Reciprocal Homoeologous Replacement of MYB Factors Drives the Evolutionary Trade-off between Yield and Fiber Quality in Cotton

YiFan Xu , Yanchao Xu , Zaoyang Gong , Qinglin Shen , Rongzheng Zhao , Chuangkang Cheng , Wanting Su , Yibing Li , Xingrui Yang , Xing Ruan , Fengyun Li , Dajun Liu , Kai Guo , Xueying Liu , Zhonghua Teng , Fang Liu , Zhengsheng Zhang , Dexin Liu.

Abstract

During the domestication of allotetraploid cotton, high-yielding *Gossypium hirsutum* (Upland cotton) and superior-quality *Gossypium barbadense* (Sea Island cotton) developed marked phenotypic divergence. Although the negative correlation between yield and fiber quality is well recognized, the genomic structural basis driving this differentiation and its cascading effects on transcriptional regulation remain largely elusive. Here, we integrated multi-tissue transcriptomics with pan-genome structural variation analysis to explore the genomic mechanisms underlying this phenotypic divergence. Transcriptomic profiling revealed that while the transcriptional programs of vegetative organs are highly conserved between the two species, significant spatiotemporal expression remodeling occurred during key windows of fiber development. Specifically, *G. barbadense* maintained a highly active transcriptional network during the transition from fiber elongation to thickening (20 DPA), exhibiting unique "delayed elongation" transcriptional features. Further pan-genomic analysis identified a pair of key transcription factors—*MYB111* (A12) and *MYB93* (A11)—that underwent "Reciprocal Homoeologous Replacement" (RHR). Structural variation evidence indicates that *G. hirsutum* lost the elongation-promoting *MYB111* via a physical deletion on chromosome A12, potentially leading to the premature termination of the elongation program. As an evolutionary trade-off, *G. hirsutum* specifically retained the root development-related *MYB93* on chromosome A11, suggesting a potential strategy to optimize root system architecture for broad environmental adaptation. We propose that these genomic structural variations not only reshaped tissue-specific transcriptional networks ("software") but also serve as the fundamental genetic basis ("hardware") driving the differentiation between "high yield" and "superior quality" in cotton. These findings provide new insights into trait evolution in polyploid crops and offer valuable genomic targets for breaking genetic bottlenecks in cotton breeding.

Introduction

Cotton stands as a global cornerstone of the textile industry, serving as the premier source of renewable natural fiber. Among cultivated cottons, two allotetraploid species dominate: *Gossypium hirsutum* (Upland cotton) and *G. barbadense* (Sea Island cotton) (Wen et al. 2023). Upland cotton, represented by the standard genetic line TM-1, accounts for over 90% of global fiber production due to its exceptional yield potential and broad environmental adaptability; however, its fiber quality traits—such as length, strength, and fineness—remain moderate (Avci et al. 2013). In stark contrast, Sea Island cotton (e.g., H7124) is renowned for its superior "Extra Long Staple" fibers but is hampered by low yield and stringent ecological requirements (Alagarsamy 2023). This marked phenotypic divergence between "high yield" and "superior quality" reflects the distinct selection pressures imposed during their long-term history of "Divergent Domestication". Consequently, decoupling this genetic negative correlation remains a

central challenge in modern cotton breeding (Liu et al. 2024).

The rapid advancement of high-throughput sequencing has ushered in a new era for cotton genomics. Since the release of the first draft genome of Upland cotton in 2015 (Li et al. 2015; Zhang et al. 2015) , high-quality reference genomes for both *G. hirsutum* (TM-1) and *G. barbadense* (H7124, 3-79) have been successively published (Yuan et al. 2015; Hu et al. 2019). More recently, the completion of "Telomere-to-Telomere" (T2T) assemblies (Hu et al. 2025a; Hu et al. 2025b) has provided unprecedented resolution for dissecting complex allotetraploid genomes. These genomic resources have unveiled large-scale chromosomal rearrangements (such as inversions and translocations) and transposon amplification events associated with polyploidization (Wang et al. 2022; Peng et al. 2022). However, current comparative genomic studies have predominantly focused on global synteny or single-nucleotide polymorphisms (SNPs). There remains a lack of systematic understanding regarding how "Gene Presence/Absence Variations" (PAVs)—genes lost or gained during domestication—reshape downstream tissue-specific transcriptional networks, particularly those governing fiber development.

Cross-species comparative transcriptomics has emerged as a powerful tool for dissecting the evolutionary mechanisms of complex traits. Yao et al. (2022) , through a systematic comparison of large-scale multi-tissue transcriptomes in humans and cattle, revealed the evolutionary conservation and divergence of tissue-specific regulatory networks, demonstrating the critical role of multi-tissue atlases in resolving the genetic basis of complex traits. Yet, such a holistic strategy remains rarely applied to polyploid crop domestication. Although extensive pan-genomic studies in cotton have identified numerous structural variations linked to agronomic traits (Meng et al. 2025; Wang et al. 2022) , existing comparative transcriptomic efforts are largely confined to single tissues—such as fibers (Chen et al. 2012) or leaves (Cheng et al. 2025) —or specific developmental stages, lacking a systematic analysis of co-evolution patterns at the whole-plant level. Unlike diploid mammals, the domestication of allotetraploid cotton involves complex subgenome interactions and homoeologous chromosome differentiation. Therefore, integrating spatiotemporal multi-tissue transcriptomics with pan-genome structural variations is essential to reveal how genomic "hardware" drives the remodeling of transcriptomic "software".

Within the complex transcriptional regulatory networks of plants, the MYB transcription factor superfamily occupies a central position due to its vast membership and functional diversity (Dubos et al. 2010; Zhang et al. 2025a). As one of the largest transcription factor families in the plant kingdom, the R2R3-MYB subfamily has been extensively implicated in regulating secondary metabolism (e.g., flavonoid biosynthesis) and epidermal cell differentiation (Stracke et al. 2007; Tan et al. 2013). In cotton, MYB factors are regarded as pivotal regulatory hubs governing fiber development. For instance, *GhMYB25* and *GhMYB25-like* are critical for fiber initiation (Machado et al. 2009; Walford et al. 2011), while other members orchestrate fiber elongation and secondary cell wall thickening (Li et al. 2016; Zhang et al. 2025b). Despite functional characterizations of specific MYB genes, it remains largely unexplored whether this superfamily underwent systematic copy number variations or asymmetric losses during the divergent domestication of Upland and Sea Island cotton, and how such genomic alterations drove the phenotypic differentiation between "yield" and "quality."

Despite decades of research, the precise molecular basis driving this phenotypic divergence between "high yield" and "superior quality" remains incompletely understood (Zhang et al. 2025a). Existing studies have predominantly relied on single nucleotide polymorphism (SNP)-based genome-wide association studies (GWAS) and QTL mapping. However, these approaches often encounter the "missing heritability" dilemma when explaining complex quantitative traits (Manolio et al. 2009), as they struggle to capture large-scale genomic structural variations (SVs), such as presence/absence variations (PAVs) and large-segment inversions (Pinglay et al. 2025). Particularly within allopolyploid genomes, structural variations often serve as a primary engine driving subgenome differentiation and adaptive evolution. These "hard" genomic alterations can lead to the definitive gain or loss of key regulators, thereby establishing insurmountable "trait barriers" between species. Currently, a critical knowledge gap remains: whether and how specific chromosomal structural variations—especially asymmetric variations between homoeologous chromosomes — reshape downstream transcriptional regulatory networks, thereby determining the divergent evolutionary trajectories of Upland and Sea Island cotton in terms of "yield" versus "quality" (Wendel 2000).

To address this question, this study integrated multi-tissue, whole-growth-period transcriptomic atlases of *G. hirsutum* (TM-1) and *G. barbadense* (H7124) with pan-genome structural variation analysis to systematically dissect their molecular evolutionary logic during domestication. We aim to bridge genomic "hardware" variations (PAVs) with transcriptomic "software" remodeling (spatiotemporal expression differences) to uncover the "Reciprocal Homoeologous Replacement" (RHR) mechanism driving the differentiation of yield and quality in cotton. This work provides a theoretical basis and precise genomic targets for breeding new cotton varieties that combine high yield with superior quality.

Materials and Methods

Plant Materials and Data Acquisition

In this study, we utilized the Telomere-to-Telomere (T2T) reference genome of Gossypium hirsutum (TM-1) released by Yan et al. (2025) and the high-quality reference genome of G. barbadense (H7124) sequenced by Hu et al. (2019) . The genomic data for G. barbadense were retrieved from the CottonMD database (Yang et al. 2023). Furthermore, to facilitate pan-genomic evolutionary analysis, chromosome-level genome assemblies for 26 Gossypium species, comprising both wild diploids and tetraploids, were collected from CottonGen (Yu et al. 2021) and the NCBI database (Supplementary Table S1).

Transcriptomic Data Acquisition and Preprocessing

The gene expression data employed in this study were retrieved from the CottonMD database (Yang et al. 2023). The raw high-throughput sequencing data originated from the genome sequencing project of *G. hirsutum* (TM-1) and *G. barbadense* (H7124) conducted by Hu et al. (2019) (BioProject: PRJNA490626). We obtained normalized gene expression matrices (TPM values) and selected a subset of data covering 21 representative tissues/stages, including roots, stems, leaves, floral organs, and developing ovules and fibers at various stages (Supplementary Table S2). To enhance the reliability of the analysis and mitigate background noise, genes with consistently low expression levels (TPM < 0.1) across all samples were filtered out prior to

subsequent analyses.

Gene Expression Characterization and Comparative Transcriptomics

To evaluate global expression patterns across samples, we performed dimensionality reduction and clustering analyses based on $\log_2$(TPM+1) transformed expression matrices. Principal Component Analysis (PCA) was visualized using the ggplot2 package (Wickham 2016) in R (R Core Team 2025). Additionally, t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) were executed using the Rtsne and umap packages, respectively.

To quantify transcriptional conservation between species, expressed genes were screened using a threshold of TPM≥0.1. The Pearson correlation coefficient (R) for the number of expressed genes between corresponding tissues of TM-1 and H7124 was calculated using the cor.test function in R. Furthermore, to dissect divergence in expression abundance, expressed genes in each tissue were ranked by TPM values and categorized into five expression windows (Very Low, Low, Medium, High, and Very High). The number of "shared genes" falling within the same expression window in homologous tissues of both species was quantified using a custom Python script. Finally, to account for differences in total gene counts, we calculated the relative proportions of common genes versus species-specific genes within each tissue.

For tissue specificity analysis, the Tissue Specificity Index (Tau) was calculated following the method described by Yanai et al. (2005) . A Tau value approaching 1 indicates stronger tissue specificity. We defined specific genes as those with a Specificity Measure (SPM) >0.5. The intersections of specific genes across different tissues were visualized using UpSet plots generated with the R package UpSetR (Conway et al. 2017).

Weighted Gene Co-expression Network Construction and Unbiased Functional Annotation

To mine key co-expression gene modules driving the delayed fiber elongation in Sea Island cotton (Avci et al. 2013) , we constructed a weighted gene co-expression network using the WGCNA package in R (Langfelder and Horvath 2008). To optimize computational efficiency and minimize noise, we sorted genes by variance and selected the top 6,000 genes exhibiting the highest variation in fiber development samples as input data. First, a Pearson correlation matrix was calculated, and a scale-free network was constructed based on a soft-thresholding power. Subsequently, the adjacency matrix was converted into a Topological Overlap Matrix (TOM), and co-expression modules were identified using the Dynamic Tree Cut algorithm. Module expression patterns were visualized via Module Eigengenes (ME) to identify target modules specifically highly expressed in G. barbadense fibers at 20 DPA.For unbiased functional dissection of the target module, we extracted the protein sequences of all genes within the module and performed a comprehensive scan against the Pfam-A database (Mistry et al. 2021) using the hmmscan program in HMMER 3.0 (Eddy 2011). The screening threshold was set to an E-value < 1e−5. Based on the scan results, we cataloged the domain types for each gene and ranked gene families by domain frequency to determine the most enriched category of transcriptional regulators within the module.

