Long non-coding RNAs and their proposed functions in fibre development of cotton (*Gossypium* spp.)

Maojun Wang[1], Daojun Yuan[1], Lili Tu[1], Wenhui Gao[1], Yonghui He[1], Haiyan Hu[1], Pengcheng Wang[1], Nian Liu[1], Keith Lindsey[2] and Xianlong Zhang[1*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, Hubei, China

[2]Integrative Cell Biology Laboratory, School of Biological and Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, United Kingdom

*Corresponding author: Xianlong Zhang

E-mail: xlzhang@mail.hzau.edu.cn

Tel: +86-27-87280510

Fax: +86-27-87280196

Summary

Long non-coding RNAs (lncRNAs) are transcripts of at least 200 bp in length, that possess no apparent coding capacity and are involved in various biological regulatory processes. Until now, no systematic identification of lncRNAs has been reported in cotton (*Gossypium* spp.).

Here, we describe the identification of 30,550 long intergenic non-coding RNA (lincRNA) loci (50,566 transcripts) and 4,718 long non-coding natural antisense transcript (lncNAT) loci (5,826 transcripts). LncRNAs are rich in repetitive sequences and preferentially expressed in a tissue-specific manner. The detection of abundant genome-specific and/or lineage-specific lncRNAs indicated their weak evolutionary conservation. Approximately 76% of homoeologous lncRNAs exhibit biased expression patterns towards the At or Dt subgenomes. Compared with protein-coding genes, lncRNAs showed overall higher methylation levels and their expression was less affected by gene body methylation.

The expression validation in different cotton accessions and co-expression network construction helped identify several functional lncRNA candidates involved in cotton fibre initiation and elongation. Analysis of integrated expression from the subgenomes of lncRNAs generating miR397 and its targets due to genome polyploidization indicated their pivotal functions in regulating lignin metabolism in domesticated tetraploid cotton fibres.

This study provides a first comprehensive resource of lncRNAs in *Gossypium*.

Keywords: cotton lncRNAs methylation polyploidization fibre development

Introduction

Generally, long non-coding RNAs (lncRNAs) are transcripts of at least 200 bp in length, possess no apparent coding capacity but are involved in various biological regulatory processes (Rinn and Chang, 2012). On the basis of their genomic localization with respect to protein-coding genes, lncRNAs can be classified as long intergenic non-coding RNAs (lincRNAs), long non-coding natural antisense transcripts (lncNATs), long intronic non-coding RNAs and overlapping lncRNAs that partially overlap with protein-coding genes (Derrien *et al*., 2012). Compared to protein-coding genes and even small non-coding RNAs, most lncRNAs lack strong sequence conservation between species (Marques and Ponting, 2009; Necsulea *et al*., 2014). LncRNAs are usually expressed at low levels and often exhibit tissue-specific patterns (Cabili *et al*., 2011), raising the possibility that lncRNAs regulate tissue development. In animals, lncRNAs have been demonstrated to be involved in chromatin modification, transcriptional regulation and post-transcriptional regulation (Geisler and Coller, 2013; Cech and Steitz, 2014). A recent study shows that lncRNAs may play an important role in *de novo* protein evolution (Ruiz-Orera *et al*., 2014).

With the rapid advances in sequencing technology and transcriptomic analysis, thousands of lncRNAs have been now identified in several plant species. In *Arabidopsis*, more than 6,000 lincRNAs have been identified using Tiling Array and RNA-seq (Liu *et al*., 2012). More recently, 37,238 lncNATs were identified and their responses to light were characterized (Wang *et al*., 2014). In a study of the origins of small RNAs, Zhou *et al*. (2009) identified more than 7000 lncNATs in rice. In maize, 20,163 lincRNAs were identified by integrating public EST databases and RNA-seq data (Li *et al*., 2014). The public databases PLncDB and PlantNATsDB store lincRNAs from *Arabidopsis* and lncNATs from 69 plant species, respectively (Chen *et al*., 2012; Jin *et al*., 2013).

While many sequences have been identified, a detailed functional analysis of plant lncRNAs is still in its infancy. For example, lncNAT COOLAIR and intronic lncRNA COLDAIR have been demonstrated to be vital for vernalization in *Arabidopsis* (Swiezewski *et al*., 2009; Wang *et al*., 2014). Viroids, a class of sub-viral plant-pathogenic lncRNAs, can regulate gene expression through a small RNA-guided pathway after their degradation (Navarro *et al*., 2012). LDMAR in rice was found to

be required for normal pollen development under long-day conditions (Ding *et al*., 2012). In addition, the DNA-dependent RNA Polymerase V (Pol V)-dependent lncRNAs are involved in RNA-directed DNA methylation (RdDM) by acting as scaffold RNAs (He *et al*., 2014; Matzke and Mosher, 2014).

Cotton (*Gossypium* spp.) is widely cultivated and utilized for its single-celled fibre in the textile industry and is also an important oilseed crop. *Gossypium* belongs to the Malvaceae and diverged from a common ancestor with *Theobroma cacao* (Paterson *et al*., 2012; Wang *et al*., 2012). Generally, the genus *Gossypium* is categorized into 45 diploid species (A-G,K; 2n = 2x = 26) and 5 tetraploid species (AADD, 2n = 4x = 52), with genome sizes varying about 3-fold, from ~880 Mb to ~2.5 Gb (Hawkins *et al*., 2006; Wendel *et al*., 2010). The tetraploid species were formed approximately 1-2 million years ago by the reunification of two divergent diploid species *Gossypium arboretum* (A2) and *Gossypium raimondii* (D5) (Senchina *et al*., 2003). Human domestication has produced the high-yielding tetraploid *Gossypium hirsutum* (Upland cotton, AADD, AD1 genome), whereas *Gossypium barbadense* (Sea-Island cotton, AADD, AD2 genome) is exploited for the superior length, strength, and fineness of the fibres (Kim and Triplett, 2001). Because of its excellent genetic and genomic resources, cotton is regarded as a good model to study genome polyploidization (Paterson *et al*., 2012), and the cotton fibre is an excellent experimental system for studying cell fate determination, cell elongation and cell wall formation (Guan and Chen, 2013).

Studies on non-coding RNAs in cotton have been largely limited to small RNAs until now, and RNA sequencing has helped identify hundreds of small non-coding RNAs. For example, Wei *et al*. (2013) identified miRNAs expressed during anther development in genetic male sterile and wild type cottons and Yang *et al*. (2013) identified miRNAs in cotton somatic embryogenesis. Gong *et al*. (2013) identified 33 miRNA families that were conserved between the A and D genomes. Xue *et al*. (2013) confirmed the expression of 79 miRNA families and identified 257 novel miRNAs related to cotton fibre elongation. Functional analysis of miR828 and miR858 identified roles in the regulation of homoeologous MYB2 in allotetraploid *G. hirsutum* fibre development (Guan *et al*., 2014). Recent transgenic analysis of miRNA156/157 indicated a fundamental role in fibre elongation (Liu *et al*., 2014).

We aimed to identify lncRNAs in the allotetraploid cotton species *G. babardense*, following genomic and RNA sequencing. We integrated 162 public unstranded transcriptomic sequencing datasets and generated 9 stranded transcriptomic sequences representing the main tissues of cotton to identify lncRNAs. In total, we identified 50,566 lincRNAs and 5,826 lncNATs in *G. babardense*. To assign these lncRNAs to subgenomes, we studied their homoeologous expression bias, and characterized the methylation profiles of lncRNAs and compared them with protein-coding genes. We went on to identify functional lncRNA candidates by differential expression analysis and co-expression network construction during cotton fibre development.

Materials and Methods

Plant material, library construction and sequencing

Plant seeds of cotton accession 3-79 (*Gossypium barbadense*) were sown in the glasshouse. When two fully expanded leaves appeared, root, hypocotyl and leaf were excised separately, frozen immediately in liquid nitrogen and stored at -70℃ until use. To collect cotton fibre samples, plants were grown in the field in Wuhan, China. Flowers were tagged at the day of blooming (0 day post anthesis, 0 DPA), and bolls were collected at 10 DPA and 20 DPA (Table S2). Samples from different plants were pooled. Total RNA was isolated from these samples using the Spectrum Plant Total RNA Kit (Sigma-Aldrich). Libraries were constructed using the Illumina TruSeq Stranded RNA Kit following the kit's recommendation. Strand-specific sequencing was performed on the Illumina HiSeq 2000 system (paired end 100 bp reads).

