cottons could contribute to morphological and physiological changes, including fiber length, reduction in seed dormancy, and photoperiod sensitivity [28], during domestication of *G. hirsutum* (Upland) and *G. barbadense* (Pima) cotton.

Among the putative epialleles, we selected a cotton *CONSTANS-LIKE2D* homoeolog (*COL2D*) and an *Arabidopsis CONSTANS* (*CO*) homolog [55] for functional analysis. In *Arabidopsis*, *CO* and *CO-LIKE* (*COL*) genes control photoperiodic flowering through induction of *FLOWERING LOCUS T* (*FT*), and *CO* regulates *FT* expression through diurnal rhythms [55–57]. Phylogenetically, eight *G. hirsutum COLs* (*GhCOLs*) (1–8, group I *COL*) were in the same clade with *Arabidopsis CO* (Fig. 5c; Additional file 9: Table S8). Among *GhCOLs*, only *GhCOL2* exhibited similar expression rhythms with *GhFT*, indicating that *GhCOL2* is a major regulator of *GhFT* (Fig. 5d; Additional file 2: Figure S8).

Loss of photoperiod sensitivity is a major "domestication syndrome" trait [35] of Upland or American cotton (*G. hirsutum* L.) and Pima or Egyptian cotton (*G. barbadense* L.) that accounts for >95% and ~5%, respectively, of these annual cotton crops worldwide [31]. We predicted that DMRs of *COL2D* could lead to the loss of photoperiod sensitivity in domesticated cottons and promote global cotton production. Indeed, lower CG methylation levels of *COL2D* were associated with its higher expression levels in cultivated and photoperiod-insensitive *G. hirsutum* and *G. barbadense* than in their wild relatives (photoperiod sensitive) (Fig. 5e). To confirm the relationship between loss of DNA methylation and expression of *COL2D*, we treated the wild *G. hirsutum* (TX2095) seedlings using 5-aza-2′-deoxycytidine (5-aza-dC), a chemical inhibitor for DNA methylation [58]. Consistent with the hypothesis that demethylation of *COL2D* leads to increased levels of expression, the 5-aza-dC treatment reduced DNA methylation and increased expression levels of *COL2D* (Additional file 2: Figure S9). Although we cannot rule out possible toxic and side effects of 5-aza-dC, both genome-wide data and 5-aza-dC treatment results indicate that DNA methylation loss in the *COL2D* could promote its expression in domesticated cottons. Notably, *COL2D* was heavily methylated and silenced in *G. raimondii* that is photoperiod sensitive, while *COL2A* was hypomethylated and highly expressed in cultivated *G. arboreum* (Fig. 5e), which is photoperiod insensitive [59]. Interestingly, the