Identification of the MYB Gene Family

Genome-wide identification of the MYB gene family in TM-1 and H7124 was performed using the Bitacora pipeline (V1.4.2) (Vizueta et al. 2020). Briefly, this pipeline integrates Hidden Markov

Model (HMM) searches based on HMMER (Eddy 2011) with homology-based alignment using BLASTP (Camacho et al. 2009). The HMM seed file for the MYB domain (PF00249) was retrieved from the Pfam database (Mistry et al. 2021).

Identification and Rigorous Screening of Presence/Absence Variations (PAVs)

Species-specific genes (PAVs) were identified based on a whole-genome alignment strategy using the NUCmer program within the MUMmer4 package (Marçais et al. 2018). NUCmer parameters were set to --maxmatch -c 1000 -l 100. The alignment results were filtered using delta-filter (parameters: -m -i 90 -l 100) to retain high-confidence alignment chains. Subsequently, to further minimize the false discovery rate, candidate sequences were subjected to a secondary filtration using Minimap2 (Li 2018) with parameters -c -x asm10, eliminating sequences that still retained high similarity to the reference genome.

To rigorously exclude false-positive deletions resulting from assembly gaps or sequence fragmentation, we implemented a strict "Two-Step Validation" strategy:

(1) Sequence Extraction: The Coding DNA Sequences (CDS) of candidate species-specific genes from H7124 or TM-1 were extracted.

(2) Reverse BLAST Validation: The candidate sequences were aligned back to the counterpart genome using BLASTN (Camacho et al. 2009) with an E-value threshold of < 1e–5. If a matching fragment with Identity > 95% and Coverage > 80% was detected in the counterpart genome, the gene was designated as a false positive and discarded. Only genes with no valid matches were retained as high-confidence species-specific PAVs.

Genomic Synteny and Structural Variation Analysis

Whole-genome alignments between TM-1 and H7124 were generated using the NUCmerprogram in MUMmer4 (Marçais et al. 2018) with parameters --maxmatch -l 100 -c 1000. The best-hit alignment chains were retrieved via delta-filter (parameters: -m -i 90 -l 100) for structural variation analysis. Based on these alignments, structural variations (SVs) were systematically identified using SyRI (Synteny and Rearrangement Identifier) (Goel et al. 2019). The identified SVs encompassed syntenic regions, inversions, translocations, and local variations, including non-aligned regions (NOTAL) and highly diverged regions (HDR). Specifically, we interrogated the genomic classifications of the *MYB93* and *MYB111* loci to determine whether they corresponded to physical deletions (NOTAL) or sequence collapse due to high divergence (HDR). The final synteny and structural variation landscapes were visualized using the Python tool plotsr (Goel and Schneeberger 2022).

Pan-genome Homology Search and Evolutionary Trajectory Analysis

To trace the retention and loss history of the target genes (*MYB93* and *MYB111*) throughout *Gossypium* evolution, we utilized their CDS sequences as queries to perform TBLASTN and BLASTN searches against 26 representative *Gossypium* genomes using the BLAST+ package (Camacho et al. 2009). The retention threshold was set to an E-value . Best-hit loci in each species were extracted using a custom Python script to calculate sequence identity and coverage. Finally, the distribution patterns across different evolutionary clades (wild diploids, wild tetraploids, and cultivated species) were visualized as bubble plots using the ggplot2 package (Wickham 2016) in R (R Core Team 2025).

结果

Global Transcriptomic Landscape of G. hirsutum and G. barbadense Across Multiple Tissues

To systematically dissect the genome-wide transcriptional regulatory landscape of *G. hirsutum*(TM-1) and *G. barbadense* (H7124), we performed dimensionality reduction and clustering analyses based on the expression matrices of genome-wide orthologous genes using three distinct algorithms: PCA, t-SNE, and UMAP (Fig. 1). All three algorithms revealed a highly consistent sample topology, underscoring the robustness of the intrinsic data structure.

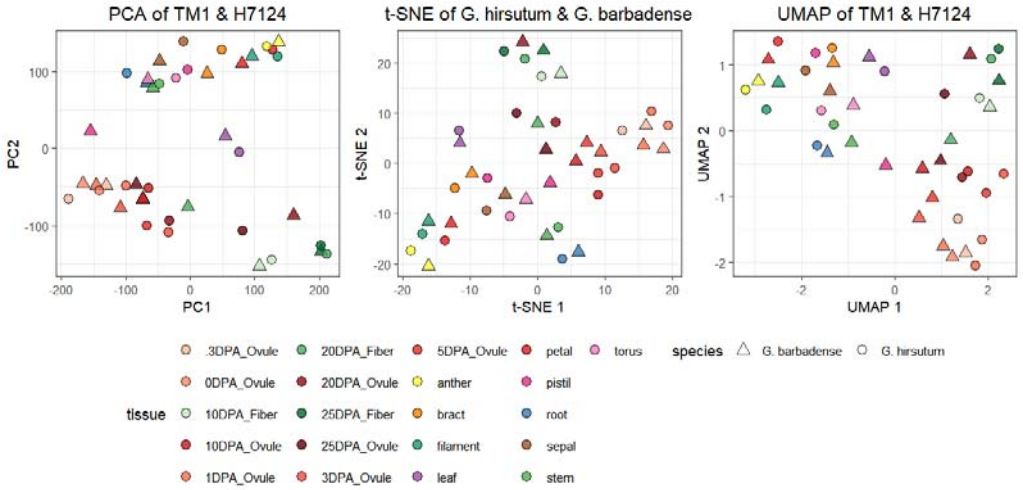

Fig. 1 Global transcriptomic landscape of *G. hirsutum* (TM-1) and *G. barbadense* (H7124) across multiple tissues. Dimension reduction analyses of gene expression profiles using (a)Principal Component Analysis (PCA), (b) t-Distributed Stochastic Neighbor Embedding (t-SNE), and (c) Uniform Manifold Approximation and Projection (UMAP). Each point represents a sample. Shapes denote the species: circles for *G. hirsutum* (TM-1) and triangles for *G. barbadense*(H7124). Colors denote the different tissues or developmental stages as indicated in the legend below. Samples from the same tissue cluster tightly together regardless of species, indicating that tissue specificity is the dominant factor driving transcriptomic variance.

The analysis demonstrated that samples clustered predominantly by tissue identity rather than species origin. Specifically, *G. hirsutum* and *G. barbadense* samples from the same tissue (e.g., leaf) clustered tightly together in the low-dimensional space, forming distinct clusters that were clearly separated from other tissues (e.g., root or petal). Clear boundaries were observed between vegetative organs (root, stem, leaf) and reproductive organs (flower, ovule, fiber). Furthermore, developmentally related tissues, such as ovules and fibers at various stages, exhibited continuous trajectory distributions.

Notably, despite the substantial divergence in morphology and agronomic traits between the two species, we did not observe overt species-specific separation at the whole-genome level. This global expression landscape conveys two critical insights. First, the evolutionary constraints governing tissue development are significantly stronger than species divergence. This implies that *G. hirsutum* and *G. barbadense* share a highly conserved "Transcriptional Chassis" to maintain fundamental physiological functions, such as photosynthesis and nutrient absorption. Second, it suggests that the distinct phenotypic differentiation between "yield" and "quality" does not stem from systemic genomic remodeling. Instead, it is likely driven by the spatiotemporal "fine-tuning"

or "modular rewiring" of a limited number of key regulatory networks. Indeed, this high similarity in the macroscopic background underscores the biological significance of the fiber-specific structural variations (e.g., the reciprocal replacement of MYB factors) that we identified as being concealed within this conserved landscape.

Conservation and Specificity of Expressed Gene Counts Between G. hirsutum and G.barbadense

To investigate interspecific differences in transcriptomic breadth, we defined "expressed genes" as transcripts with a TPM threshold of . Based on this criterion, we compared the number of expressed genes across 21 homologous tissues between TM-1 and H7124 (Fig.2).

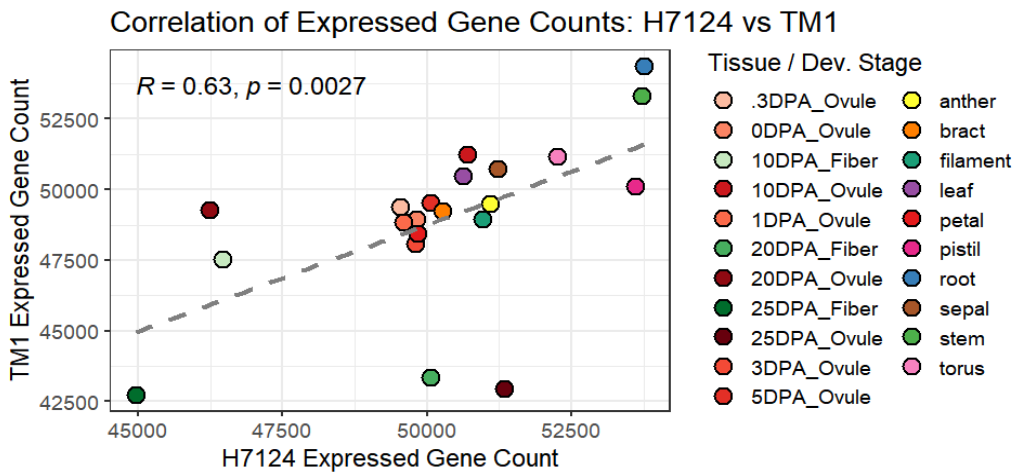

Fig. 2 Correlation of expressed gene counts between *G. hirsutum* (TM-1) and *G. barbadense* (H7124) across 21 tissues. The scatter plot compares the number of expressed genes (TPM≥0.1) in homologous tissues of the two species. The dashed line represents the linear regression fit. Pearson's correlation coefficient () and -value are indicated. Points deviating from the diagonal indicate tissues with species-specific expansion or contraction of the active transcriptome (e.g., Root/Anther for TM-1; 20 DPA Fiber for H7124).

Pearson correlation analysis revealed a significant moderate positive correlation (R=0.63, p=0.0027) between the two species. This indicates that despite extensive divergent domestication, the transcriptional activity across most tissues retains a degree of evolutionary conservation. Specifically, sample points representing vegetative organs (root, stem, leaf) and floral parts (petal, sepal, etc.) clustered tightly around the regression line, suggesting that these foundational tissues mobilize comparable genomic resources in both species.

However, we observed marked species-biased dispersion in specific tissues. In roots and anthers, the number of expressed genes in *G. hirsutum* (TM-1) was significantly higher than in *G. barbadense*. Conversely, during the fiber elongation-to-thickening transition (20 DPA Fiber) and late ovule development (25 DPA Ovule), *G. barbadense* (H7124) exhibited a greater abundance of expressed genes. This reciprocal shift in gene repertoire size underscores a significant divergence in the complexity of transcriptional regulation during specific developmental windows.