Publicly available datasets used in this study

We downloaded 154 RNA datasets of *Gossypium* species from the NCBI Sequence Read Archive collection sequenced on the Illumina platform, which include Zebularine-treated RNA and control datasets released by the Plant Industry of Commonwealth Scientific and Industrial Research Organisation (CSIRO) (Table S1). We downloaded 13 *Gossypium* 454 long reads sequencing datasets from the NCBI Sequence Read Archive and integrated all the public ESTs of cotton (Table S3). We also obtained 4 whole genome DNA methylation sequencing datasets released by

Joshua A. Udall laboratory (SRX331701). The 7 small RNA and 3 degradation sequencing datasets of cotton fibre tissues were from our laboratory (Liu *et al*., 2014).

lncRNA identification

All the RNA datasets were processed by removing adaptors and trimming low-quality bases (Q >20). The clean sequencing reads were mapped independently to the *Gossypium barbadense* genome using the spliced read aligner Tophat (Trapnell *et al*., 2009). We then applied two iterations of Tophat alignments proposed by Cabili *et al*. (2011) to maximize the splice junction site information from all samples. We separately assembled the transcriptomes using Cufflinks (Trapnell *et al*., 2010). The Cuffcompare procedure was applied to compare all the assemblies to the genome annotation of *G. barbadense*.

We then adopted 6 steps to identify *bona fide* lncRNAs from the novel and antisense transcripts of transcriptome assemblies: 1) transcripts were removed that were detected in fewer than two experiments; 2) transcripts with mapping coverage less than half of transcript length were removed; 3) transcripts were removed that derived from rRNA and tRNA (cutoff E-value 0.001); 4) transcripts with length less than 200 bp were removed; 5) transcripts were searched against the Swiss-Prot and Pfam databases to eliminate transcripts encoding proteins and protein-coding domains (cutoff E-value 0.001); 6) transcripts were removed that did not pass protein-coding-score test by the Coding Potential Calculator (CPC) and Coding-Non-Coding Index (CNCI) softwares (Sun *et al*., 2013). The optimized parameters of Coding-Non-Coding Index were trained using a lncRNA dataset from *Arabidopsis* (Liu *et al*., 2012). To verify the lncRNA identification, the public 454 datasets and ESTs were mapped to the lncRNA transcripts by blastn (E-value cutoff 1e-10, coverage >0.8).

Expression analysis

We employed the Tophat software (with -G parameter) to map all clean RNA-seq reads to the *G. barbadense* genome. The normalized expression of lncRNA and protein-coding transcripts were estimated using all mapped reads by Cufflinks. The multi-read and fragment bias correction methods embedded in Cufflinks were adopted to improve the accuracy of expression level estimation. The differentially expressed

genes were identified using DESeq package (adjusted p value 0.01 and at least two-fold change) (Anders and Huber, 2010).

Nearest neighbour analysis

Based on the genome location of the lncRNAs and protein-coding genes, the nearest protein-coding genes around each lincRNA at upstream and downstream positions within 5 kb were identified. For lncNATs, we identified the protein-coding genes on the antisense strand. Pearson correlation was employed to explore the expression relationship between these lincRNA/protein-coding gene and lncNAT/protein-coding gene pairs. The GO terms of nearest protein-coding genes with highly similar expression patterns were mapped to lincRNAs for enrichment analysis, similar to the method described by Pauli *et al*. (2012).

Tissue specificity analysis

To determine the tissue specificity of lncRNAs and protein-coding genes, we followed the entropy-based measure suggested by Cabili *et al*. (2011). Expression values of genes in samples were firstly normalized to density vectors. Then, the distance between two tissue expression patterns was defined by JS divergence. Finally, we defined the tissue specificity score per transcript using the maximal tissue specificity score of all tissues.

Genome synteny of lncRNA

The scaffolds of the At and Dt subgenomes were aligned to *G. arboretum* and *G. raimondii* diploid genomes using LASTZ respectively (Harris, 2007). The best mapping results allowing at least 60% coverage were sorted along the diploid chromosomes to construct pseudochromosomes. The syntenic blocks with at least five genes between At and Dt subgenomes were identified using MCScanX software (Wang *et al*., 2012). We referred the homoeologous lincRNA pairs based on the overlapping of these transcript loci to syntenic blocks and also evidenced by blastn reciprocal best hits with coverage of at least 90%.

Methylation data analysis

After clipping adapters and trimming low quality reads, the clean bisulphate-treated DNA sequencing reads were aligned to the *G. barbadense* genome using Bismark software (-N 1, -L 30) (Krueger and Andrews, 2011). Only unique mapping reads were retained for further analysis. Methylated cytosines covered by at least three reads were identified using binomial distribution (p value cutoff 1e-5). Customized Perl scripts were programmed to calculate the CG, CHG and CHH ratio per transcript.

miRNA prediction

The clean data of small RNA sequencing (miRNAs and small RNAs, smRNA) were mapped to *G. barbadense* using Bowtie, which allowed 200 multiple mapping positions and zero mismatch for each read. We adopted structure-based annotation and probability-based annotation to predict miRNA loci as suggested by Paterson *et al*. (2012). For the structure-based annotation, RNAfold was employed to predict secondary structures and miRcheck was used to evaluate secondary structures (Jones-Rhoades and Bartel, 2004). We then utilized miRDP to filter the putative precursors of the structure-based annotation (Yang and Li, 2011). All the annotated mature miRNAs were searched against the miRBase (Release 20) to categorize them into cotton conserved and non-conserved miRNA gene families (Kozomara *et al*., 2013). We also employed the CleaveLand pipeline to predict putative miRNA targets based on the degradation data (Addo-Quaye *et al*., 2009). The *bona fide* miRNA targets were detected based on the criteria suggested by Addo-Quaye *et al*. (2008).

Network construction

Weighted gene co-expression analysis (WGCNA) was employed to construct the network (Langfelder *et al*., 2008). The framework for network construction can be summarized as: 1) defining a gene co-expression similarity by the pearson correlation; 2) applying an adjacency function to transform the co-expression similarities to connection strengths with a soft thresholding power of 10; 3) identifying network modules consisting of the highly correlated gene expression patterns using the hierarchical clustering with topological overlap matrix. Non-module genes were categorized by a 'grey' colour. All the steps for network analysis were completed using language R. The software VisANT was used to graphically visualize networks (Hu *et al*., 2013).

Quantitative Real-Time PCR

RNA samples from ovules at -1, 0, 4 and 5 DPA and fibres at 10 DPA and 20 DPA were collected, and quantitative real-time PCR was performed as described previously and the expression levels were normalized using UB7 (Tan *et al*., 2013). The PCR products at 10 DPA and 20 DPA fibres were cloned into the pGEM-T vector and the randomly selected 100 clones were each sequenced.

RLM-RACE

The RLM-RACE was performed to validate the splicing site of miRNA target genes using GeneRacer kit (Invitrogen, https://www.lifetechnologies.com). Total RNA (5 μg) from 10 DPA and 20 DPA fibres were ligated to RNA adapter without calf intestinal phosphatase treatment. Further PCR reactions using 5' adaptor primers and 3' gene-specific primers were guided by the manufacturer's instructions.

Data access

The stranded RNA-seq data have been submitted to the NCBI Sequence Read Archive under the Bioproject ID PRJNA266265. The lncRNA sequences and genome coordinate files can be accessed from our genome website at http://cotton.cropdb.org/cotton/download/data.php.