COL2A homoeolog was hypermethylated and repressed in cultivated Upland and Pima cottons, while the *COL2D* homoeolog was highly expressed (Fig. 5e), suggesting that *COL2A* in the allotetraploid cottons was either silenced after polyploid formation or originated from a progenitor that was different from the extant *G. arboreum* species. The high-level expression of *COL2D* is likely associated with positive selection of *COL2D* but not with *COL2A* during domestication of Upland and Pima cottons [60].
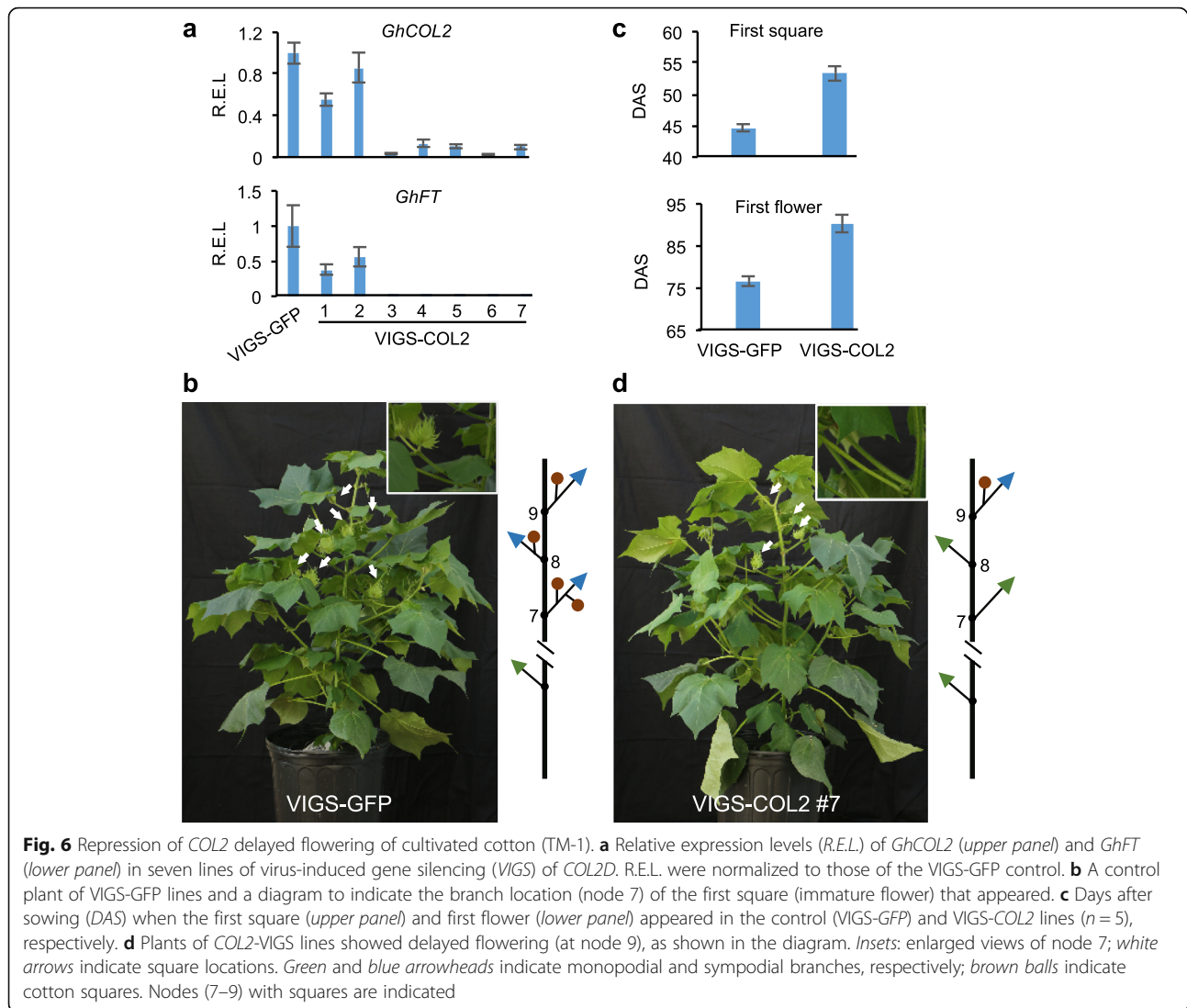
Consistent with the role of this epiallele in allotetraploid cotton domestication, locus-specific DNA methylation loss and expression increase of *COL2D* were associated with nine cultivated accessions but not with nine wild accessions of *G. hirsutum* and *G. barbadense* which were randomly selected for this test (Fig. 5f). As a result, cotton *FT* was expressed at higher levels in the cultivated accessions than in their wild relatives (Fig. 5g).

If *COL2* regulates photoperiodic flowering in cotton, reducing *COL2* expression in Upland cotton (accession TM-1) would delay flowering. This should mimic the methylation effect on a specific gene, which cannot be easily tested. Using virus-induced gene silencing (VIGS) [61], *COL2* was specifically down-regulated (Fig. 6a), while expression of other *COLs* remained unchanged relative to the control plants that expressed VIGS of GFP (VIGS-GFP) (Additional file 2: Figure S10). Expression of the downstream gene *GhFT* was also inhibited in seven independently derived VIGS lines of *COL2* (VIGS-COL2 #1-7), mimicking the repression pattern of *COL2D* (Fig. 6a). In addition, knocking down expression of other *COLs* such as *GhCOL3* and *GhCOL6* did not affect *GhFT* expression (Additional file 2: Figure S11), further validating this specific role of *COL2* in regulating *GhFT* expression in cotton. Down-regulating *COL2* and *GhFT* expression has delayed flowering time in five VIGS-GhCOL2 lines (#3–7) analyzed (Fig. 6b, c). In the control plants, the first square appeared at ~44 days after sowing (DAS) on node 7 of the main stem in the long-day condition (Fig. 6b, c), while the first squares emerged by ~53 DAS on node 9 in the VIGS-COL2 lines. Consequently, VIGS-COL2 also flowered later than the control plants (Fig. 6d). As wild cottons normally do not flower in the long-day condition [60], delayed flowering by knocking down *COL2* expression indicates that other genes could also mediate photoperiodic flowering in cotton. This notion is supported by several quantitative trait loci (QTLs) that are associated with the loss of photoperiod sensitivity during cotton domestication [62, 63].