Interspecific Conservation and Tissue-Specific Divergence in Gene Expression Abundance

To further dissect the regulatory characteristics of abundance underlying the correlation in gene counts, we compared the distribution of gene expression levels across tissues between H7124 and TM-1 (Fig. 3). The results showed that the expression abundance distributions (violin shapes

and medians) in vegetative organs (root, stem, leaf) and floral organs (anther, petal, etc.) exhibited high consistency between the two species1. This indicates that the transcriptional homeostasis maintaining fundamental physiological functions is highly evolutionarily conserved. However, significant interspecific differences were observed during fiber and ovule development. Notably, during the fiber development stage from 10 to 25 DPA, *G. barbadense* (H7124) displayed a broader high-expression interval ($\log_2$TPM > 5), suggesting the mobilization of a more abundant transcriptional regulatory network during this period2. Furthermore, a comparison of synchronous tissues revealed that the overall transcriptional abundance in 10 DPA fibers was significantly higher than that in 10 DPA ovules, reflecting a metabolic activity in fiber cells during rapid elongation that far exceeds that of the corresponding ovule tissues.

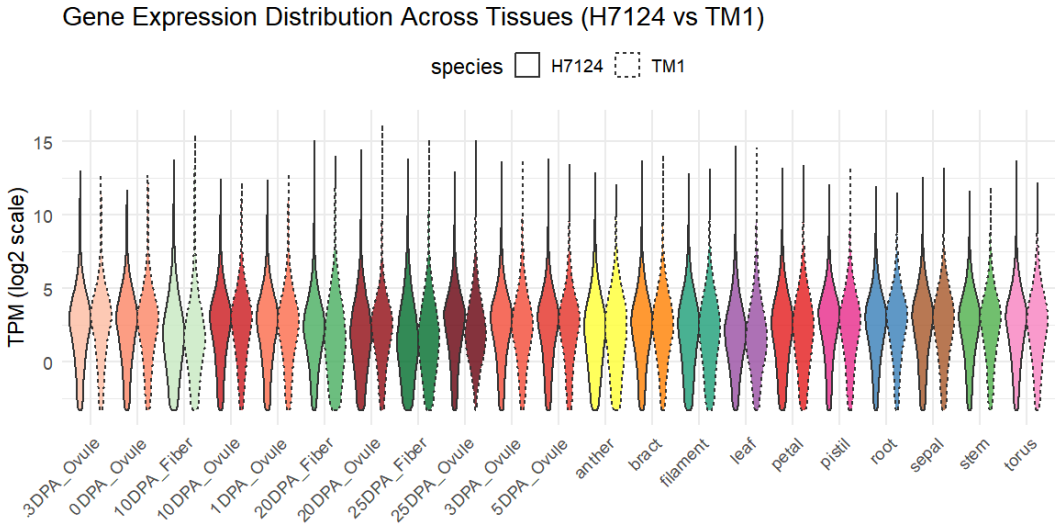

Fig. 3 Global distribution of gene expression abundance across tissues. Violin plots showing the density distribution of gene expression levels [$\log_2$(TPM + 1)] for *G. barbadense* (H7124, solid outline) and *G. hirsutum* (TM-1, dashed outline). The width of the violin represents the density of genes at different expression levels. While vegetative and floral tissues show overlapping distributions, distinct shifts in high-abundance expression are observed in developing fibers (e.g., 10–25 DPA).

Further, to quantify the degree of conservation in these expression patterns, we stratifiedgenes into five abundance tiers (Very Low to Very High) and enumerated the number of shared genes within the same expression window between the two species (Fig.4).

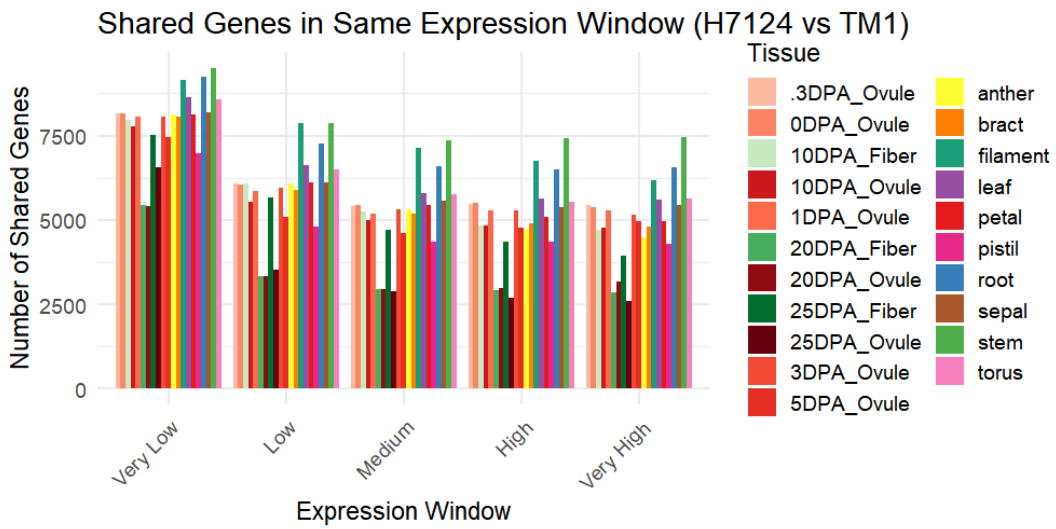

Fig. 4 Conservation of gene expression strata between species. Bar charts displaying the number of shared genes between H7124 and TM-1 within five defined expression windows (Very Low to Very High). Genes were grouped based on their TPM ranking in each tissue. Note the consistently high number of shared genes in the "Very Low" window across all tissues, contrasting with the reduced number of shared genes in the active expression windows (Medium to High) for developing fibers and late-stage ovules compared to vegetative tissues (e.g., leaf, root).

The analysis revealed two distinct features. First, the "Very Low" expression window exhibited the strongest conservation. Across all tissues, the number of shared genes in this window was the highest (mean > 7,500), indicating that regulatory mechanisms governing "gene silencing" or basal transcriptional maintenance are the most conserved between the two species. Second, transcriptional divergence intensified in fibers and ovules. Compared to vegetative organs such as leaves and roots, late-stage developing ovules (20–25 DPA) and fibers (10–25 DPA) showed a systemic reduction in the number of shared genes within the "Medium" to "Very High" expression windows. This result suggests that, in contrast to conserved housekeeping genes, the transcriptional regulatory programs governing the late stages of fiber and ovule development have undergone significant species-specific remodeling during domestication.

Distribution Characteristics and Species-Specific Bias of Tissue-Specific Genes

To systematically evaluate the breadth and specificity of gene expression, we calculated the Tissue Specificity Index (Tau) for all expressed genes genome-wide (Fig. 5). The Tau values for both *G. hirsutum* (TM-1) and *G. barbadense* (H7124) exhibited a typical bimodal distribution pattern1111. The primary peaks were located in the regions of Tau < 0.25 (broadly expressed housekeeping genes) and Tau > 0.90 (highly tissue-specific genes). This highly consistent distribution morphology confirms that Gossypium species maintain a similar functional partitioning landscape at the genomic level.

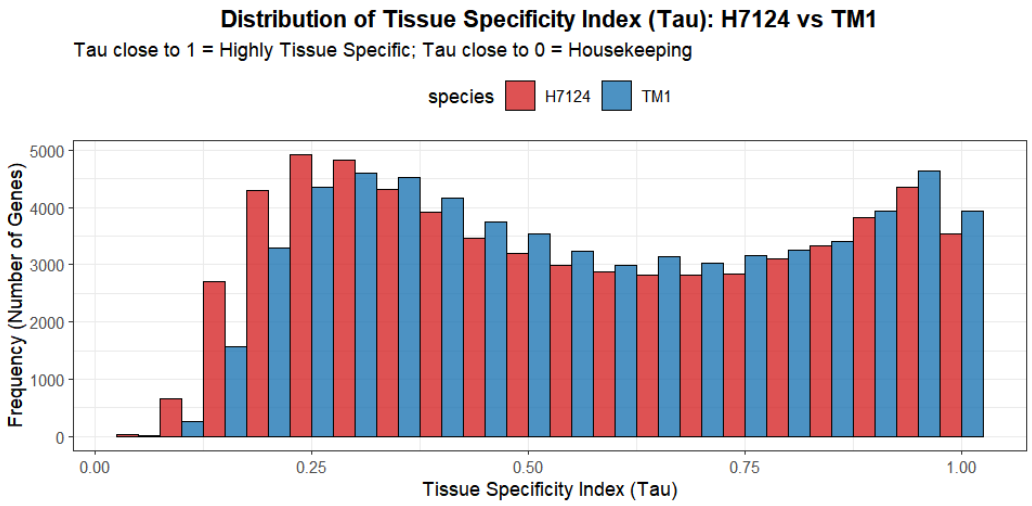

Fig. 5 Distribution of Tissue Specificity Index (Tau) in *G. barbadense* and *G. hirsutum*.Frequency histograms showing the bimodal distribution of Tau values for H7124 (red) and TM-1 (blue). Tau values range from 0 (broadly expressed housekeeping genes) to 1 (highly tissue-specific genes). While the overall patterns are similar, TM-1 shows an expansion in the number of tissue-specific genes (Tau > 0.5), whereas H7124 retains a higher proportion of broadly expressed genes (Tau < 0.4).

However, we observed significant species bias in the two extreme regions. In the specific expression interval (Tau>0.5), the gene frequency in TM-1 was generally higher than in H7124, indicating that the Upland cotton genome contains a larger repertoire of specific genes executing specialized functions in single organs. Conversely, in the broad expression interval (Tau<0.4), the gene frequency in H7124 was significantly higher than in TM-1, suggesting that Sea Island cotton retains a higher proportion of broadly expressed genes associated with basal metabolism.

Further correlation analysis revealed a robust positive correlation between the Tau indices of orthologous genes in H7124 and TM-1 (Fig. 6). The vast majority of genes clustered near the diagonal, indicating that expression breadth has maintained high evolutionary conservation following speciation. Nonetheless, we identified a subset of "outlier genes" deviating from the diagonal. For instance, certain genes manifested as highly specific (Tau > 0.8) in H7124, while transitioning to broad expression (Tau < 0.3) in TM-1, and vice versa. These loci exhibiting drastic shifts in specificity imply local functional remodeling of transcriptional regulatory networks between the species.

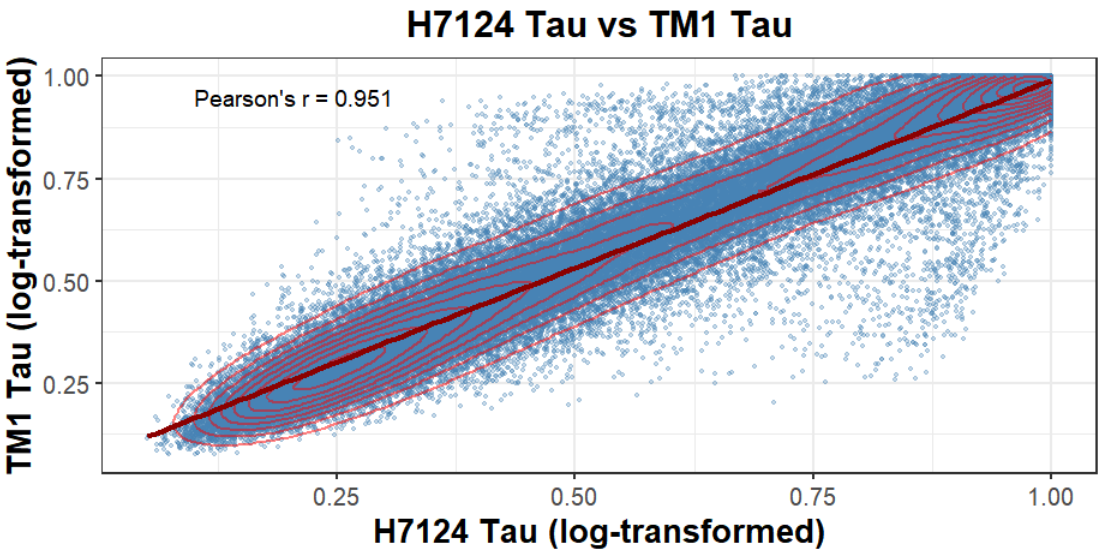

Fig. 6 Correlation of Tau indices between orthologous genes. Scatter plot comparing the Tau index of each orthologous gene pair between H7124 (x-axis) and TM-1 (y-axis). The red line represents the linear regression fit. The high Pearson correlation coefficient (*r*) indicates global conservation of tissue specificity. Points deviating from the diagonal (Outliers) represent genes that have undergone divergent regulation of expression breadth between the two species.