Results

Identification and characterization of cotton lncRNAs

In order to develop a comprehensive catalogue of lncRNAs in *Gossypium*, a prerequisite is to integrate a high-quality and high-depth RNA-seq dataset. We collected 154 public and 8 in-house Illumina transcriptomes (Table S1). To determine the orientation of transcripts accurately, we also generated 9 transcriptomes covering different developmental stages of *G. babardense* using the stranded sequencing method (Table S2). In total, this collection represents more than 5 billion clean reads for lncRNA identification.

We mapped RNA-seq data from diploids and tetraploids to the subgenomes and the whole genome of *G. babardense* independently (data from our unpublished *G.*

barbadense genome sequence, 29,751 scaffolds, N50 260.06 kb, encoding 80,876 protein-coding genes) in order to perform *de novo* transcript assembly using the Tophat-Cufflinks pipeline. Some filtering steps were conducted to retain *bona fide* lncRNAs (Fig. 1a). This pipeline provided 30,550 lincRNA loci (50,566 transcripts) and 4,718 lncNAT loci (5,826 transcripts).

To verify the reliability of prediction, we aligned all the lincRNAs to 425,526 public cotton ESTs. A total of 2,929 lincRNAs (5.8%) were supported by at least one EST. We also aligned all the lincRNAs to the collected 454 sequencing reads (Table S3) and observed that 12,029 lincRNAs (23.8%) were supported by at least one read. Attributing lncRNAs to subgenomes showed that the number of lncRNAs in the At subgenome was approximately 2,900 larger than that in the Dt subgenome (Table S4). The exon number distribution of lncRNAs showed that the *G. barbadense* genome encoded 63% single-exonic lincRNAs and 77% single-exonic lncNATs, which are significantly higher proportions than those of protein-coding transcripts (15%; Fig. 1b). The mean transcript length of lncRNAs was typically shorter than protein-coding genes (average length: 504 bp for lincRNAs, 713 bp for lncNATs and 1,621 bp for protein-coding transcripts; Fig. 1c).

GC content is believed to be related to the biased intergenomic nonreciprocal DNA exchanges in the tetraploid cotton genomes (Guo *et al*., 2014). In this study, we observed that both the distributions of GC content amongst lncRNAs (lincRNAs and lncNATs) and protein-coding genes exhibit no apparent differences between the At and Dt subgenomes (Kolmogorov-Smirnov test, lncRNA p-value 0.1486, protein-coding genes p-value 0.1803; Fig. 1d). However, lincRNAs show the lowest GC content (median 37.1%), followed by lncNATs (median 40.6%), and protein-coding genes (median 41.8%) the highest in each subgenome.

The *G. barbadense* genome is highly enriched for repetitive sequences (70%), with the At subgenome at 74% and the Dt subgenome at 63%. Overlapping coordinates of lncRNAs with transposable elements (TE), we found that 55.8% of lincRNAs contained TE, corresponding to the At subgenome with 58.1%, the Dt subgenome with 54.8% and ungrouped scaffolds with 48.8% (Fig. 1e). The fraction of TE-containing lncNATs was less than half relative to lincRNAs, with At subgenome at 23.2%, Dt subgenome at 21.7% and ungrouped scaffolds at 23.7%. This result is comparable to the studies in animals, such as mouse, zebrafish and human (Kapusta *et*

297 *al*., 2013). The LTR retrotransposons of the Gypsy family occupied a dominant
298 proportion of repetitive sequences in lincRNAs, which was the same as its distribution
299 at the genome level (Fig. 1f). Long interspersed nuclear elements (LINE) only
300 occupied 6% of the genome, but showed an increased abundance to 14% in lincRNAs,
301 and up to 37% in lncNATs.
302
303 **Expression of cotton lncRNAs among tissues**
304 The stranded RNA-seq data were adopted to systematically explore lncRNA
305 expression among 9 different tissues/samples. The results showed that the highly
306 differentiated tissues anther and cotton fibres at 20 DPA expressed fewer genes than
307 others (Fig. 2a). The overall expression levels of both lincRNAs and lncNATs were
308 lower than of protein-coding transcripts (Fig. 2b), consistent with a previous study
309 (Cabili *et al*., 2011). Given that lncRNAs may function in regulating adjacent
310 protein-coding genes and thus possess similar expression patterns, we examined this
311 possibility by computing the Pearson correlation coefficients (r_p) between lincRNAs
312 and the nearest protein-coding genes (within 5 kb) (lincRNA-PCgene); lncNATs and
313 the corresponding protein-coding genes on the opposite strand (lncNAT-PCgene); and
314 the nearest protein-coding pairs lacking an intervening gene (PCgene-PCgene). In
315 total, we identified 10,749 lincRNA-PCgene pairs, 5,826 lncNAT-PCgene pairs and
316 25,449 PCgene-PCgene pairs. Compared with randomly sampled transcript pairs, we
317 observed high ratios of extremely positive correlations between lincRNA-PCgene (16%
318 vs. 6%, r_p>0.8), lncNAT-PCgene (35% vs. 4%, r_p>0.8) and PCgene-PCgene (24% vs.
319 6%, r_p>0.8) pairs (Fig. 2c). The expression relationships between these pairs provide
320 candidates to be tested in further functional studies.
321 To evaluate the tissue specificity of expression, the JS scores (an entropy-based
322 measure) of transcripts were calculated (Cabili *et al*., 2011). The density distributions
323 of lincRNAs and lncNATs were significantly different from protein-coding transcripts
324 (Kolmogorov-Smirnov test, p value < 2.2e-16; Fig. 2d). Using a JS score of 0.5 as a
325 cutoff, we found that 42% of lincRNA and 51% of lncNAT transcripts were
326 tissue-preferentially expressed, dramatically higher than the percentage of
327 protein-coding transcripts (18%) across the 9 tissues/samples. Further quantitative
328 analysis showed that anther expressed the largest number of tissue-preferential genes
329 (3,140 protein-coding transcripts, 3,925 lincRNAs and 787 lncNATs) though the total

number of expressed transcripts was smaller than for other samples (Fig. 2e). In contrast, fibres at 20 DPA expressed a relatively small number of specific genes (973 protein-coding transcripts, 852 lincRNAs and 230 lncNATs), slightly higher than for stigma. Randomly selected tissue-preferentially expressed lncRNAs were verified by RT-PCR (Fig. 2f). These results indicate that a large number of lncRNAs were expressed preferentially in particular tissues.

Evolution history and subgenome expression partition

It is believed that the sequences of lncRNAs are less conserved than protein-coding transcripts (Marques and Ponting, 2009; Necsulea *et al*., 2014), and we were interested to know how many cotton lncRNAs are inherited from closely related species.

We firstly aligned the lncRNAs of the At and Dt subgenomes to each reciprocally, then to the diploid A and D genomes, and also to the closely related species *Theobroma cacao* and the more distant dicot *Vitis vinifera* (Jaillon *et al*., 2007; Argout *et al*., 2011). Using all the lncRNA transcripts in the At subgenome as queries, we found that 99.5% had homologous copies in the diploid A genome, 76.7 % in the Dt subgenome and 75.6 % in the diploid D genome (Fig. 3a). However, only 6.8% of the lncRNAs in the At subgenome were found to match homologous regions in the *T. cacao* genome and 2.6 % in the *V. vinifera* genome. Similar results were observed when lncRNAs in the Dt-subgenome were used as query sequences (Fig. S1a). These results suggest that the vast majority of lncRNAs were species-specific or limited to closely related species.

As relatively highly expressed neighbour protein-coding genes may have functional relationships with lncRNAs, we mapped the GO terms of such protein-coding genes ($r_p > 0.9$) to lncRNAs in order to predict their possible functions. The results showed that the At subgenome-specific lncRNAs were enriched in ribosome assembly, spermine biosynthesis process and microtubule cytoskeleton organization (Fig. 3b). Dt subgenome-specific lncRNAs were enriched in lignin catabolic process, response to biotic stimulus and carbon utilization (Fig. 3c). The conserved lncRNAs in *T. cacao* and *V. vinifera* were enriched in fundamental biological processes, such as translation elongation, peroxisome organization and L-phenylalanine catabolism (Fig. S1b).