Discussion and conclusions

DNA methylation affects many biological processes including disease-associated syndromes in humans [64]. Natural variation of epialleles has emerged to play a role in plant evolution [65], morphological diversity in plants [11], and selection and breeding of agronomic traits in

Fig. 6 Repression of *COL2* delayed flowering of cultivated cotton (TM-1). **a** Relative expression levels (*R.E.L.*) of *GhCOL2* (*upper panel*) and *GhFT* (*lower panel*) in seven lines of virus-induced gene silencing (*VIGS*) of *COL2D*. R.E.L. were normalized to those of the VIGS-GFP control. **b** A control plant of VIGS-GFP lines and a diagram to indicate the branch location (node 7) of the first square (immature flower) that appeared. **c** Days after sowing (*DAS*) when the first square (*upper panel*) and first flower (*lower panel*) appeared in the control (VIGS-*GFP*) and VIGS-*COL2* lines (*n* = 5), respectively. **d** Plants of *COL2*-VIGS lines showed delayed flowering (at node 9), as shown in the diagram. *Insets*: enlarged views of node 7; *white arrows* indicate square locations. *Green* and *blue arrowheads* indicate monopodial and sympodial branches, respectively; *brown balls* indicate cotton squares. Nodes (7–9) with squares are indicated

crops [12, 13]. The evolutionary history of cotton species consists of four intervened phases: genome variation among diploid species including A- and D-genome progenitors, interspecific hybridization and allotetraploid formation, natural selection and adaptation, and artificial selection and domestication (Fig. 1a). In this study, we compared DNA methylation changes between Gr and Ga, A2D5 hybrid and Gr/Ga, wild tetraploid cottons (wGh, wGb, Gt, Gd, and Gm) and diploid cottons (Gr and Ga), and cultivated cottons (Gh and Gb) and wild tetraploid cottons (wGh and wGb) and examined the role of DNA methylation in four evolutionary phases. We found that the rate of DNA methylation changes exceeds that of neutral sequence substitutions during cotton evolution and domestication (Fig. 1d, e). The methylated genes remain methylated with a high rate of DNA methylation changes in the coding sequences, whereas unmethylated genes stay unmethylated with

little or no change in DNA methylation. This probably occurs because, for unmethylated genes, DNA methylation is dispensable for expression regulation, and *cis*-acting elements and/or *trans*-acting factors may regulate expression of these genes. For methylated genes, DNA methylation could change and alter gene expression in response to internal (genetic perturbation) and external (environmental variation) signals during evolution and adaptation. This suggests that the DNA methylation status of genes in ancestral progenitor species could be selected and maintained during evolution.

The content and organization of repeats and transposons in the genome are correlated with DNA methylation patterns, which could affect transcription activity of neighboring genes [66, 67]. Among diploid species, *G. raimondii* has more TEs and higher DNA methylation levels in genic regions than *G. arboreum*, although *G. arboreum* has more TEs in the whole genome [33, 41].

Song *et al. Genome Biology* (2017) 18:99

Page 10 of 14

The asymmetrical expansion of TEs and DNA methylation is probably a rule for the evolution of ancestral cotton genomes. Interestingly, although TEs are associated with higher DNA methylation levels than other regions, DNA methylation changes between diploid and allotetraploid cottons were largely excluded from TEs in conserved sequences between diploid and allotetraploid cottons (Fig. 4a), indicating that DNA methylation in TEs is relatively stable during evolution. To exclude a potential effect of sequence variation on DNA methylation analysis, we used the conserved sequences to examine DNA methylation changes between diploid and allotetraploid cottons. DNA methylation changes in non-conserved sequences should be examined in the future by longer reads, such as single-molecule real-time (SMRT) sequencing [68].

Hybridization and polyploidy play key roles in plant speciation and diversification [4, 8]. Hybridization induces genome-wide DNA methylation changes [27, 44]; however, it is unclear whether these changes could be maintained in the long-term evolution of allopolyploids. Remarkably, we observed that the majority of DNA methylation changes induced in the interspecific hybrids are maintained in one or more allotetraploid cottons, some of which are shared between wild and domesticated cottons, suggesting that hybridization-induced DNA methylation changes are transmittable during polyploid speciation. Furthermore, we uncovered homoeologous DNA methylation changes, which correlate with homoeolog-expression bias, providing an epigenetic basis for gene expression and evolutionary novelty in the polyploids [69]. This is reminiscent of homoeologous sequence or chromosomal exchanges between subgenomes that occur in some polyploids like *Brassica* [51] or *Tragopogan* [70]. Homoeologous DNA methylation changes (hDMCs) in allotetraploid cottons could occur from genomic exchanges, although this type of intergenomic change is rather rare in polyploid cotton [71] or *Arabidopsis* [72, 73]. Alternatively, small interfering RNAs (siRNAs) produced from either or both homoeologs could target a different homoeolog to induce de novo methylation via RdDM [74, 75] which is maintained during evolution.

Consistent with the stable maintenance of DNA methylation, many candidate epialleles could contribute to morphological and physiological changes during cotton evolution and domestication, providing a valuable resource for epigenetic engineering and breeding better-yielding crops. This example of the domestication-induced epiallele *COL2D* that contributes to the loss of photoperiod sensitivity by demethylation in cultivated cottons provides novel insights into the role of DNA methylation in domestication of cotton and many other polyploid crops. Functional analysis of these polyploidy-induced epialleles should help develop biotechnological tools for manipulating DNA methylation and epialleles through breeding, selection, and ultimate improvement of plants and animals.