Divergent Patterns of Tissue-Specific Gene Repertoires

To further dissect the distributional architecture of tissue-specific genes across distinct developmental modules, we visualized the identified specific genes (SPM>0.5) using UpSet plots (Figs.7and 8).

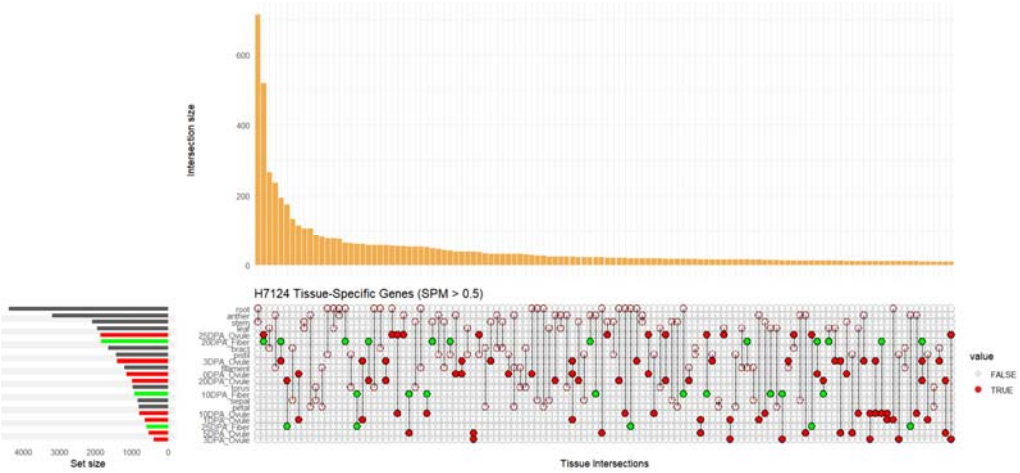

Fig. 7 Intersection of tissue-specific genes in *G. barbadense* (H7124). UpSet plot visualizing the number of tissue-specific genes (SPM > 0.5) across 21 tissues in H7124. The horizontal bars (left) represent the total number of specific genes in each tissue, while the vertical bars (top) show the size of intersection sets. Note the substantial number of specific genes retained in late-stage fibers (e.g., 20 DPA, 25 DPA) and ovules compared to TM-1.

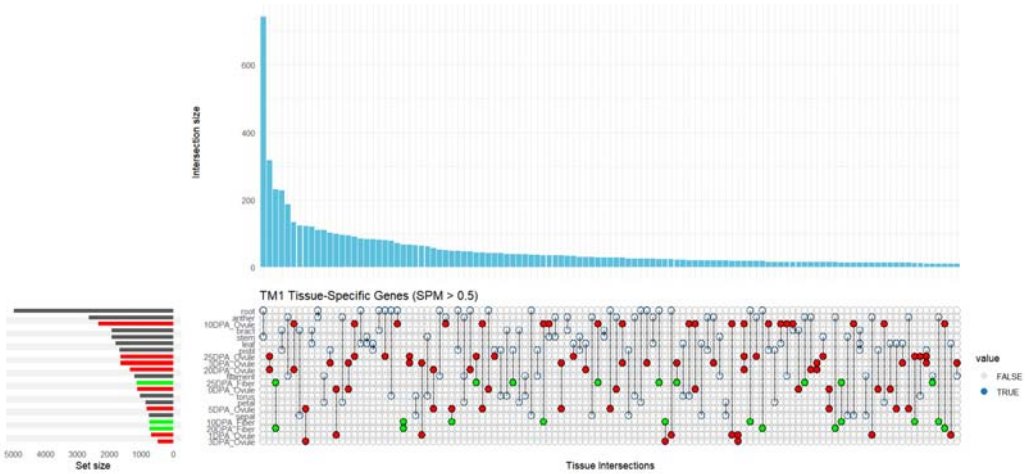

Fig. 8 Intersection of tissue-specific genes in *G. hirsutum* (TM-1). UpSet plot for tissue-specific genes in TM-1. In contrast to H7124, TM-1 exhibits a predominant expansion of specific genes in roots and anthers, while the number of specific genes in developing fibers (e.g., 10–25 DPA) is notably reduced. This distinct distribution pattern highlights the divergence in tissue-specific transcriptional regulation between the two species.

In *G. barbadense* (H7124) (Fig. 7), while roots and anthers retained a substantial number of specific genes, the most prominent feature was the sustained retention of specific genes during the late stages of fiber and ovule development. Specifically, H7124 maintained a massive specific gene repertoire during key developmental windows—such as 20 DPA Fiber, 25 DPA Fiber, and 25 DPA Ovule—where the count of specific genes was significantly higher than in the synchronous tissues of *G. hirsutum*. This indicates that *G. barbadense* sustains a more active and specialized transcriptional regulatory program during the critical transition from fiber elongation to secondary cell wall thickening, as well as during ovule maturation.

In sharp contrast, *G. hirsutum* (TM-1) exhibited a distinctly different distribution pattern (Fig. 8). A burst-like expansion in the number of root-specific genes constituted the most striking feature of its profile. Conversely, the specific gene repertoire in the mid-to-late stages of fiber development (10–25 DPA Fiber) underwent a marked contraction. This interspecific disparity reveals a directional differentiation of tissue-specific transcriptomes: *G. barbadense* tends to preserve high-complexity specific regulation in tissues governing fiber and ovule quality, whereas *G. hirsutum* has undergone significant expansion of specific gene families in tissues associated with nutrient acquisition (roots) and male reproduction (anthers).

Cross-Species Conservation and Spatiotemporal Expression Bias of Tissue-Specific Genes

While the aforementioned UpSet plots revealed absolute disparities in gene counts across specific tissues, to normalize against fluctuations in total gene numbers between tissues, we further calculated the relative proportions of "Common" versus "Species-specific" genes within each tissue (Fig. 9). This normalized analysis clearly delineates the boundaries between evolutionary conservation and divergence within the functional landscape of the cotton genome.

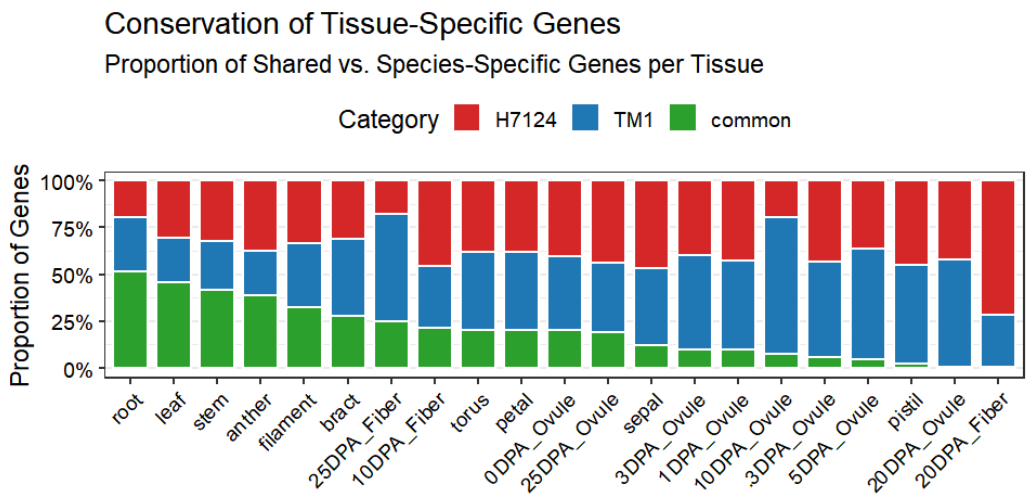

Fig. 9 Proportion of shared versus species-specific genes across tissues. Stacked bar chart showing the relative composition of tissue-specific genes. Genes are categorized as Common(green, shared by both species), H7124-specific (red), or TM-1-specific (blue). Note the high proportion of Common genes in vegetative tissues (e.g., root, leaf) indicating evolutionary conservation. In contrast, developing fibers exhibit high divergence with a distinct temporal shift: H7124-specific genes dominate the elongation-to-thickening transition phase (20 DPA), whereas TM-1-specific genes become predominant during the late secondary wall synthesis phase (25 DPA).

First, vegetative organs (root, leaf, stem) and certain floral organs (anther, petal, filament) exhibited the highest degree of evolutionary conservation. In these tissues, the proportion of common specific genes (green bars) was generally sustained between 30% and 50%, indicating that the core transcriptional programs maintaining the fundamental framework of plant growth and reproduction remain relatively stable between the two species.

In stark contrast, the proportion of common genes in developing fiber and ovule tissues was significantly reduced (generally <25%). Notably, the pistil, despite being a floral organ, also exhibited extremely low conservation (<5%), implying that reproductive development and fiber differentiation have undergone drastic transcriptional remodeling.

Crucially, we observed a significant phenomenon of "Spatiotemporal Inversion" regarding species bias during the progression of fiber development. During the transition phase from fiber elongation to secondary cell wall synthesis (10 DPA and 20 DPA Fiber), *G. barbadense* (H7124)-specific genes occupied a position of absolute dominance. Specifically at 20 DPA, the proportion of H7124-specific genes (red bars) reached approximately 70%, while the proportion of common genes was compressed to less than 5%. However, upon entering the phase of rapid secondary cell wall synthesis (25 DPA Fiber), this trend was reversed: the proportion of G. hirsutum (TM-1)-specific genes (blue bars) increased significantly, surpassing *G. barbadense* and assuming dominance.

This dynamic transition from "Sea Island-dominated mid-stage" to "Upland-dominated late-stage" (Fig. 9) reveals fundamental regulatory disparities in the fiber development programs of the two species. Given that the transcriptomes of vegetative organs are highly conserved, this

drastic spatiotemporal expression remodeling during fiber development must be driven by specific, tightly regulated upstream gene networks. However, identifying the core regulators responsible for this differentiation based solely on macroscopic transcriptomic comparisons is challenging. To deconstruct the key functional modules driving "delayed elongation in Sea Island cotton" from complex genome-wide expression data and to uncover the underlying core drivers in an unbiased manner, we subsequently employed Weighted Gene Co-expression Network Analysis (WGCNA).

WGCNA Unveils Species-Specific Regulatory Modules Driving Delayed Fiber Elongation

To systematically dissect the gene regulatory networks underlying the spatiotemporal transcriptomic remodeling observed during fiber development (Fig. 9), we constructed a weighted gene co-expression network (WGCNA) using the top 6,000 genes exhibiting the highest transcriptional variance in fiber tissues of *G. barbadense* (H7124) and *G. hirsutum* (TM-1). The analysis identified multiple co-expression modules characterized by distinct spatiotemporal expression patterns (Fig. 10a).