Despite rapid gene fractionation, the majority of lncRNAs were conserved between the At and Dt subgenomes. Using data from the recently released *G. arboreum* and *G. raimondii* genomes, we ordered the scaffolds of At and Dt subgenomes to pseudochromosomes based on whole genome alignment (Fig. S2). Through genome-wide synteny analysis, we identified 377 syntenic blocks between the At and Dt subgenomes representing 9,262 protein-coding gene pairs (Fig. 3d). Overlapping lncRNAs with these syntenic blocks and using a reciprocal best hit alignment (coverage cutoff 0.9), we identified 1,090 homoeologous lincRNA pairs between the At and Dt subgenomes, of which 900 pairs were anchored on pseudochromosomes. Genomic landscape analysis showed that both of lncRNAs and protein-coding genes were preferentially located in regions with poor repetitive sequences assuming as a negative correlation (Fig. S3), especially for the protein-coding genes (Fig. 3d).

As highlighted in recent studies, the non-additivity of gene expression, also referred as 'transcriptomic shock', appears to be widespread in newly formed allopolyploids (Yoo *et al*., 2013). Hierarchical clustering of homoeologous lincRNAs showed that those from a total of 8 tissues/samples were clustered in a subgenome-specific manner with the exception of those derived from anther (Fig. S4a), contrasted with the result by clustering protein-coding genes (Fig. S4b). The averaged expressions of lincRNA pairs across tissues were compared (Fig. 3d). This led to the identification of 196 pairs expressed dominantly in At-subgenome and 188 pairs expressed dominantly in Dt subgenome. However, the overall comparison ignored the detailed bias in patterns in different tissues, and so we categorized the expression patterns into four types.

Based on these analyses, the expression of 305 pairs were At-biased, 315 pairs were Dt-biased and 67 pairs were chimeric-biased. Therefore, we conclude that expression bias of lincRNAs was extensive in tetraploid cottons in a subgenome-specific manner, and the numbers of bias-expressed pairs in each subgenome were comparable.

Methylation of lncRNAs

DNA methylation is widespread as a means of regulating protein-coding gene transcription in diverse organisms. To characterize the methylation patterns of

lncRNAs, we obtained 4 bisulphate-converted DNA sequencing datasets of petal in cotton species, including diploid *G. arboretum*, *G. raimondii*, an F1-hybrid between *G. arboretum* and *G. raimondii*, and the natural tetraploid *G. hirsutum*. The clean reads were uniquely mapped to the *G. barbadense* genome to dissect cytosine methylation (Table S5). The numbers of methylated sites in the At and Dt subgenomes were summarized using each dataset and the percentages of DNA methylation in CG, CHG and CHH contexts were compared (Table S6).

At the chromosomal level, highly methylated regions showed preferentially a particular abundance of TEs, seen as a broadly positive correlation. However, protein-coding genes in these regions were expressed at generally low levels (Fig. 4a). This phenomenon was observed in all the four datasets used to analyse diploids and tetraploids. Compared with protein-coding genes, lincRNAs showed higher methylation levels in CG and CHG contexts, but comparable methylation levels in a CHH context (Fig. 4b; Fig. S5). Specifically, the CG methylation levels in exon regions of protein-coding genes rapidly increased when departing from the transcription starting sites and termination sites. However, no such obvious methylation patterns were seen for lincRNAs. For CHG and CHH methylation, the upstream, exon and downstream regions of lincRNAs showed no obvious differences.

Many studies have found that the methylation levels of upstream and genic sequences are negatively correlated with the expression levels of protein-coding genes. However, few studies have focused on the relationship between DNA methylation and lncRNA expression. To investigate this, we used RNA-seq data from the same sample as bisulphite-converted DNA sequencing to quantify expression levels of lincRNAs in petals. It was found that in all the three methylation contexts, genes with very high expression levels displayed low methylation levels while highly methylated genes displayed low expression levels, indicating a negative correlation between DNA methylation and gene expression for both of lincRNAs and protein-coding genes (Fig. 4c). Specifically, in upstream regions, the scatter-plots of protein-coding genes tended to cover lincRNAs in all three methylation contexts. Interestingly, for gene body methylation, protein-coding genes showed a tighter distribution of methylation levels in each of the three contexts than did lincRNAs. Analysis of accumulated frequency distribution of methylation levels to the relative gene number demonstrated that gene body methylation of lincRNAs in each methylation context was significantly different

from that for protein-coding genes, whereas upstream methylation showed no significant differences. These studies suggest that gene body methylation has a generally stronger effect on protein coding gene expression than for lincRNAs.

To reveal the direct effects of methylation on lncRNA expression, we collected RNA-seq data from Upland cotton ovules at 0 DPA treated with zebularine, a DNA methylation inhibitor forming a covalent complex with DNA methyltransferases (Zhou *et al*., 2002). After analysing the quality of RNA-seq (Fig. S6a), we observed that the expression levels of lincRNAs were quite variable and up-regulated expression was clearly consistent along each chromosome after zebularine treatment, while the expression levels of protein-coding genes varied less (Fig. S7).

We then conducted a differential gene expression analysis (Fig. 4d). The results showed that a total of 9,917 lincRNA transcripts were differentially expressed, among which the majority (94.4%) were highly expressed in treated ovule samples. In contrast, only 52.2% of differentially expressed protein-coding transcripts were highly expressed in treated samples. Intriguingly, the 86% of up-regulated lincRNAs in the At subgenome and 87% in the Dt subgenome contained repetitive sequences (Fig. 4e), which was a value much higher than for the down-regulated lincRNAs (32% of the At subgenome and 36% of the Dt subgenome) and also higher than ratios of all the lincRNAs in the At and Dt subgenomes (58% of the At subgenome and 55% of the Dt subgenome). Further functional enrichment of the differentially expressed transcripts revealed that up-regulated lincRNAs in treated samples were enriched in DNA integation, cytoskeleton organization, regulation of pH and cell death, while down-regulated lincRNAs were enriched in respiratory gaseous exchange, protein ubiquitination and nucleoside metabolic process (Fig. S6b).

Small RNAs generated by lncRNAs

lncRNAs can be small RNA precursors and can also negatively regulate miRNA maturation (Plosky, 2014). We collected 7 sets of small RNA sequencing data for *G. barbadense* fibres, representing three important developmental stages (-3 DPA, 0 DPA, 3 DPA for fibre initiation stage, 7 DPA and 12 DPA for fibre elongation stage, 20 DPA and 25 DPA for fibre secondary cell wall synthesis stage) to identify putative small RNA precursors. The miRNA prediction resulted in a total of 318 conserved miRNAs and 227 non-conserved miRNAs (Table S7, S8). All the lincRNAs were

then overlapped to precursors of miRNAs from genome-wide miRNA predictions. We found 128 lincRNAs as possible precursors of conserved miRNAs related to 25 families and 101 lincRNAs as possible precursors of non-conserved miRNAs (Table S9). Three well-known miRNAs were covered in this study and presented as examples (Fig. S8). In addition to functioning as miRNA precursors, abundant lncRNA transcripts may be degraded to form smRNAs. The mapping of smRNA reads showed that 4,707 lincRNA transcripts (9.3%) were mapped sense and 4,131 (8.2%) were mapped antisense to endo-smRNA reads (Table S9). Future experimental studies are necessary to demonstrate the function of these lincRNAs, but are beyond the scope of the current work.

Functional lncRNA candidates in cotton fibre development

Cotton fibre initiation is a fundamental stage determining the fate of the fibre cell. Lint fibres are believed to appear on the day of anthesis (0 DPA) and fuzz fibres develop on the fourth day post anthesis (4 DPA) (Zhang *et al*., 2007). To identify putative functional lncRNAs contributing to the initiation of lint and fuzz fibres, the expression of 20 randomly selected lncRNAs that were highly expressed in ovules of *G. barbadense* 3-79 was determined in 8 different genotypes of Upland cotton (*G. hirsutum*). These cotton accessions include 3 lint-fuzz (TM-1, Xuzhou-142 and YZ1) wild types, 2 lintless-fuzzless mutants (Xuzhou-142 lintless-fuzzless (XZ142WX) and Xinxiangxiaoji lintless-fuzzless (XinWX)) and 3 linted-fuzzless mutants (n2, GZnn and GZNn) (Fig. 5a).