Methods

Plant materials

G. raimondii (Gr, D5-3), *G. arboreum* (Ga, A2), interspecific hybrid (A2D5) between *G. arboreum* and *G. raimondii*, wild *G. hirsutum* (wGh, TX2095), wild *G. barbadense* (wGb, Gb-706), *G. tomentosum* (Gt, AD3-30), *G. mustelinum* (Gm, AD4-11), *G. darwinii* (Gd, AD5-31), cultivated *G. hirsutum* (cGh, TM-1), and cultivated *G. barbadense* (cGd, Pima-S6) were grown in the greenhouses at College Station and Austin, Texas. Leaves of each genotype were harvested with three biological replications for MethylC-seq and RNA-seq libraries. DNA methylation and expression levels of GhCOL2_D were further analyzed in five wild *G. hirsutum* accessions (TX701, TX1039, TX2092, TX2095, and TX2096), five cultivated *G. hirsutum* (TM-1, SA-308, SA-508, SA-528, and SA-1475), four wild *G. barbadense* (Gb-472, Gb-470, Gb-617, Gb-716), and four cultivated *G. barbadense* (Pima S2, Pima S6, Phytogen 800, 3-79), which were grown in the greenhouse under the light/dark (L/D) cycle of 16 h/L at 24 °C and 8 h/D at 20 °C. To exclude effects of development stage and circadian rhythm on gene expression change, the first true leaves at 16 days after sowing were harvested at ZT15 (zeitgeber time, ZT0 = dawn, 6 am) for DNA and RNA extraction.

mRNA-seq library construction

After DNase treatment, total RNA (~1 μg) was subjected to construct strand-specific mRNA-seq libraries with two biological replications using NEBNext® Ultra™ Directional RNA Library Prep Kit (New England Bioloabs (NEB), Ipswich, MA, USA) according to the manufacturer's instructions. For each genotype, mRNA-seq libraries were constructed with two biological replicates and were paired-end sequenced for 126 cycles.

MethylC-seq library construction

Total genomic DNA (~5 μg) was fragmented to 100–1000 bp using Bioruptor (Diagenode, Denville, NJ, USA). End repair (NEBNext® End Repair Module) was performed on the DNA fragments followed by adding an "A" base to the 3′ end (NEBNext® dA-Tailing Module), and the resulting DNA fragments were ligated to the methylated DNA adapter (NEXTflex™ DNA Barcodes, Bioo Scientific, Austin, TX, USA). The adapter-ligated DNA of 200–400 bp was purified using AMPure beads (Beckman Coulter, Brea, CA, USA), followed by sodium bisulfite conversion using the MethylCode™ Bisulfite Conversion Kit (Life Technologies, Foster City, CA, USA). The bisulfite-converted DNA was amplified by 12 cycles

Song *et al. Genome Biology* (2017) 18:99

Page 11 of 14

of PCR using LongAmp® Taq DNA Polymerase (NEB) and subject to purification using AMPure beads (Beckman Coulter). For each genotype, MethylC-seq libraries were constructed with two biological replicates and paired-end sequenced for 126 cycles.

qRT-PCR

After DNase treatment, total RNA (2 μg) was used to produce cDNA with the Omniscript RT Kit (Qiagen, Valencia, CA, USA). The cDNA was used as the template for qRT-PCR using FastStart Universal SYBR Green Master (Roche, Indianapolis, IN, USA). The reaction was run on the Light-Cycler® 96 System (Roche, Pleasanton, CA, USA). The relative expression level of a gene was quantified using the expression value of cotton *GhUBQ10* as an internal control using the primers (Additional file 10: Table S9).

Analysis of mRNA-seq data

mRNA-seq reads of *G. raimondii* and *G. arboreum* were mapped to genome sequences of *G. raimondii* and *G. arboreum,* respectively [40, 41]. mRNA-seq reads of other species were mapped to combined genome sequences of *G. raimondii* and *G. arboreum*. TopHat software with options (–library-type fr-firststrand –b2-score-min L, 0, -0.6) was used for read mapping [76]. Uniquely mapped reads were extracted and analyzed by Cufflinks to determine transcript values [77]. The differentially expressed genes (DEGs) were identified using both the fold-change (>twofold) and analysis of variance (ANOVA) tests ($P < 0.01$) with two biological replicates.