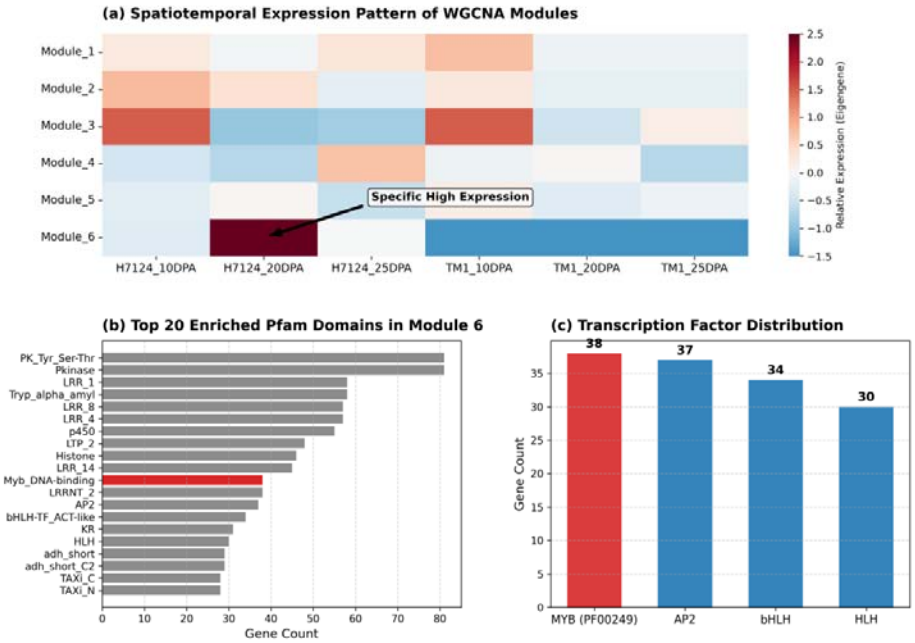

Fig. 10 Identification of the key regulatory module for fiber elongation. (a) Heatmap of module eigengenes identified by WGCNA. The color scale represents the relative expression level (red: high; blue: low). Module 6 (arrow) shows specific high expression in H7124 at 20 DPA. (b) Top 20 enriched Pfam domains within Module 6 genes. The module is rich in signaling components (Pkinase, LRR) and transcriptional regulators. (c) Distribution of transcription factor families within Module 6. The MYB family (containing PF00249) represents the most abundant class of transcriptional regulators.

Among these modules, Module 6 exhibited critical features highly correlated with the "delayed elongation" phenotype. The expression profile of the module eigengene for Module 6 displayed a striking species-specific pattern: it specifically maintained robust transcriptional activity in *G.*

barbadense (H7124) fibers at 20 DPA, a stage when fiber elongation in TM-1 typically nears cessation (Fig. 10a). In contrast, this module appeared to be markedly suppressed in the corresponding tissues of TM-1. This "Sea Island-present/Upland-absent" expression pattern suggests that Module 6 represents the core transcriptional machinery responsible for extending the developmental window of fiber elongation in *G. barbadense*.

To dissect the functional architecture of this module without a priori assumptions, we performed an unbiased genome-wide domain scan against the Pfam database for all genes within Module 6 (n=3151). Domain abundance analysis revealed a bi-level functional organization within the module (Fig. 10b). At the effector level, the module was enriched with numerous genes involved in signal transduction and metabolism, characterized by a high abundance of Protein Kinase (Pkinase, 81 genes), Leucine-Rich Repeat (LRR, 57 genes), and Cytochrome P450 (p450, 55 genes) domains. This indicates that Module 6 orchestrates an active metabolic state and a sustained signal perception network, serving as the physiological prerequisite for continuous cell wall biosynthesis.

Crucially, when shifting focus to upstream Transcriptional Regulation, we identified the MYB gene family as the dominant class of transcription factors within the module. Specifically, 38 genes were identified as containing the Myb_DNA-binding domain (PF00249), ranking first among all transcription factor families (Fig. 10c), followed closely by the AP2 (37 genes) and bHLH (34 genes) families. Given that MYB transcription factors are known to form MBW complexes with bHLH proteins to synergistically regulate cell fate, the co-enrichment of MYB and bHLH factors further supports the existence of a coordinated regulatory network. The significant enrichment of the MYB family implies that this family acts as the primary "regulatory hub" driving the *G. barbadense*-specific elongation program within Module 6.

Based on these findings, we hypothesized that the species-specific expression of Module 6 might be driven by genetic variations in core MYB members.

Genomic Evolutionary Dynamics of Key MYB Transcription Factor Families

To trace the "genomic origins" driving the observed transcriptomic remodeling in fibers, we focused on the MYB transcription factor superfamily, which is intimately associated with the regulation of fiber development. Utilizing pan-genomic approaches, we performed genome-wide mining, PAV identification, and structural variation analysis within *G. hirsutum* (TM-1) and *G. barbadense* (H7124, 3-79) (Fig. 11).

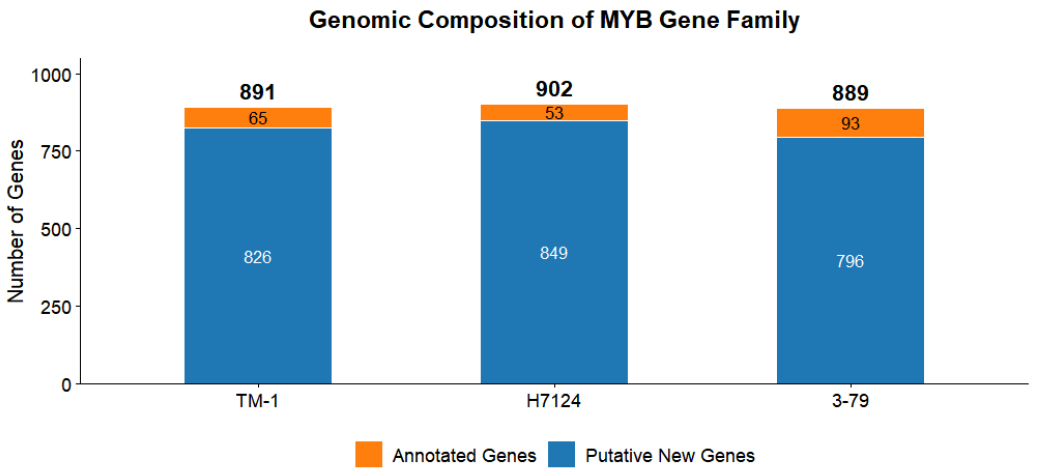

Fig. 11 Genomic composition of the MYB gene family in *G. hirsutum* and *G. barbadense*. Stacked bar chart showing the total copy number and composition of the MYB gene family in *G. hirsutum* (TM-1) and *G. barbadense* (H7124, 3-79). Blue bars (Annotated Genes): Genes present in the original reference genome annotations. Orange bars (Putative New Genes): Novel genes identified by the Bitacora pipeline that were absent in the original GFF3 files. Numbers on top of the bars indicate the total gene count for each family in each accession. Note the high conservation of total gene counts across the three genomes, suggesting that structural variations of specific members rather than global family expansion drive the evolutionary divergence.

The analysis revealed that the MYB family exhibits high evolutionary conservation among *Gossypium* species. The total count of MYB genes in *G. hirsutum* (TM-1) was 891, which is remarkably similar to that of *G. barbadense* accessions H7124 (902) and 3-79 (889). Although the Bitacora pipeline identified a substantial proportion (6%–10%) of unannotated "Putative New Genes" in each genome—thereby significantly refining the existing genomic annotations—no large-scale family expansion or contraction was observed between the species.

This quantitative stability at the whole-genome level suggests that the phenotypic differentiation between *G. hirsutum* and *G. barbadense* is not driven by a transcription factor "Dosage Effect." Instead, it is more likely attributable to the "functional replacement" or "asymmetric loss" of specific key members. This finding prompted us to narrow our focus to a microscopic genomic analysis of Presence/Absence Variations (PAVs).

Distribution Characteristics of Genomic PAVs and Identification of Key MYB Factors

To investigate the genomic basis underlying the significant agronomic trait differences between the two species, we identified Presence/Absence Variations (PAVs) between *G. hirsutum* (TM-1) and *G. barbadense* (H7124) using a whole-genome alignment strategy. Statistical analysis revealed a remarkable balance in the number of species-specific genes: a total of 1,130 H7124-specific PAV genes and 1,084 TM-1-specific PAV genes were identified.

To mine key factors regulating fiber and root development, we intersected the PAV gene set with the MYB transcription factor family and implemented a rigorous "false-positive elimination" workflow (Fig. 12).

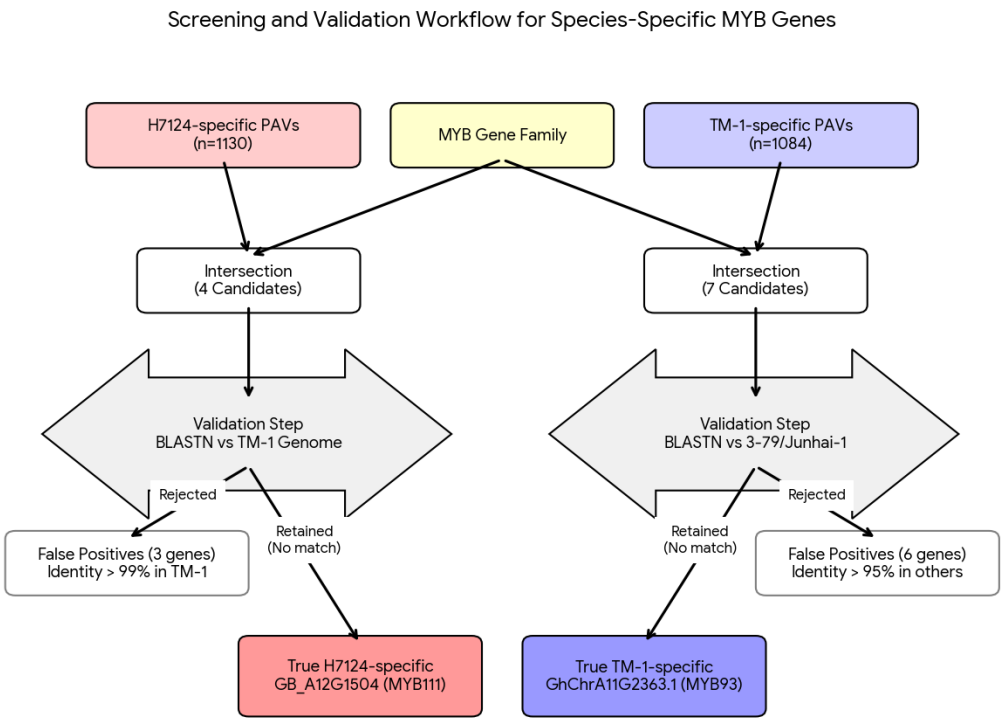

Fig. 12 Schematic workflow for identifying high-confidence species-specific MYB genes.

For the screening of *G. barbadense* (H7124)-specific genes: We aligned the CDS sequences of the four initial H7124-specific MYB candidates back to the TM-1 reference genome. BLASTN validation revealed that three of these genes possessed highly homologous sequences (Identity >99%) in TM-1 and were thus designated as false positives and discarded. Only GB_A12G1504 (*MYB111*) survived the validation: this gene is intact in H7124 but is unambiguously absent from the corresponding position on the homologous chromosome (A12) in TM-1, with only a homeolog exhibiting significant sequence divergence (Identity 96%) present in the D-subgenome. Integrating transcriptomic data (Fig. 13), GB_A12G1504 displayed extremely high fiber specificity, with its expression peak appearing precisely at 10 DPA, coinciding with the rapid fiber elongation phase.