Hierarchical clustering analysis showed that most lncRNAs were preferentially expressed in lint-fuzz cotton ovules at -1 and 0 DPA or 4 and 5 DPA (Fig. 5b, c). Specifically, the expression of one lncRNA (LINC02) was highlighted, the expression of which might in part underlie the development of lint and fuzz fibres. This lncRNA produced significantly higher transcription levels in lint-fuzz/linted-fuzzless cottons than that in lintless-fuzzless cottons (p-value < 0.05), but no different transcription levels were seen between lint-fuzz and linted-fuzzless cotton ovules at -1 or 0 DPA ovules (Fig. 5d). We also observed the higher transcription levels in lint-fuzz cottons than that in lintless-fuzzless/linted-fuzzless cottons at 4 DPA or 5 DPA (p-value < 0.05) (Fig. 5e).

To predict the functional roles of lncRNAs in the 'fibre elongation' and 'secondary cell wall synthesis' stages of fibre development, we applied a weighted gene co-expression network analysis (WGCNA) using published cotton fibre transcriptomes at 10 DPA and 20 DPA (Fig. S9). After removing the low-expressed transcript pairs, 720 lincRNA pairs and 6,858 protein-coding gene pairs were retained for network construction. The network was partitioned into 17 modules (Fig. 6a). Hierarchical clustering and functional enrichment of these modules showed they displayed different characteristics (Fig. S10, S11).

The module M12 is highlighted here (Fig. 6c). Transcripts in this module were At-biased in their expression and significantly enriched in heterocyclic metabolic and cofactor metabolic processes (Fig. S10). Hub genes often play founder roles and can define the functional foci in networks (Langfelder *et al.*, 2008). The phosphoenolpyruvate carboxylase-related kinase 2, involved in protein phosphorylation, and a ubiquitin-specific protease were regarded as two hub genes. Interestingly, one lincRNA pair, designated as P1, was highlighted as a hub gene, suggesting a vital functional role in this module (Fig. 6c).

Another module, M16, was highlighted as a representative of a Dt-bias expression module (Fig. 6b). This module involved 18 lincRNA pairs and was enriched in oxidation-reduction and small molecule metabolic processes. Previous studies have showed that regulation of reactive oxygen species levels plays a pivotal role in the formation of spinnable cotton fibre (Hovav *et al.*, 2008). Consistent with this, we found that key genes related to reactive oxygen species metabolism, such as 2-oxoglutarate (2OG) and Fe (II)-dependent oxygenase, flavin-binding monooxygenase and alpha-helical ferredoxin, were involved in this module. The RabGAP domain-containing protein related to small GTPase mediated signal transduction, categorized as 'small molecule metabolic process', was also involved (Fig. 6d).

Integrated expression of lncRNAs generating miR397 and their targets in cotton fibre development

Comparative analysis of lncRNAs with small RNA sequencing data helped identify one pair of lncRNAs preferentially expressed in fibres, that were precursors of miR397 from the At and Dt subgenomes (Fig. S12). The Dt-derived lncRNA was

highly expressed, and suppressed its At sugenome homoeologue at 10 DPA (Fig. 7a). Conversely, at 20 DPA, the expression of At-subgenome copy reached a very high level, while the expression of Dt-subgenome copy was reduced to a quite low level. This observation was confirmed by the sequencing of 100 randomly picked PCR clones (Fig. 7b). Moreover, the expression level of At-subgenome copy at 20 DPA was significantly higher (~10 fold) than that of the highly expressed Dt-subgenome copy at 10 DPA, which was verified by qRT-PCR detecting the total expression at 10 DPA and 20DPA (Fig. 7c).

The expression of these two lncRNAs was further analysed in two diploid progenitors and in domesticated and wild tetraploid cottons, using public RNA-seq data. In both diploids, we found the At-subgenome and Dt-subgenome copies were highly expressed in 20 DPA fibres (Fig. S13). We also found that the expression pattern of the At-subgenome copy in all the domesticated and wild Upland and Sea-Island cotton accessions was consistent with the observation in Sea-Island cotton 3-79 (Fig. 7d). For the Dt-subgenome copy, we observed the same expression pattern in domesticated Upland and the other 1 Sea-Island cottons, but a reverse expression pattern between 10 DPA and 20 DPA fibres in wild cottons. These results showed that strong directional human selection for enhancing fibre yield has prioritized the expression of the Dt-subgenome copy of lncRNA generating miR397 at 10 DPA, but retained the expression pattern of the At subgenome copy as the same as the diploid A genome and wild tetraploid cottons.

MiR397 was validated to target laccase (LAC) transcripts which are important regulators in lignin metabolism (Wang *et al*., 2012). We detected two types of such *LAC* genes (*LAC4a* and *LAC4b*; one gene locus in the At subgenome and one locus in the Dt subgenome for each type) in tetraploid cotton genomes (Fig. 7e). RNA-seq data showed that *LAC4a* in the At and Dt subgenomes retained the same expression pattern as diploid progenitors. Nevertheless, the Dt subgenome copy of *LAC4b* underwent an expression transition event the same as the Dt subgenome lncRNA (Fig. 7e). *LAC4b* was highly expressed at 20 DPA in *G. raimondii* (proposed Dt subgenome progenitor), which suppressed the expression level at 10 DPA. However, in tetraploid cotton, the Dt subgenome *LAC4b* (Gbscaffold30529.8.0) was highly expressed at 10 DPA and reduced at 20 DPA. These results were validated by qRT-PCR and random clone sequencing analysis (Fig. S14). Degradome sequencing

data showed an obvious cleavage activity of miR397 in *LAC4a* (Fig. 7f), indicating that miR397 could repress *LAC4a* by guiding mRNA degradation. In contrast, no cleavage signal was detected in *LAC4b*. Sequence alignment showed a SNP at the tenth site, which was crucial for miRNA-guided mRNA cleavage (Zheng *et al*., 2012), of miRNA binding region between *LAC4a* and *LAC4b*. The RLM-RACE results confirmed this finding (Fig. 7f).

To study the putative mechanisms of expression transition of the Dt subgenome *LAC4b*, we aligned its promoter and downstream regions with the diploid *G. raimondii* genome. Intriguingly, little evolutionary variations were observed at the upstream region (3 kb; Fig. 7g). However, an approximate 500 bp transposon inserted into the region downstream of *LAC4b* in the Dt subgenome, which was coming from a region downstream of the At subgenome *LAC4b* and might induce the expression transition (Fig. 7g, h). We confirmed this observation by directly sequencing these two regions from the At and Dt subgenomes.

Discussion

Increasing numbers of functional studies on protein-coding genes and small non-coding RNAs are revealing the high level of complexity of eukaryotic transcriptomes, especially when we consider the extensive abundance of long non-coding RNAs (Kapusta and Feschotte, 2014). However, limited data are available for plants. One of the reasons is the poor availability of complete reference genomes and high-depth transcriptome datasets. In cotton, several studies have identified small non-coding RNAs through small RNA sequencing but there are no data presented for lncRNAs. The recent publication of genome sequences and the accumulation of RNA-seq data make it feasible for genome-wide identification of lncRNAs.