Mapping of MethylC-seq reads with two biological replicates

MethylC-seq reads of *G. raimondii* and *G. arboreum* were mapped to genome sequences of *G. raimondii* and *G. arboreum,* respectively [40, 41]. MethylC-seq reads of the A2D5 hybrid were mapped to the combined genome sequences of *G. raimondii* and *G. arboreum*. MethylC-seq reads of all allopolyploid cottons were mapped to the genome sequence of *G. hirsutum* (TM-1) [33]. Bismark with parameters (–score_min L,0,-0.2 -X 1000 –no-mixed –no-discordant) was used for read mapping [78]. Only reads mapped to the unique sites were retained and used for further analysis. The reads mapped to the same site were collapsed into a single consensus read to reduce clonal bias.

Conserved cytosines between diploid and allotetraploid cottons

To exclude the bias, we selected the conserved regions (2 kb or longer) by aligning *G. hirsutum* (tetraploid) genome sequences with *G. arboreum* and *G. raimondii* genome sequences (Additional file 2: Figure S1). Specifically, genome sequences of *G. raimondii* and *G. arboreum*

were aligned to the genome sequence of *G. hirsutum* (TM-1) using LAST (version 545) software (*lastal –q3 – e35 –m50*) [79]. The unique best alignments with score >500 were extracted (*last-split –m1 –s200*). The 1-to-1 alignments were generated by swapping the sequences (*maf-swap*) and subjected to extracting best alignments again (*last-split –m1 –s200*) until the alignments reached a score over 2000, and the conserved cytosines between diploid and allopolyploid cottons were extracted using Python scripts. Thus, these conserved regions are present in both diploid and tetraploid cottons. DMRs identified between allotetraploids and diploid species were present in all allotetraploid and diploid cottons.

Conserved cytosines among allotetraploid cottons

To avoid the base bias among different species, only the conserved cytosines were used for the analysis. Although most cytosines were changed to uracil in bisulfite conversion, guanine in MethylC-seq reads could be used to confirm cytosine in the complementary strand (Additional file 2: Figure S1c). After mapping MethylC-seq reads of wild *G. hirsutum*, *G. barbadense*, *G. tomentosum*, *G. mustelinum*, and *G. darwinii* to the reference genome of cultivated *G. hirsutum* (TM1) [33], guanines (G) in uniquely mapped MethylC-seq reads were used to confirm cytosines (C) in the complementary strand. All conserved cytosines were called when the same bases were covered by at least three reads and used for further analysis.

Phylogenetic tree

Methylation levels of the cytosines conserved in all species were obtained in each species. We calculated the distance between every two species using Euclidean distances. For two species P and Q, $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ in n CG sites, p_n is the methylation level in the nth CG site in the P species; q_n is the methylation level in the nth CG site in the Q species. The general Euclidean distance between P and Q was calculated as $((p_1-q_1)^2 + (p_2-q_2)^2 + ... + (p_n-q_n)^2)^{0.5}$. The Euclidean distances were used to construct the distance matrix for all species. Phylogenetic trees were constructed based on the distance matrix using the 'ape' R package with Neighbor-Joining algorithm and 1000 bootstraps [80].

Differentially methylated cytosines (DmCs)

Only cytosines covered by at least three reads in both species in comparison were considered for testing. The methylation level of a cytosine site was calculated as $C/(C + T)$ [81]. C indicates the number of reads with cytosine for this site. T indicates the number of reads with thymine for this site. A cytosine was considered as a DmC between two species if it showed an

Song *et al. Genome Biology* (2017) 18:99

Page 12 of 14

absolute change in methylation level of at least 0.5 and $P < 0.01$ by a one-way ANOVA test between species using two biological replicates.

Differentially methylated regions (DMRs)

DMRs were identified using 100-bp sliding windows. The mean methylation level was calculated for each window [81]. For CG and CHG DMRs, windows containing at least four cytosines in the CG or CHG context covered by at least three reads were selected for identifying DMRs. For CHH DMRs, windows containing at least 16 cytosines in the CHH context covered by at least three reads were selected. DMRs between two species in comparison were determined using an ANOVA test with two biological replications ($P < 0.05$) and cut-off values of methylation levels (0.5 for CG and CHG DMRs; 0.2 for CHH DMRs).

Identification of orthologous genes

Orthologous genes were obtained from a published paper [33]. Orthologous genes with more than 40% cytosines in the gene body covered by at least three reads were selected to calculate Ks and DmCG ratios between methylated and unmethylated orthologous genes.

Treatment of cotton seedling using 5-aza-2′-deoxycytidine (5-aza-dC)

In a beaker, cotton seeds (TX2095) were placed on sterile gauze and soaked in water with or without 5-aza-dC (18 mg/L). The beaker was capped with plastic wrap and placed in a climate incubator under a light/dark (L/D) cycle of 16 h/L at 30 °C and 8 h/D at 24 °C. At 3 days after sowing (DAS), the old gauze was replaced with new gauze soaked in water with or without 5-aza-dC. From 6 DAS, we replaced the old gauze with new gauze soaked in 1/2 Murashige and Skoog medium with or without 5-aza-dC (18 mg/L) every 3 days. The first true leaves at 16 DAS were harvested at ZT15 with five biological replicates for DNA and RNA extraction.