For the screening of *G. hirsutum* (TM-1)-specific genes: Employing the same strategy, we validated seven TM-1-specific MYB candidates. After cross-validation against auxiliary Sea Island cotton genomes (3-79, Junhai-1) to exclude false positives, only *GhChrA11G2363.1* (*MYB93*) was confirmed as a high-confidence TM-1-specific PAV gene. This gene is present in TM-1 but is completely lost in the *G. barbadense* genome.

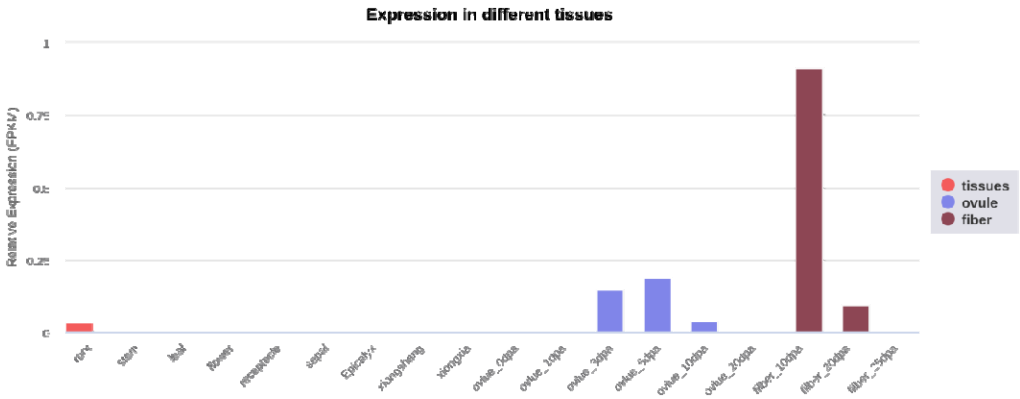

Fig. 13 Tissue-specific expression pattern of *GB_A12G1504* (*MYB111*) in H7124. The bar chart (based on FPKM) shows that this gene is explicitly expressed in fibers, with a distinct peak at 10 DPA, coinciding with the rapid elongation stage.

Genome-wide Structural Variation Analysis Corroborates Reciprocal Homoeologous Replacement of Key MYB Genes

To validate whether the species-specific genes identified by BLAST originated from authentic genomic physical variations rather than assembly artifacts, we employed the SyRI tool to perform a high-resolution analysis of synteny and structural variations (SVs) between *G. hirsutum* (TM-1) and *G. barbadense* (H7124) at the whole-genome level. This analysis not only confirmed macroscopic synteny at the genome-wide level (Fig. S1) but, more importantly, unveiled two distinct variation mechanisms driving the differentiation of the MYB family at the micro-scale (Fig.14).

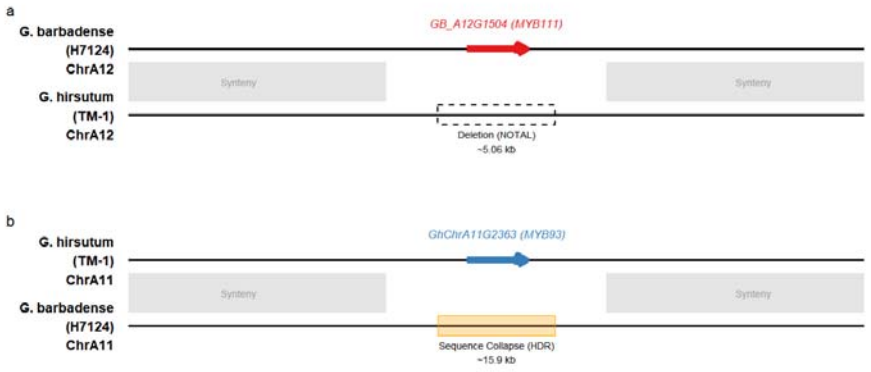

Fig. 14 SyRI-based validation of structural variations in key MYB loci. (a) Visualization of the physical deletion (NOTAL) of *MYB111* in TM-1. (b) Visualization of the structural collapse (HDR) of *MYB93* in H7124.

First, regarding GB_A12G1504 (MYB111), which is specifically retained in *G. barbadense*, SyRI analysis using H7124 as the reference (Supplementary Fig. 1) demonstrated that the gene is located intact on chromosome A12. However, in the corresponding syntenic position of *G. hirsutum* (TM-1), the region was identified as a 5.06 kb non-aligned region (NOTAL)(coordinates: ChrA12: 77,954,240–77,959,302). This result definitively confirms that *MYB111* underwent a

clear physical deletion in the A-subgenome of Upland cotton, rather than being a result of sequence alignment failure.

Conversely, regarding *GhChrA11G2363.1* (*MYB93*), which is specifically retained in *G. hirsutum*, the analysis using TM-1 as the reference revealed a contrasting landscape (Supplementary Fig. 2). Although TM-1 and H7124 maintained macroscopic synteny in this region of chromosome A11, the MYB93 locus (coordinates: chrA11: 47,300,518–47,301,823) corresponded to a 15.9 kb Highly Diverged Region (HDR) in *G. barbadense* (coordinates: chrA11: 47,297,796–47,313,692). This indicates that while *G. barbadense* retained the corresponding chromosomal scaffold, this region underwent drastic sequence rearrangement or degeneration, resulting in the structural collapse of the orthologous gene.

Collectively, structural genomic evidence demonstrates that *G. hirsutum* lost MYB111 via physical deletion (Deletion/NOTAL), whereas *G. barbadense* lost MYB93 via sequence collapse (Divergence/HDR). These genomic remodeling events, occurring on homoeologous chromosomes A11 and A12 via distinct mechanisms, constitute a quintessential phenomenon of "Reciprocal Homoeologous Replacement (RHR)."

Pan-Genomic Comparison Reveals Reciprocal Homoeologous Replacement Patterns of Key Genes

Having confirmed the physical Presence/Absence Variations (PAVs) of MYB93 and MYB111 in TM-1 and H7124, we sought to explore the distribution of these variations throughout the evolutionary history of the *Gossypium* genus. We extended our analysis to the genomes of 26 representative *Gossypium* species, utilizing the CDS sequences of these two genes as queries for pan-genomic homology searches and coverage analyses (Fig. 15).

Regarding MYB93 on chromosome A11, the alignment results (Fig. 15, bottom row) indicated that a high-coverage, high-identity A-subgenome specific copy (Red dots, Identity) was detectable in the A-genome progenitors (*G. herbaceum*, *G. arboreum*) and all wild tetraploids (*G. tomentosum*, *G. darwinii*, *G. stephensii*). This high-homology feature was completely preserved within the Upland Cotton Clade (*G. hirsutum* accessions). However, in all accessions of the Sea Island Cotton Clade (*G. barbadense*), we failed to detect high-homology sequences from the A-subgenome; only D-subgenome homoeologs with significant sequence divergence (Blue dots, Identity) were identified. This confirms that MYB93 underwent a lineage-specific loss in the Sea Island clade.

Regarding MYB111 on chromosome A12, the analysis revealed a diametrically opposite pattern (Fig. 15, top row). All accessions within the Sea Island Clade retained the A-subgenome copy (Red dots) that was highly consistent with that of the wild tetraploids. In sharp contrast, across all accessions in the Upland Clade, only homoeologous sequences derived from the D-subgenome could be detected at this locus (Blue dots). This indicates that the A-subgenome MYB111

underwent a fixed physical deletion event during the domestication of Upland cotton.

Collectively, pan-genomic evidence demonstrates that these two functionally distinct MYB homoeologs exhibit a strict pattern of "Reciprocal Retention" during the differentiation of cultivated species: the Upland lineage specifically retained MYB93, whereas the Sea Island lineage specifically retained MYB111.

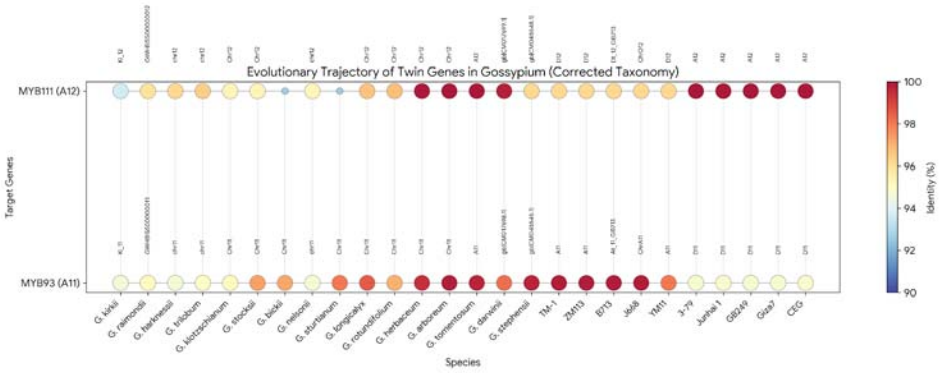

Fig. 15 Evolutionary trajectories of *MYB93* and *MYB111* across 26 *Gossypium* genomes.Bubble plots displaying the sequence identity (color) and coverage (size) of (top) *MYB111* (A12) and (bottom) *MYB93* (A11) across the *Gossypium* pan-genome. Species are arranged phylogenetically: Wild Diploids, Wild Tetraploids, Upland Clade (*G. hirsutum* accessions), and Sea Island Clade (*G. barbadense* accessions). Red dots (Identity > 97%) indicate the presence of an intact A-subgenome specific copy. Blue dots (Identity 90–97%) indicate the absence of the A-subgenome copy, with only D-subgenome homologs or diverged sequences detected. Absence of dots indicates complete deletion. The patterns reveal a reciprocal loss event: *MYB93* is lost in the Sea Island lineage, while *MYB111* is lost in the Upland lineage.

Discussion

Genomic and Transcriptomic Architecture of the Yield-Quality Trade-off

This study integrates multi-tissue transcriptomics, pan-genomics, and structural variation analysis to unravel the potential genetic basis of the "negative correlation between yield and fiber quality," a perennial conundrum in cotton breeding (Liu et al. 2024). Our data suggestthat this phenotypic differentiation may not be solely the result of stochastic drift, but appears to be intricately linked to the mechanism of "Reciprocal Homoeologous Replacement (RHR)" involving key MYB transcription factor families. This phenomenon reflects the asymmetric remodeling of subgenome functions driven by distinct structural variation events on homoeologous chromosomes.

Evolutionary Constraints and Adaptive Differentiation: An Evolutionary Trade-off between "Resource Acquisition" and "Sink Development Timing"

Our comparative transcriptomic landscape unveils a fundamental paradigm in cotton

domestication: the high evolutionary conservation of vegetative functions versus the extreme differentiation of reproductive strategies.

Despite millions of years of interspecific divergence and independent histories of artificial domestication, *G. hirsutum* (TM-1) and *G. barbadense* (H7124) exhibit remarkable conservation in the expression patterns of vegetative organs (root, stem, leaf) and floral organs (petal, sepal). The high correlation of gene expression counts (Fig. 2) and the consistency of Tau index distributions (Fig. 6) in these tissues indicate that the core transcriptional programs maintaining basic physiological activities—such as photosynthesis (involving genes like Rubisco), long-distance nutrient transport, and floral organ morphogenesis—have been subject to strong Evolutionary Constraints or Canalization. These "housekeeping" networks constitute the essential "Chassis" for the survival of *Gossypium* species, ensuring stability in maintaining basic survival capabilities across diverse ecological environments.