In this study, we integrated high-quality RNA-seq data with high depth stranded RNA sequencing to explore lncRNAs. We obtained 50,566 lincRNA and 5,826 lncNAT transcripts. Due to the tetraploid genomic characteristics and large genome size of cotton, the number of lncRNAs is larger than previous identifications in *Arabidopsis* and maize (Liu *et al*., 2012; Li *et al*., 2014). We also believe that more lncRNAs may be identified using stressed plants, as reported for *Arabidopsis* (Liu *et al*., 2012). After attributing these lncRNAs to the At and Dt subgenomes, we observed

592 the number encoded by the At subgenome was 2,900 larger than in the Dt subgenome.
593 Further homoeologous sequence alignments showed that the At subgenome encoded
594 nearly 23% specific lncRNAs (Dt subgenome 17%), which is higher than the ratio of
595 protein-coding genes between these two genomes (Li *et al*., 2014). When compared
596 with data for the *T. cacao* and *V. vinifera* genomes, we found that lncRNAs diverged
597 quickly among closely related species and even in different genomes of *Gossypium*.
598 Further studies should be conducted to elucidate the functional roles of specific
599 lncRNAs in the At and Dt subgenomes, and those of other species.
600 Genome-wide methylation characterization of protein-coding genes has been
601 explored widely in animals and plants, but few systematic analyses of lncRNAs have
602 been carried out (Zemach *et al*., 2010). Therefore, we characterized the methylation of
603 lncRNAs using bisulphite-converted DNA sequencing data. It was found that the
604 methylation levels of lncRNAs were higher overall than for protein-coding genes. A
605 large proportion of differentially expressed lincRNAs were up-regulated in ovule
606 samples when treated with methyltransferase inhibitor, and the majority of these
607 lincRNAs overlapped with transposable elements.
608 Furthermore, the genome landscape of averaged gene expression levels in 500 kb
609 windows showed that the expression levels of lncRNAs were more obviously changed
610 compared to protein-coding genes. These results are consistent with the fact that more
611 than half of lncRNAs originated from transposable elements, which are generally
612 heavily methylated (Fedoroff, 2012), indicating that a large number of lincRNAs are
613 silenced in developing cotton ovules due to DNA methylation. These results suggest a
614 functional relationship between transposable elements, lncRNAs and DNA
615 methylation.
616 Functional characterization of lncRNAs is still in its infancy. High-throughput
617 methods, such as Chromatin isolation by RNA purification (ChIRP) and RNA
618 immunoprecipitation (RIP), have proved to be useful and have been utilized in many
619 studies (Chu *et al*., 2011; Quinn *et al*., 2014). In this study, we identified several
620 differentially expressed lncRNAs in cotton fibre initiation stage in different cotton
621 accessions, which might be in part associated with the development of lint and fuzz
622 fibres. These lncRNAs represent functional candidates for future experimental studies.
623 We then used a co-expression network strategy to predict function in cotton fibre

elongation and secondary cell wall synthesis stages by combining the expression of homoeologous protein-coding genes and lncRNAs across the At and Dt subgenomes.

We systematically explored the expression of one lncRNA pair generating miR397. The function of miR397 has been well studied in rice by down-regulating its target laccase-like gene transcripts (Zhang *et al*., 2013). The target of miR397, *LAC4*, can promote constitutive lignification in *Arabidopsis* (Berthet *et al*., 2011). In cotton fibres, accumulation of lignin will reinforce the fibre cell walls (Han *et al*., 2013). Therefore, we focused on the expression of lncRNAs and their target *LAC4* in developing cotton fibres.

The expression of two lncRNAs were biased in their subgenomes at different stages, and analysis in diploids and several domesticated and wild tetraploid cottons suggested that human domestication changed the expression pattern of the Dt subgenome lncRNA. Intriguingly, the expression pattern of the Dt subgenome *LAC4b* was also changed in the same manner as for the lncRNA. We speculate that the expression transition of the Dt subgenome *LAC4b* was induced by a TE insertion from the At subgenome. The finding of a SNP in the miRNA binding region between *LAC4a* and *LAC4b* suggests that *LAC4b* might be regulated by miR397 via translational inhibition (Li *et al*., 2013). Our study provides a framework to explore gene expression bias in tetraploid cotton and the molecular basis of miR397-guided lignin metabolism in fibre development.

In summary, our study is the first to characterize lncRNAs in *Gossypium* using high-depth RNA-seq data, although we were able to verify only part of lncRNAs by expression analysis. Future work will aim to dissect their biological functions in relation to cotton development and the genetics underpinning improved agronomic traits. In allopolyploid organisms, such as cotton, wheat and rapeseed, gene expression is to a significant level likely to be regulated by diverse epigenetic modifications (Chen, 2007), and therefore studies on lncRNAs are imperative, as some are most likely involved in epigenetic regulation, such as through chromatin modification and RNA-directed DNA methylation (RdDM). Our study provides new information that underpins the functional characterization of lncRNAs in allopolyploid plants.

Acknowledgements

We are very grateful to the laboratory of Dr Joshua A Udall for releasing the bisulphite converted DNA and transcriptome sequencing data in petals. We are also very grateful to the laboratory of Dr Elizabeth S. Dennis for releasing cotton ovule RNA-seq data treated with DNA methyltrasferases inhibitor. This work was financially supported by National Natural Science Foundation of China (NO. 31230056 and NO. 31201251) and Huazhong Agricultural University Independent Scientific & Technological Innovation Foundation (NO. 2014bs03).

References

Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ. 2008. Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome. *Current Biology* **18**, 758-762.

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

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

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN *et al*. 2011. The genome of *Theobroma cacao*. *Nature Genetics* **43**, 101-108.

Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cézard L, Le Bris P, Borrega N, Hervé J, Blondet E, Balzergue S, *et al*. 2011. Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of *Arabidopsis thaliana* stems. *Plant Cell* **23**, 1124-1137.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Development* **25**, 1915-1927.

Cech TR, Steitz JA. 2014. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94.

Chen DJ, Yuan CH, Zhang J. *et al*. 2012. PlantNATsDB: a comprehensive database of plant natural antisense transcripts. *Nucleic Acids Research* **40**, 1187-1193.

Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annual Review of Plant Biology* **58**, 377-406.

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. 2011. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Molecular Cell* **44**, 667-678.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG. *et al*. 2012. The GENCODE v7 catalog

of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22**, 1775-1789.

Ding JH, Lu Q, Ouyang YD, Mao HL, Zhang PB, Yao JL, Xu CG, Li XH, Xiao JH, Zhang QF. 2012. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proceedings of the National Academy of Sciences, USA* **109**, 2654-2659.

Fedoroff NV. 2012. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758-767.

Geisler S, Coller J. 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Review Molecular Cell Biology* **14**, 699-712.

Gong L, Kakrana A, Arikit S, Meyers BC, Wendel JF. 2013. Composition and expression of conserved microRNA genes in diploid cotton (*Gossypium*) species. *Genome Biology and Evolution* **5**, 2449-2459.

Guan X, Chen ZJ. 2013. Cotton Fiber Genomics. In Seed Genomics, pp. 203-216. Wiley-Blackwell.

Guan X, Pang M, Nah G, Shi X, Ye W, Stelly DM, Chen ZJ. 2014. miR828 and miR858 regulate homoeologous MYB2 gene functions in *Arabidopsis* trichome and cotton fibre development. *Nature Communication* **5**, 3050.

Guo H, Wang X, Gundlach H, Mayer KFX, Peterson DG, Scheffler BE, Chee PW, Paterson AH. 2014. Extensive and biased intergenomic nonreciprocal DNA exchanges shaped a nascent polyploid genome, *Gossypium* (Cotton). *Genetics* **197**, 1153-1163.

Han LB, Li YB, Wang HY, Wu XM, Li CL, Luo M, Wu SJ, Kong ZS, Pei Y, Jiao GL, *et al*. 2013. The dual functions of WLIM1a in cell elongation and secondary wall formation in developing cotton fibers. *Plant Cell* **25**, 4421-4438.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. Ph.D. thesis, Pennsylvania State University.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research* **16**, 1252-1261.

He XJ, Ma ZY, Liu ZW. 2014. Non-coding RNA transcription and RNA-directed DNA methylation in *Arabidopsis*. *Molecular Plant* **7**, 1406-1414.

Hovav R, Udall JA, Chaudhary B, Hovav E, Flagel L, Hu G, Wendel JF. 2008. The evolution of spinnable cotton fiber entailed prolonged development and a novel metabolism. *PLoS Genetics*. **4**, e25.

Hu Z, Chang YC, Wang Y, Huang CL, Liu Y, Tian F, Granger B, DeLisi C. 2013. VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic Acids Research* **41**, 225-231.