Bisulfite sequencing of *COL2D* fragment

Approximately 500 ng of genomic DNA was used for bisulfite conversion using the MethylCode™ Bisulfite Conversion Kit (Life Technologies). Bisulfite-treated DNA was then amplified by ZymoTaq DNA polymerase (Zymo Research, Irvine, CA, USA) and primers (5′-TTATTTGTAGTGTTGATGTAGTATTATTTTG-3′ and 5′-TTTCCAAACTCAAACAATAACCAAAAATCCAT TTC-3′) targeting the *COL2D* fragment (D03: 32225460-32225724). The purified amplicons were cloned into a pGEM-T vector (Promega, Madison, WI, USA). For each plant genotype, an average of 15 clones was randomly chosen for sequencing in two biological replicates.

Virus-induced gene silencing (VIGS)

G. hirsutum acc. TM-1 grew in a greenhouse under the light/dark (L/D) cycle of 16 h/L at 24 °C and 8 h/D at 20 °C. Fragments of *GhCOL2*, *GhCOL3*, and *GhCOL6* cDNA were amplified by PCR using primers (Additional file 10: Table S9) and subsequently cloned into the pYL156 (pTRV-RNA2) vector as pYL156-GhCOL2, pYL156-GhCOL3, and pYL156-GhCOL6, respectively. The plasmid pYL156-GFP was used as a control [61]. The plasmids pTRV-RNA1, pYL156-GFP, pYL156-GhCOL2, pYL156-GhCOL3, and pYL156-GhCOL6 were transformed into *Agrobacterium tumefaciens* strain GV3101 by electroporation. The *Agrobacterium tumefaciens* strain GV3101 was incubated overnight in Luria-Bertani medium containing 10 mM MES and 20 uM acetosyringone and finally resuspended in infiltration buffer (10 mM MES, 0.2 mM acetosyringone, 10 mM $MgCl_2$) to a final concentration of OD600 = 1.0. Cell suspensions were incubated at room temperature for 3 h. Equal amounts of different bacterial suspensions (pTRV-RNA1 with pYL156-GFP, pYL156-GhCOL2, pYL156-GhCOL3, or pYL156-GhCOL6) were infiltrated into the fully expanded cotyledons of the 10-day-old cottons with a needleless syringe.

Additional files

Additional file 1: Table S1. Summary of MethylC-seq reads. (XLSX 10 kb)

Additional file 2: Supplemental Figures S1–S11. (PDF 438 kb)

Additional file 3: Table S2. DMRs between A2D5 hybrid and the parents (Gr/Ga). (XLSX 2969 kb)

Additional file 4: Table S3. DMRs between polyploid and diploid (Gr/Ga) cottons. (XLSX 19969 kb)

Additional file 5: Table S4. Differentially methylated genes between wGh and diploid cottons (Gr/Ga). (XLSX 47 kb)

Additional file 6: Table S5. Homoeologs with hDMCs showing expression bias. (XLSX 17 kb)

Additional file 7: Table S6. DMRs between cultivated and wild cottons. (XLSX 579 kb)

Additional file 8: Table S7. Genes overlapped with shared DMRs between wild and cultivated cotton. (XLSX 54 kb)

Additional file 9: Table S8. Gene ID for *GhCOLs*. (XLSX 9 kb)

Additional file 10: Table S9. Primers for qRT-PCR and VIGS. (XLSX 9 kb)

Availability of data and materials

Sequence data have been deposited in the NCBI Nucleotide and Sequence Read Archive (SRA) under [SRA:SRP071640].

Song *et al. Genome Biology* (2017) 18:99

Page 13 of 14

Authors' contributions

Competing interests

Ethics approval

Publisher's Note

Author details

[1]Department of Molecular Biosciences, Institute for Cellular and Molecular Biology, and Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX 78712, USA. [2]State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China. [3]Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 78743, USA.