In stark contrast, the two species exhibit significant asymmetry in the distribution of specific gene repertoires. This reflects that, when facing selection pressures for "yield" versus "quality," they evolved distinct adaptive strategies:

- ***G. hirsutum*: A Strategy of "Adaptability & Reproductive Potential"**

The domestication of *G. hirsutum* (TM-1) for "high yield" essentially engineered a systemic infrastructure ranging from underground support to a reproductive burst.

First, the foundation of environmental adaptability. TM-1 exhibited a significant expansion of specific genes in the root system (Fig. 8) and specifically retained MYB93. This not only endowed it with a robust root architecture for resource acquisition but is likely intrinsically linked to its broad ecological adaptability. In contrast to the stringent environmental dependency of *G. barbadense*, this genetic foundation ensures that Upland cotton can establish a robust vegetative growth architecture across diverse global planting regions.

Second, the reproductive burst through sink capacity establishment. The core of the yield disparity lies in the establishment of "Sink Capacity." Upland cotton typically possesses 4–5 locules, resulting in a significantly higher number of seeds per boll compared to the 3-locule structure of Sea Island cotton (Viot and Wendel 2023). Our data precisely captured the transcriptomic imprints of this trait, which manifests not only as increased male investment but also as a fundamental remodeling of the female developmental program.

On one hand, TM-1 showed significant specific gene amplification in anthers (Figs. 2, 8), potentially enhancing pollen viability and fertilization efficiency to ensure multi-ovule fertilization. On the other hand, our common gene analysis (Fig. 9) revealed an even more drastic female differentiation: the pistil exhibited the lowest proportion of shared genes ($< 5\%$) among all organs, excluding fibers and ovules. As the maternal tissue determining the number of locules

(carpels), the high specificity of the pistil transcriptome suggests that Upland cotton may have breached the "3-locule" genetic constraint of Sea Island cotton by reconstructing the developmental network of female reproductive organs, thereby establishing the "4–5 locule" high-yield architecture.

This co-evolution of "Anther Expansion (Fertilization Assurance)" and "Pistil Remodeling (Locule Increment)" collectively underpins the yield advantage of Upland cotton in terms of boll number and seed count per plant (Zhang et al. 2015; Hu et al. 2019).

- ***G. barbadense*: A Strategy of "Sink Developmental Extension"**

Conversely, *G. barbadense* (H7124) retained a highly complex repertoire of specific genes in developing fibers and ovules—the ultimate storage sinks for photosynthates (Fig. 7). In particular, the retention of MYB111 enabled Sea Island cotton to bypass conventional developmental checkpoints, maintaining a highly active fiber elongation network even at 20 DPA. This achieved a "Temporal Extension" of the developmental window.

This strategy is not merely about increasing sink size, but rather about achieving refined structuralization and quality enhancement by prolonging the developmental duration of sink cells. This suggests that the "superior quality" domestication of Sea Island cotton involved a concentrated and sustained investment in specific reproductive sink organs, potentially at the cost of partial redundancy in root development or robustness in environmental adaptation.

The differentiation of these transcriptomic strategies—*G. hirsutum* prioritizing "Robust Acquisition & High Biomass" versus *G. barbadense* prioritizing "Refinement & Extended Duration"—provides a systems biology-level explanation for the breeding phenomenon where "high yield and superior quality are difficult to combine." These two traits rely on fundamentally different, and perhaps energetically competing, genomic strategies (White et al. 2016) that have been fixed within the species' genetic blueprints.

Spatiotemporal Transcriptomic Remodeling and the "Molecular Trigger" of Fiber Quality Formation

Throughout the domestication history of cotton, distinct selection pressures targeting fiber length (*G. barbadense*) versus fiber biomass (*G. hirsutum*) have not only sculpted the static genomic landscape but also induced profound Spatiotemporal Expression Remodeling of developmental programs. Our multi-tissue transcriptomic analysis precisely captured the pivotal inflection point of this dynamic process: 20 Days Post Anthesis (20 DPA).

During this critical transition window of fiber development—shifting from rapid elongation to secondary cell wall thickening (Haigler et al. 2012)—we observed a striking species-specific

"Phase Inversion" (Fig. 9). At this stage, *G. barbadense* maintained a dominance of up to ~70% in species-specific gene expression, whereas *G. hirsutum*-specific genes were largely suppressed. Integrating the specific high-expression peak of GB_A12G1504 (MYB111) at 10 DPA (Fig. 12), we propose a regulatory model of "Delayed Elongation":

The "Relay" Strategy of *G. barbadense*: The transcriptional burst of MYB111 at 10 DPA likely acts as a "Molecular Trigger." Its primary function is to activate or sustain downstream elongation-related gene networks, such as the Expansin family (Li et al. 2016) or cytoskeleton protein genes. This continuous signal input potentially enables *G. barbadense* to bypass the conventional elongation termination checkpoint, maintaining cell wall extensibility even at 20 DPA, thereby achieving the "extra-long" fiber phenotype.

The "Early-Maturing" Strategy of *G. hirsutum*: In contrast, due to the physical deletion of A-subgenome *MYB111* (detailed below), *G. hirsutum* lacks this critical delay signal. Consequently, its fiber cells rapidly exit the elongation program at 20 DPA and switch to activating the secondary cell wall synthesis network (evidenced by the rapid surge and dominance of *G. hirsutum*-specific gene abundance at 25 DPA, Fig. 9). While this "early-maturing" strategy limits fiber length, the premature initiation of cellulose deposition significantly shortens the developmental cycle and increases fiber specific gravity (micronaire). Theoretically, this aligns better with the breeding objectives of "high ginning rate and high boll weight."

This finding suggests that the disparity in cotton fiber quality is not merely the result of mutations in a few structural genes (e.g., CesA) (Liu et al. 2024). Rather, it is profoundly influenced by a "Developmental Window Shift" triggered by the temporal expression divergence of upstream regulators (such as *MYB111*).

Reciprocal Homoeologous Replacement (RHR): The Genomic Mechanism of Divergent Domestication

If spatiotemporal transcriptomic remodeling represents the "software" manifestation of phenotypic differentiation, then genomic structural variation provides the underlying "hardware" foundation. Our pan-genomic and micro-synteny analyses identified a distinct evolutionary pattern, which we term "Reciprocal Homoeologous Replacement (RHR)."

On the homoeologous chromosomes A11 and A12, the two cultivated species appear to have achieved "reciprocal selection" for potential key agronomic traits via distinctly different structural variation mechanisms:

- "Quality Sacrifice" on Chromosome A12: G. hirsutum lost the chromosomal segment containing

MYB111 through a definitive physical deletion (Deletion/NOTAL) event (Fig. 13a). *MYB111* is generally regarded as a key regulator of flavonol biosynthesis in the plant kingdom (Stracke et al. 2007), and flavonoids have been proven to be highly correlated with cotton fiber development (Tan et al. 2013). Given the specific expression of *MYB111* during the fiber elongation phase in *G. barbadense*, this deletion, while fixed in the pan-genome (Fig. 14), likely represents a critical

strategy of "Resource Reallocation." By eliminating the energy-demanding program of excessive fiber elongation, carbon sources are conserved for seed development and secondary wall thickening. More importantly, this "early exit from the elongation program" directly endows *G. hirsutum* with early maturity, ensuring that its massive reproductive sink (numerous bolls and seeds) can be fully filled within a limited growing season.

- "Root System Remodeling" on Chromosome A11: Conversely, *G. barbadense* lost MYB93on

chromosome A11 via sequence collapse and high divergence (Divergence/HDR)(Fig. 13b), whereas the gene remains intact in *G. hirsutum*. Notably, *MYB93* has been reported to act as a negative regulator of lateral root development in other plants (Gibbs et al. 2014; Zhang et al. 2025b). Assuming a conserved function in cotton, the specific retention of this "lateral root inhibitor" in *G. hirsutum* implies a strategy for optimizing root system architecture. By limiting redundant lateral root growth, it facilitates deeper taproot penetration or optimizes the root-to-shoot ratio. This mechanism helps reduce unnecessary biomass consumption in the underground parts while maintaining nutrient absorption efficiency, thereby supporting high-yield reproductive growth.

Therefore, the RHR model proposed herein provides a plausible genomic explanation: polyploid genomes may achieve "functional compensation" and "genetic trade-offs" for complex traits through asymmetric structural variations between homoeologous chromosomes. *G. hirsutum* retains root-related genes while losing fiber elongation genes, and vice versa for *G. barbadense*. This genomic interplay is likely one of the fundamental genetic roots underlying the perennial breeding dilemma of the "negative correlation between yield and quality."

Implications for Functional Gene Mining and Molecular Breeding

The genomic and transcriptomic divergences revealed in this study provide precision genetic targets for decoupling the negative correlation between yield and fiber quality in cotton. Based on the RHR model, we propose two complementary molecular breeding strategies aimed at achieving the optimal assembly of the "Chassis" (broad adaptability) and "Modules" (specific traits):

First, for the adaptability improvement of *G. barbadense*: The *G. hirsutum*-specific "High-Yield & Adaptability Module" offers valuable genetic resources. For decades, the defects of "low yield and difficult cultivation" have severely restricted the widespread adoption of Sea Island cotton. *MYB93* (a potential regulator of root system architecture), which we identified in TM-1, serves as an ideal candidate target for enhancing its environmental adaptability. More critically, the specific gene clusters significantly amplified in the anthers of TM-1 (Fig. 8) likely harbor the genetic code for improving pollen viability and fruit set rate. Future breeding efforts could employ gene editing (e.g., CRISPR/Cas9) or Marker-Assisted Selection (MAS) to synergistically introduce these "Robust Root (Stress Tolerance)" and "High-Fertility Anther" elements into *G. barbadense* (Sheri et al. 2025). This strategy promises to overcome its agronomic drawbacks of being "fragile (low yield/narrow adaptability)" while preserving its superior fiber quality.

Second, for the quality enhancement of *G. hirsutum*: The *G. barbadense*-specific "Delayed Elongation Module" (Avci et al. 2013) represents a critical breakthrough point. The physical deletion of *MYB111* is likely the bottleneck limiting fiber length in Upland cotton. Utilizing genetic engineering to knock-in this gene, or employing strong promoters to reconstruct its specific expression peak at 10–20 DPA, could potentially reactivate the "delayed elongation" program in Upland cotton fibers. However, we must be vigilant regarding the potential "late-maturity risks" and additional carbon cost associated with such improvements. Therefore, future breeding by design should focus on developing Inducible Smart Expression Systems, striving to precisely extend the fiber elongation window without significantly prolonging the growth period, thereby achieving an effective coupling of "high yield" and "superior quality."

Conclusion

In summary, by integrating multi-omics analyses, this study constructs a systematic genetic model underlying the differentiation of yield and quality in cotton. Our findings demonstrate that the root cause of the phenotypic divergence between *G. hirsutum* and *G. barbadense* lies in the transcriptomic "software"—specifically, the spatiotemporal remodeling of the fiber developmental program—driven by the genomic "hardware," namely, the Reciprocal Homoeologous Replacement (RHR) events occurring on homoeologous chromosomes.