Jaillon O. *et al*. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467.

Jin J, Liu J, Wang H, Wong L, Chua NH. 2013. PLncDB: plant long non-coding RNA database. *Bioinformatics* **29**, 1068-1071.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular Cell* **14**, 787-799.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of Vertebrate long noncoding RNAs. *PLoS Genetics* **9**, e1003470.

Kapusta A, Feschotte C. 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics* **30**, 439-452.

Kim HJ, Triplett BA. 2001. Cotton fiber growth in planta and in vitro. Models for plant cell elongation and cell wall biogenesis. *Plant Physiology* **127**, 1361-1366.

Kozomara A, Griffiths-Jones S. 2013. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42**, 68-73.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559.

760 **Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C. *et al*.**
761 **2014.** Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nature*
762 *Genetics* **46**, 567-572.
763 **Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA,**
764 **Cole RA, Fowler JE. *et al*. 2014.** Genome-wide discovery and
765 characterization of maize long non-coding RNAs. *Genome Biology* **15**, R40.
766 **Li S, Liu L, Zhuang X, Yu Y, Liu X, Cui X, Ji L, Pan Z, Cao X, Mo B *et al*. 2013.**
767 MicroRNAs inhibit the translation of target mRNAs on the Endoplasmic
768 Reticulum in *Arabidopsis*. *Cell* **153**, 562-574.
769 **Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH.**
770 **2012.** Genome-wide analysis uncovers regulation of long intergenic
771 noncoding RNAs in *Arabidopsis*. *Plant Cell* **24**, 4333-4345.
772 **Liu N, Tu LL, Tang WX, Gao WH, Lindsey K, Zhang XL. 2014.** Small RNA and
773 degradome profiling reveals a role for miRNAs and their targets in the
774 developing fibers of *Gossypium barbadense*. *Plant Journal* **80**, 331-344.
775 **Marques AC, Ponting CP. 2009.** Catalogues of mammalian long noncoding RNAs:
776 modest conservation and incompleteness. *Genome Biology* **10**, R124.
777 **Matzke MA, Mosher RA. 2014.** RNA-directed DNA methylation: an epigenetic
778 pathway of increasing complexity. *Nature Review Genetics* **15**, 394-408.
779 **Navarro B, Gisel A, Rodio ME, Delgado S, Flores R, Di Serio F. 2012.** Small
780 RNAs containing the pathogenic determinant of a chloroplast-replicating
781 viroid guide the degradation of a host mRNA as predicted by RNA silencing.
782 *Plant Journal* **70**, 991-1003.
783 **Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC,**
784 **Grützner F, Kaessmann H. 2014.** The evolution of lncRNA repertoires and
785 expression patterns in tetrapods. *Nature* **505**, 635-640.
786 **Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D,**
787 **Showmaker KC, Shu S, Udall J. *et al*. 2012.** Repeated polyploidization of
788 *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**,
789 423-427.
790 **Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L,**
791 **Sandelin A. *et al*. 2012.** Systematic identification of long noncoding RNAs
792 expressed during zebrafish embryogenesis. *Genome Research* **22**, 577-591.

Plosky BS. 2014. An ultraconserved lnc to miRNA processing. *Molecular Cell* **55**, 3-4.

Quinn JJ, Ilik IA, Qu K, Georgiev P, Chu C, Akhtar A, Chang HY. 2014. Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nature Biotechnology* **32**, 933-940.

Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry* **81**, 145-166.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523.

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF. 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Molecular biology and evolution* **20**, 633-643.

Sturn A, Quackenbush J, Trajanoski Z. 2002. Genesis: Cluster analysis of microarray data. *Bioinformatics* **18**, 207-208.

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research* **41**, e166.

Swiezewski S, Liu F, Magusin A, Dean C. 2009. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* Polycomb target. *Nature* **462**, 799-802.

Tan JF, Tu LL, Deng FL, Hu HY, Nie YC, Zhang XL. 2013. A genetic and metabolic analysis revealed that cotton fiber cell development was retarded by flavonoid naringenin. *Plant Physiology* **162**, 86-95.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515.

Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, Xu J, Chua NH. 2014. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*. *Genome Research* **24**, 444-453.

826 **Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S.**
827 ***et al*. 2012.** The draft genome of a diploid cotton *Gossypium raimondii*.
828 *Nature Genetics* **44**, 1098-1103.
829 **Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B,**
830 **Guo H. *et al*. 2012.** MCScanX: a toolkit for detection and evolutionary
831 analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49.
832 **Wang ZM, Xue W, Dong CJ, Jin LG, Bian SM, Wang C, Wu XY, and Liu JY.**
833 **2012.** A comparative miRNAome analysis reveals seven fiber
834 initiation-related and 36 novel miRNAs in developing cotton ovules.
835 *Molecular Plant* **5**, 889-900.
836 **Wang ZW, Wu Z, Raitskin O, Sun Q, Dean C. 2014.** Antisense-mediated FLC
837 transcriptional repression requires the P-TEFb transcription elongation factor.
838 *Proceedings of the National Academy of Sciences, USA* **111**, 7468-7473.
839 **Wei MM, Wei HL, Wu M. *et al*. 2013.** Comparative expression profiling of miRNA
840 during anther development in genetic male sterile and wild type cotton. *BMC*
841 *Plant Biology* **13**, 66.
842 **Wendel JF., Brubaker CL, Seelanan T. 2010.** The origin and evolution of
843 *Gossypium*. 1-18.
844 **Xue W, Wang Z, Du M, Liu Y, Liu JY. 2013.** Genome-wide analysis of small
845 RNAs reveals eight fiber elongation-related and 257 novel microRNAs in
846 elongating cotton fiber cells. *BMC Genomics* **14**, 629.
847 **Yang X, Li L. 2011.** miRDeep-P: a computational tool for analyzing the microRNA
848 transcriptome in plants. *Bioinformatics* **27**, 2614-2615.
849 **Yang XY, Wang LC, Yuan DJ, Lindsey K, Zhang XL. 2013.** Small RNA and
850 degradome sequencing reveal complex miRNA regulation during cotton
851 somatic embryogenesis. *Journal of experimental botany* **64**, 1521-1536.
852 **Yoo MJ, Szadkowski E, Wendel JF. 2013.** Homoeolog expression bias and
853 expression level dominance in allopolyploid cotton. *Heredity* **110**, 171-180.
854 **Zemach A, McDaniel IE, Silva P, Zilberman D. 2010.** Genome-wide evolutionary
855 analysis of eukaryotic DNA methylation. *Science* **328**, 916-919.
856 **Zhang DY, Zhang TZ, Sang ZQ, Guo WZ. 2007.** Comparative development of lint
857 and fuzz using different cotton fiber-specific developmental mutants in
858 *Gossypium hirsutum*. *Journal of Integrative Plant Biology* **49**, 1038-1046.

Zhang YC, Yu Y, Wang CY, Li ZY, Liu Q, Xu J, Liao JY, Wang XJ, Qu LH, Chen F. *et al*. 2013. Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nature Biotechnology* **31**, 848-852.

Zheng Y, Li YF, Sunkar R, Zhang W. 2012. SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Research* **40**, e28.

Zhou L, Cheng X, Connolly BA, Dickman MJ, Hurd PJ, Hornby DP. 2002. Zebularine: A novel DNA methylation inhibitor that forms a covalent complex with DNA methyltransferases. *Journal of Molecular Biology* **321**, 591-599.

Zhou X, Sunkar R, Jin H, Zhu JK, Zhang W. 2009. Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*. *Genome Research* **19**, 70-78.