References

1. Comai L. The advantages and disadvantages of being polyploid. Nat Rev Genet. 2005;6:836–46.
2. Otto SP. The evolutionary consequences of polyploidy. Cell. 2007;131:452–62.
3. Chen ZJ. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol. 2007;58:377–406.
4. Soltis DE, Visger CJ, Soltis PS. The polyploidy revolution then…and now: Stebbins revisited. Am J Bot. 2014;101:1057–78.
5. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016;17:37.
6. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473:97–100.
7. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10:725–32.
8. Stebbins GL. Chromosomal evolution in higher plants. London: Edward Arnold; 1971.
9. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR. Transgenerational epigenetic instability is a source of novel methylation variants. Science. 2011;334:369–73.
10. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016;17:194.
11. Cubas P, Vincent C, Coen E. An epigenetic mutation responsible for natural variation in floral symmetry. Nature. 1999;401:157–61.
12. Quadrana L, Almeida J, Asis R, Duffy T, Dominguez PG, Bermudez L, Conti G, da Silva JV C, Peralta IE, Colot V, et al. Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat Commun. 2014;5:3027.
13. Manning K, Tor M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet. 2006;38:948–52.
14. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11:204–20.
15. Kankel MW, Ramsey DE, Stokes TL, Flowers SK, Haag JR, Jeddeloh JA, Riddle NC, Verbsky ML, Richards EJ. Arabidopsis MET1 cytosine methyltransferase mutants. Genetics. 2003;163:1109–22.
16. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, Jacobsen SE. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science. 2001;292:2077–80.
17. Bartee L, Malagnac F, Bender J. Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. Genes Dev. 2001;15:1753–8.
18. Cao X, Jacobsen SE. Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. Curr Biol. 2002;12:1138–44.
19. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol. 2014;21:64–72.
20. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell. 2013;153:193–205.
21. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69:915–26.
22. Stokes TL, Kunkel BN, Richards EJ. Epigenetic variation in Arabidopsis disease resistance. Genes Dev. 2002;16:171–82.
23. Hu L, Li N, Xu C, Zhong S, Lin X, Yang J, Zhou T, Yuliang A, Wu Y, Chen YR, et al. Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. Proc Natl Acad Sci U S A. 2014;111:10642–7.
24. Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. Genome-wide demethylation of Arabidopsis endosperm. Science. 2009;324:1451–4.
25. Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, Ecker JR, Whelan J, Lister R. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. Elife. 2015;4. doi: 10.7554/eLife.09343.
26. Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L. Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic Arabidopsis allotetraploids. Plant Physiol. 2002;129:733–46.
27. Lee HS, Chen ZJ. Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. Proc Natl Acad Sci U S A. 2001;98:6753–8.
28. Wang J, Tian L, Madlung A, Lee HS, Chen M, Lee JJ, Watson B, Kagochi T, Comai L, Chen ZJ. Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. Genetics. 2004;167:1961–73.
29. Salmon A, Ainouche ML, Wendel JF. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol Ecol. 2005;14:1163–75.
30. Kim KD, El Baidouri M, Abernathy B, Iwata-Otsubo A, Chavarro C, Gonzales M, Libault M, Grimwood J, Jackson SA. A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. Plant Physiol. 2015;168:1433–47.
31. Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM, Rabinowicz PD, Town CD, et al. Toward sequencing cotton (Gossypium) genomes. Plant Physiol. 2007;145:1303–10.
32. Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. Adv Agron. 2003;78:139–86.
33. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33:531–7.
34. Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, Gore MA, Wendel JF. Re-evaluating the phylogeny of allopolyploid Gossypium L. Mol Phylogenet Evol. 2015;92:45–52.
35. Olsen KM, Wendel JF. Crop plants as models for understanding plant adaptation and diversification. Front Plant Sci. 2013;4:290.
36. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 2007;39:61–9.
37. Takuno S, Gaut BS. Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol Biol Evol. 2012;29:219–27.
38. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328:916–9.
39. Song Q, Guan X, Chen ZJ. Dynamic roles for small RNAs and DNA methylation during ovule and fiber development in allotetraploid cotton. PLoS Genet. 2015;11:e1005724.
40. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature. 2012;492:423–7.
41. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, et al. Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet. 2014;46:567–72.
42. Witte CP, Le QH, Bureau T, Kumar A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A. 2001;98:13778–83.

Song *et al. Genome Biology* (2017) 18:99

Page 14 of 14

43. Gao D, Li Y, Kim KD, Abernathy B, Jackson SA. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome Biol. 2016;17:7.

44. Song Q, Chen ZJ. Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol. 2015;24:101–9.

45. Greaves IK, Groszmann M, Ying H, Taylor JM, Peacock WJ, Dennis ES. Trans chromosomal methylation in *Arabidopsis* hybrids. Proc Natl Acad Sci U S A. 2012;109:3570–5.

46. Shen H, He H, Li J, Chen W, Wang X, Guo L, Peng Z, He G, Zhong S, Qi Y, et al. Genome-wide analysis of DNA methylation and gene expression changes in two *Arabidopsis* ecotypes and their reciprocal hybrids. Plant Cell. 2012;24:875–92.

47. Zheng D, Ye W, Song Q, Han F, Zhang T, Chen ZJ. Histone modifications define expression bias of homoeologous genomes in allotetraploid cotton. Plant Physiol. 2016;172:1760–71.

48. Greenberg MV, Deleris A, Hale CJ, Liu A, Feng S, Jacobsen SE. Interplay between active chromatin marks and RNA-directed DNA methylation in *Arabidopsis thaliana*. PLoS Genet. 2013;9:e1003946.