G. hirsutum adopted a strategy of "concomitant addition and subtraction": by retaining root- and anther-related modules (e.g., *MYB93*) and remodeling the highly specific developmental network of the pistil, it established a systemic yield foundation characterized by "Broad Adaptability (Root), High Fertility (Anther), and Multi-locule Sink Capacity (Pistil)." Simultaneously, by deleting the fiber elongation gene (*MYB111*), it sacrificed length in exchange for "Early Maturity and Filling Efficiency." In contrast, *G. barbadense* retained an intact delayed elongation module, establishing a high-quality architecture that trades "Time for Quality."

This discovery not only theoretically dissects the genomic mechanisms of divergent domestication in polyploid crops but also provides a clear blueprint for creating novel cotton germplasm combining "high yield and superior quality" through Precision Genomic Design—the intelligent assembly of a "High-Yield Chassis" with "High-Quality Modules."

Funding

This work was supported in part by the High Performance Computing(HPC) clusters at Southwest University.

References

Wen X, Chen Z, Yang Z, et al (2023) A comprehensive overview of cotton genomics, biotechnology

and molecular biological studies. Sci China Life Sci 66:2214–2256.

https://doi.org/10.1007/s11427-022-2278-0
Avci U, Pattathil S, Singh B, et al (2013) Cotton fiber cell walls of Gossypium hirsutum and Gossypium barbadense have differences related to loosely-bound xyloglucan. PLoS ONE 8:e56315. https://doi.org/10.1371/journal.pone.0056315
Alagarsamy M (2023) Assessing genetic variation in Gossypium barbadense L. germplasm based on fibre characters. J Cotton Res 6:15. https://doi.org/10.1186/s42397-023-00153-y
Zhang J, Yu M, Guo Y, et al (2025) Evolutionary divergence of an ethylene-responsive transcriptional cascade governs a dose-dependent balance between cotton fiber length and strength. Adv Sci. https://doi.org/10.1002/advs.202514154
Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases.

Nature 461:747–753. https://doi.org/10.1038/nature08494

Pinglay S, Lalanne J-B, Daza RM, et al (2025) Multiplex generation and single-cell analysis of structural variants in mammalian genomes. Science 387:ado5978. https://doi.org/10.1126/science.ado5978

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol 42:225–249.

https://doi.org/10.1023/A:1006392424384
Yan H, Han J, Jin S, et al (2025) Post-polyploidization centromere evolution in cotton. Nat Genet. https://doi.org/10.1038/s41588-025-02115-3
Hu Y, Chen J, Fang L, et al (2019) Gossypium barbadense and Gossypium hirsutum genomes

provide insights into the origin and evolution of allotetraploid cotton. Nat Genet 51:739–748.

https://doi.org/10.1038/s41588-019-0371-5
Yang Z, Wang J, Huang Y, et al (2023) CottonMD: a multi-omics database for cotton biological

study. Nucleic Acids Res 51:D1446–D1456. https://doi.org/10.1093/nar/gkac863

Yu J, Jung S, Cheng CH, et al (2021) CottonGen: the community database for cotton genomics, genetics, and breeding research. Plants 10:2805. https://doi.org/10.3390/plants10122805
R Core Team (2025) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York
Yanai I, Benjamin H, Shmoish M, et al (2005) Genome-wide midrange transcription profiles reveal

expression level relationships in human tissue specification. Bioinformatics 21:650–659.

https://doi.org/10.1093/bioinformatics/bti042

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605

McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426
Conway JR, Lex A, Gehlenborg N (2017) UpSetR: an R package for the visualization of intersecting

sets and their properties. Bioinformatics 33:2938–2940.

https://doi.org/10.1093/bioinformatics/btx364

Vizueta J, Sánchez-Gracia A, Rozas J (2020) BITACORA: a comprehensive tool for the identification and annotation of gene families in genome assemblies. Mol Ecol Resour 20:1445–1452. https://doi.org/10.1111/1755-0998.13202

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

Mistry J, Chuguransky S, Williams L, et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913

Marçais G, Delcher AL, Phillippy AM, et al (2018) MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Goel M, Sun H, Jiao WB, et al (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 20:277. https://doi.org/10.1186/s13059-019-1911-0

Goel M, Schneeberger K (2022) plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics 38:2922–2926. https://doi.org/10.1093/bioinformatics/btac196

Cock PJ, Antao T, Chang JT, et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423

Liu R, Xiao X, Gong J, et al (2024) Genetic linkage analysis of stable QTLs in Gossypium hirsutum RIL population revealed function of GhCesA4 in fiber development. J Adv Res 60:133–148. https://doi.org/10.1016/j.jare.2023.12.005

White AC, Rogers A, Rees M, et al (2016) How can we make plants grow faster? A source–sink perspective on growth rate. J Exp Bot 67:31–45. https://doi.org/10.1093/jxb/erv447

Haigler CH, Betancur L, Stiff MR, Tuttle JR (2012) Cotton fiber: a powerful single-cell model for cell wall research. Front Plant Sci 3:104. https://doi.org/10.3389/fpls.2012.00104

Li Y, Tu L, Pettolino FA, et al (2016) GbEXPATR, a species-specific expansin, enhances cotton fibre elongation through cell wall restructuring. Plant Biotechnol J 14:1006–1017. https://doi.org/10.1111/pbi.12450

Zhang D, Zhou H, Zhang Y, et al (2025) Diverse roles of MYB transcription factors in plants. J Integr Plant Biol. https://doi.org/10.1111/jipb.13869

Tan J, Tu L, Deng F, et al (2013) A genetic and metabolic analysis revealed that cotton fiber cell development was retarded by flavonoid naringenin. Plant Physiol 162:86–95. https://doi.org/10.1104/pp.112.212142

Stracke R, Ishihara H, Huep G, et al (2007) Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. Plant J 50:660–677. https://doi.org/10.1111/j.1365-313X.2007.03078.x

Sheri V, Mohan H, Jogam P, et al (2025) CRISPR/Cas genome editing for cotton precision breeding: mechanisms, advances, and prospects. J Cotton Res 8:4. https://doi.org/10.1186/s42397-024-00206-w

Wang M, Li J, Qi Z, et al (2022) Genomic innovation and regulatory rewiring during evolution of the cotton genus Gossypium. Nat Genet 54:1959–1971. https://doi.org/10.1038/s41588-022-01237-2

Wang M, Li J, Wang P, et al (2021) Comparative genome analyses highlight transposon-mediated genome expansion and the evolutionary architecture of 3D genomic folding in cotton. Mol Biol Evol 38:3621–3636. https://doi.org/10.1093/molbev/msab128

Grover CE, Yuan D, Arick MA, et al (2019) The genome sequence of Gossypioides kirkii illustrates a descending dysploidy in plants. Front Plant Sci 10:1541. https://doi.org/10.3389/fpls.2019.01541

Xu Y, Wei Y, Zhou Z, et al (2024) Widespread incomplete lineage sorting and introgression shaped adaptive radiation in the Gossypium genus. Plant Commun 5:100728. https://doi.org/10.1016/j.xplc.2023.100728

Huang G, Bao Z, Feng L, et al (2024) A telomere-to-telomere cotton genome assembly reveals centromere evolution and a Mutator transposon-linked module regulating embryo development. Nat Genet 56:1953–1963. https://doi.org/10.1038/s41588-024-01877-6

Chen ZJ, Sreedasyam A, Ando A, et al (2020) Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat Genet 52:525–533. https://doi.org/10.1038/s41588-020-0614-5

Peng R, Xu Y, Tian S, et al (2022) Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. Proc Natl Acad Sci USA 119:e2208496119. https://doi.org/10.1073/pnas.2208496119

Xu Z, Wang G, Zhu X, et al (2025) Genome assembly of two allotetraploid cotton germplasms reveals mechanisms of somatic embryogenesis and enables precise genome editing. Nat Genet 57:2028–2039. https://doi.org/10.1038/s41588-025-02258-3

Perkin LC, Bell A, Hinze LL, et al (2021) Genome assembly of two nematode-resistant cotton lines (Gossypium hirsutum L.). G3 Genes Genomes Genet 11:jkab276. https://doi.org/10.1093/g3journal/jkab276

Wang J, Liang Y, Gong Z, et al (2023) Genomic and epigenomic insights into the mechanism of

cold response in upland cotton (Gossypium hirsutum). Plant Physiol Biochem 205:108206. https://doi.org/10.1016/j.plaphy.2023.108206
Hu G, Wang Z, Tian Z, et al (2025) A telomere-to-telomere genome assembly of cotton provides insights into centromere evolution and short-season adaptation. Nat Genet 57:1031–1043. https://doi.org/10.1038/s41588-025-02130-4
Meng Q, Xie P, Xu Z, et al (2025) Pangenome analysis reveals yield- and fiber-related diversity and interspecific gene flow in Gossypium barbadense L. Nat Commun 16:4995. https://doi.org/10.1038/s41467-025-60254-x
Chang X, He X, Li J, et al (2024) High-quality Gossypium hirsutum and Gossypium barbadense genome assemblies reveal the landscape and evolution of centromeres. Plant Commun 5:100722. https://doi.org/10.1016/j.xplc.2023.100722
Zhang T, Hu Y, Jiang W, et al (2015) Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol 33:531–537. https://doi.org/10.1038/nbt.3207
Li F, Fan G, Lu C, et al (2015) Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol 33:524–530. https://doi.org/10.1038/nbt.3208
Yuan D, Tang Z, Wang M, et al (2015) The genome sequence of Sea-Island cotton (Gossypium barbadense) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci Rep 5:17662. https://doi.org/10.1038/srep17662
Dubos C, Stracke R, Grotewold E, et al (2010) MYB transcription factors in Arabidopsis. Trends Plant Sci 15:573–581. https://doi.org/10.1016/j.tplants.2010.06.005
Machado A, Wu Y, Yang Y, et al (2009) The MYB transcription factor GhMYB25 regulates early fibre and trichome development. Plant J 59:52–62. https://doi.org/10.1111/j.1365-313X.2009.03847.x
Walford SA, Wu Y, Llewellyn DJ, et al (2011) GhMYB25-like: a key factor in early cotton fibre development. Plant J 65:785–797. https://doi.org/10.1111/j.1365-313X.2010.04464.x
Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559
Zhang X, Tian X, Gao X, et al (2025) Integrated metabolomic and transcriptomic analyses identify MYB genes regulating key metabolites and agronomic traits in upland cotton *Gossypium hirsutum*. Nat Genet 57:2819–2830. https://doi.org/10.1038/s41588-025-02363-3
Yao Y, Liu S, Xia C, et al (2022) Comparative transcriptome in large-scale human and cattle populations. Genome Biol 23:176. https://doi.org/10.1186/s13059-022-02745-4
Chen X, Chen W, Li X, et al (2012) Molecular mechanisms of fiber differential development between *G. barbadense* and *G. hirsutum* revealed by genetical genomics. PLoS ONE 7:e30056. https://doi.org/10.1371/journal.pone.0030056
Cheng H, Liu S, Zhang Y, et al (2025) Comparative single-cell transcriptomic map reveals

divergence in leaves between two cotton species at cell type resolution. J Adv Res. https://doi.org/10.1016/j.jare.2025.04.012
Viot CR, Wendel JF (2023) Evolution of the cotton genus, *Gossypium*, and its domestication in the Americas. Crit Rev Plant Sci 42:1–33. https://doi.org/10.1080/07352689.2022.2156061
Avci U, Pattathil S, Singh B, et al (2013) Cotton fiber cell walls of *Gossypium hirsutum* and *Gossypium barbadense* have differences related to loosely-bound xyloglucan. PLoS One 8:e56315. https://doi.org/10.1371/journal.pone.0056315