Figure Legends

Fig. 1 Identification and characterization of lncRNAs in *G. barbardense*. (a) The pipeline of long non-coding RNAs (lncRNAs) identification in *G. babardense*. (b) Exon number distribution per transcript of long intergenic non-coding RNAs (lincRNAs), long non-coding natural antisense transcripts (lncNATs) and protein-coding genes (PCgenes). (c) Length density distributions of lincRNAs, lncNATs and protein-coding transcripts. (d) The GC content of lincRNA, lncNAT and protein-coding transcripts in At (GbAt), Dt (GbDt) subgenomes and ungrouped (GbUn) scaffolds of *G. barbadense* genome. (e) The percentages of lincRNA and lncNAT transcripts overlapped with repetitive sequences in At, Dt subgenomes and ungrouped scaffolds. Trancripts with at least 10 bp overlapping regions with repetitive sequences are counted. (f) The percentage of total length of different repetitive sequences in all the lincRNA and lncNAT transcripts, which were compared with At, Dt subgenomes and ungrouped scaffolds.

Fig. 2 Expression of lncRNAs across 9 tissues or developmental stages. (a) The number of expressed lncRNA and protein-coding transcripts in each tissue or stage. The FPKM cutoff for determining expressed transcripts is 0.1 for lncRNAs and 0.5 for protein-coding transcripts. (b) Boxplot shows the distribution of maximum FPKM across samples in lincRNAs, lncNATs and protein-coding transcripts. (c) Pearson correlation coefficient distribution for homoeologous transcript pairs. The lincRNA-PCgene pairs and PCgene-PCgene pairs were restricted to adjacent 5 kb regions. (d) The distributions of maximal tissue specificity scores (JS score) calculated for lncRNA and protein-coding transcripts across all tissues. (e) Venn diagram shows the numbers of tissue-preferentially expressed transcripts in each tissues. The cutoff of maximum JS score per transcript is 0.5. (f) RT-PCR validation of tissue-preferentially expressed lincRNAs (LINC1 to LINC9).

Fig. 3 Evolution history and genomic landscape of lncRNAs. The homoeologous chromosomes are in the same colour. The grey lines show syntenic blocks and coloured lines show homoeologous lincRNA pairs between At and Dt subgenomes. (a) Pie chart showing the proportions of homologous lincRNAs in closely related species.

All the At subgenome lincRNAs in *G. barbadense* are aligned to Dt subgenome, *G. raimondii*, *G. arboretum*, *T. cacao* and *V. vinifera*. (b) GO enrichment of At subgenome specific lincRNAs. (c) GO enrichment of Dt subgenome specific lincRNAs. (d) Features of lncRNAs in At (green track) and Dt (red track) subgenomes of *G. barbadense*, (a) ratio of GC content in 500 kb windows, (b) percentage of repetitive sequences in 500 kb windows, (c) number of protein-coding genes in 500 kb windows, (d) number of lncRNA loci in 500 kb windows, (e) log2 ratio of averaged FPKM values for homoeologous lincRNA pairs (log2(At/Dt) >=1). The red dots show At-biased expression, green dots show Dt-biased expression and grey dots show equivalent expression. The right panel shows the categories of biased expression of homoeologous lincRNA pairs. The grey dashed lines shows the cutoff (log2(At/Dt) >=1 or log2(At/Dt) <=-1) for determining biased expression.

Fig. 4 Characterization of lncRNA methylation. (a) The DNA methylation and gene expression levels (lincRNAs and protein-coding genes) in *G. barbadense* (At subgenome green track, Dt subgenome red track). The homoeologous chromosomes are represented by the same color. Each chromosome is divided into 500 kb windows. The four track groups represent *G. arboretum* (a), *G. raimondii* (b), F1-hybrid between *G. arboretum* and *G. raimondii* (A2 x D5) (d) and natural tetraploid (e). For each track group, the CG methylation level, CHG methylation level, CHH methylation level, averaged lincRNA expression and averaged protein-coding gene expression are depicted outside-to-inside. The track c shows the TE density along each chromosomes. (b) DNA methylation in lincRNA and protein-coding gene regions. For each gene, the up-stream 1 kb, gene body and down-stream 1 kb are characterized and divided into 50 bins, respectively. (c) Correlations of the DNA methylation in CG, CHG and CHH contexts with gene expression. For each methylation context, the averaged DNA methylation levels of up-stream 1kb and gene body were plotted against the gene expression level. The accumulated frequency distribution of transcript numbers against DNA methylation level of lincRNAs and protein-coding genes are compared on the upper-right corner. The significant levels (p value) of distribution divergence are indicated. (d) Scatter-plot shows the differentially expressed lincRNAs and protein-coding genes between zebularine-treated ovule and controls. (e) The proportions of TE-contained

up-regulated and down-regulated lincRNAs after treated with zebularine are compared to that of all the lincRNAs in At and Dt subgenomes.

Fig. 5 Identification of lncRNAs associated with cotton fibre initiation. (a) The mature fibres or naked seeds of eight Upland cottons used in this study, including three lint-fuzz wild-type genotypes (TM-1, YZ1, XZ142), two lintless-fuzzless mutant genotypes (XZ142WX, XinWX) and three linted-fuzzless mutant genotypes (n2, GZnn, GZNn). (b, c) Heatmaps show the real-time PCR validation of expression of 20 lncRNAs at -1 DPA and 0 DPA ovules (b) and 4 DPA and 5 DPA ovules (c). The relative expression levels of each gene in different samples were normalized in the same data interval (-2 to 2) and visualized using Genesis (Sturn *et al*., 2002). (d) Real-time PCR validation of the differential expression of one lncRNA (LINC02) between lint-fuzz/linted-fuzzless cottons and lintless-fuzzless cottons at -1 and 0 DPA ovules (p-value < 0.05). (e) Real-time PCR validation of the differential expression of one lncRNA (LINC02) between lint-fuzz cottons and lintless-fuzzless/linted-fuzzless cottons at 4 and 5 DPA ovules (p-value < 0.05).

Fig. 6 Functional implications of lncRNAs in cotton fibre elongation and transition to secondary cell wall synthesis stages. (a) Clustering dendrogram of homeologous gene duplets between At and Dt subgenomes and assigned modules (labeling M1 to M17). These modules are constructed using gene expression data from10 DPA and 20 DPA cotton fibre transcriptomes. (b) Heatmaps of gene pairs expression in M12 (left) and M16 (right) combined with the normalized expression of hub genes. (c) Module network of M12. The lncRNA pairs and their involved co-expression relationships with protein-coding genes are colored in red. The protein-coding genes significantly enriched in organic cyclic compound metabolic process are colored in green and orthologs in *Arabidopsis* are annotated. (d) Module network of M16. The lncRNA pairs and their involved co-expression relationships with protein-coding genes are colored as M12. The protein-coding genes significantly enriched in oxidation-reduction process are colored in blue and small molecule metabolic process in cyan.

Fig. 7 Expression and functional analysis of lncRNAs generating miR397. (a) RNA-seq mapping of the lncRNAs pair generating miR397. The mature sequences of miR397 are labeled in red boxes. (b) Ratio of the clone sequences in the At and Dt subgenomes at 10 DPA and 20 DPA. (c) Real-time PCR of the total expression of lncRNA pair in the At and Dt subgenomes. Error bars show three biological replicates. (d) Comparison of the normalized expression of lncRNA pair in domesticated and wild *G. hirsutum* and *G. barbadense* accessions by RNA-seq. (e) Phylogenetic tree of *LAC4* in diploid A and D genomes, and the At and Dt subgenomes of *G. barbadense*. The *Arabidopsis LAC4* is regarded as an outgroup. Light red symbols show genes in diploid A genome (triangle) and the At subgenome (diamond), and light green symbols show genes in diploid D genome (square) and the Dt subgenome (round). The expression of each gene at 10 DPA and 20 DPA in diploid/tetraploid cottons is indicated. (f) Degradome sequencing shows the signature abundance in the position of *LAC4* (left *LAC4a*, right *LAC4b*) targeted by miR397. The red dot shows significant signature as indicated by red arrow. The target cleavage site is identified through RLM-RACE, as shown below the target plot. The numbers indicate the cleavage frequency through clone sequencing. (g) Sequence alignment of the upstream (left) and downstream (right) 3k regions between the Dt subgenome and diploid D genome by LASTZ software. (h) Model of the TE insertion from the At subgenome to the Dt subgenome in *G. barbadense*. TSS, transcription start site; TTS, transcription termination site.