49. Li X, Wang X, He K, Ma Y, Su N, He H, Stolc V, Tongprasit W, Jin W, Jiang J, et al. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. Plant Cell. 2008;20:259–76.

50. Lorincz MC, Dickerson DR, Schmitt M, Groudine M. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol. 2004;11:1068–75.

51. Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. Plant Cell. 2007;19:3403–17.

52. Xiong Y, Sheen J. The role of target of rapamycin signaling networks in plant growth and metabolism. Plant Physiol. 2014;164:499–512.

53. Deprost D, Truong HN, Robaglia C, Meyer C. An Arabidopsis homolog of RAPTOR/KOG1 is essential for early embryo development. Biochem Biophys Res Commun. 2005;326:844–50.

54. Deprost D, Yao L, Sormani R, Moreau M, Leterreux G, Nicolai M, Bedu M, Robaglia C, Meyer C. The *Arabidopsis* TOR kinase links plant growth, yield, stress resistance and mRNA translation. EMBO Rep. 2007;8:864–70.

55. Putterill J, Robson F, Lee K, Simon R, Coupland G. The CONSTANS gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. Cell. 1995;80:847–57.

56. Samach A, Onouchi H, Gold SE, Ditta GS, Schwarz-Sommer Z, Yanofsky MF, Coupland G. Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. Science. 2000;288:1613–6.

57. Onouchi H, Igeno MI, Perilleux C, Graves K, Coupland G. Mutagenesis of plants overexpressing CONSTANS demonstrates novel interactions among *Arabidopsis* flowering-time genes. Plant Cell. 2000;12:885–900.

58. Haaf T, Werner P, Schmid M. 5-Azadeoxycytidine distinguishes between active and inactive X chromosome condensation. Cytogenet Cell Genet. 1993;63:160–8.

59. Smith CW, Cothren JT. Cotton: origin, history, technology, and production. New York: John Wiley & Sons, Inc.; 1999.

60. Zhang R, Ding J, Liu C, Cai C, Zhou B, Zhang T, Guo W. Molecular evolution and phylogenetic analysis of eight COL superfamily genes in group I related to photoperiodic regulation of flowering time in wild and domesticated cotton (*Gossypium*) species. PLoS One. 2015;10:e0118669.

61. Gao X, Wheeler T, Li Z, Kenerley CM, He P, Shan L. Silencing GhNDR1 and GhMKK2 compromises cotton resistance to Verticillium wilt. Plant J. 2011;66:293–305.

62. Waddle BM, Lewis CF, Richmond TR. The genetics of flowering response in cotton. III. Fruiting behavior of *Gossypium irsutum* race *latifolium* in a cross with a variety of cultivated American upland cotton. Genetics. 1961;46:427–37.

63. Guo YF, McCarty JC, Jenkins JN, Saha S. QTLs for node of first fruiting branch in a cross of an upland cotton, *Gossypium hirsutum* L., cultivar with primitive accession Texas 701. Euphytica. 2008;163:113–22.

64. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

65. Weigel D, Colot V. Epialleles in plant evolution. Genome Biol. 2012;13:249.

66. Martienssen R. Transposons, DNA methylation and gene control. Trends Genet. 1998;14:263–4.

67. Kato M, Miura A, Bender J, Jacobsen SE, Kakutani T. Role of CG and non-CG methylation in immobilization of transposons in *Arabidopsis*. Curr Biol. 2003;13:421–6.

68. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014;24:688–96.

69. Shi X, Zhang C, Ko DK, Chen ZJ. Genome-wide dosage-dependent and -independent regulation contributes to gene expression and evolutionary novelty in plant polyploids. Mol Biol Evol. 2015;32:2351–66.

70. Chester M, Gallagher JP, Symonds VV, da Silva AV C, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). Proc Natl Acad Sci U S A. 2012;109:1176–81.

71. Liu B, Brubaker G, Cronn RC, Wendel JF. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome. 2001;44:321–30.

72. Wang J, Tian L, Lee HS, Wei NE, Jiang H, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, Chen ZJ. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. Genetics. 2006;172:507–17.

73. Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. Plant Cell. 2000;12:1551–68.

74. Wassenegger M, Heimes S, Riedel L, Sanger HL. RNA-directed de novo methylation of genomic sequences in plants. Cell. 1994;76:567–76.

75. Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AM, Strahl BD, Patel DJ, Jacobsen SE. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature. 2013;498:385–9.

76. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.

77. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.

78. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.

79. Frith MC, Hamada M, Horton P. Parameters for accurate genome alignment. BMC Bioinformatics. 2010;11:80.

80. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.

81. Schultz MD, Schmitz RJ, Ecker JR. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. Trends Genet. 2012;28:583–5.