Genome Biology

RESEARCH

Open Access

Widespread impact of transposable elements on the evolution of post-transcriptional regulation in the cotton genus *Gossypium*

Xuehan Tian[1†], Ruipeng Wang[1†], Zhenping Liu[1], Sifan Lu[1], Xinyuan Chen[1], Zeyu Zhang[1,2], Fang Liu[3], Hongbin Li[4], Xianlong Zhang[1] and Maojun Wang[1,4*]

[†]Xuehan Tian and Ruipeng Wang contributed equally to this work.

*Correspondence: mjwang@mail.hzau.edu.cn

[1] National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China
[2] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[3] State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, Henan 455000, China
[4] College of Life Science, Shihezi University, Shihezi 832003, China

Abstract

Background: Transposable element (TE) expansion has long been known to mediate genome evolution and phenotypic diversity in organisms, but its impact on the evolution of post-transcriptional regulation following species divergence remains unclear.

Results: To address this issue, we perform long-read direct RNA sequencing, polysome profiling sequencing, and small RNA sequencing in the cotton genus *Gossypium*, the species of which range more than three folds in genome size. We find that TE expansion contributes to the turnover of transcription splicing sites and regulatory sequences, leading to changes in alternative splicing patterns and the expression levels of orthologous genes. We also find that TE-derived upstream open reading frames and microRNAs serve as regulatory elements mediating differences in the translation levels of orthologous genes. We further identify genes that exhibit lineage-specific divergence at the transcriptional, splicing, and translational levels, and showcase the high flexibility of gene expression regulation in the evolutionary process.

Conclusions: Our work highlights the significant role of TE in driving post-transcriptional regulation divergence in the cotton genus. It offers insights for deciphering the evolutionary mechanisms of cotton species and the formation of biological diversity.

Keywords: *Gossypium*, Transposable element, Direct RNA sequencing, Alternative splicing, Translation, MiRNA

Background

Evolution is a cornerstone in the field of biology, permeating research across the life sciences. The scientific community widely accepts the concept of the last universal common ancestor (LUCA), which posits that all living organisms share a common origin

[1–4]. This view is supported by a wealth of biological evidence spanning genomics, biogeography, paleobiology, and other fields of study [5–7]. A thorough investigation of evolution not only aids in understanding the origin and development of biodiversity but also reveals the adaptive evolutionary mechanisms underlying the structure, function, and behavior of organisms, providing a crucial theoretical foundation and guidance for research in the field of biology.

While genetic variation forms the foundation of biological evolution, some distantly related species exhibit high sequence similarity in their genomes [8], with their differences being regulated by gene expression. Thus, gene expression regulation plays a crucial role in species adaptive differentiation and evolution [9–11]. For instance, differential expression of *Hoxd13* regulates the adaptive evolution of tail length in deer mice [12]. Gene expression regulation is a complex process in which protein expression levels are influenced by various factors. In recent years, the emergence and advancement of omics technologies such as full-length transcriptomics [13, 14], translatomics [15–17], small RNA sequencing [18, 19], and proteomics [20, 21] have provided a more comprehensive perspective for studying posttranscriptional regulation, including the regulation of RNA splicing, modification, stability, and translation efficiency, among other aspects. In particular, alternative splicing (AS) and translation regulation have profound impacts on protein function. AS generates different mRNA isoforms from the same gene through different splicing mechanisms, which is a source of protein and phenotypic diversity [22, 23]. For example, the role of AS in regulating sex determination in turtles, as well as affecting the morphological size of plant floral organs and the length of fish fins, has been documented [24]. The ultimate goal of studying gene expression is to understand the protein level rather than the intermediate product mRNA. Proteins exhibit orders of magnitude differences from mRNAs in terms of half-life, synthesis rate, and abundance, indicating significant variability in protein levels [25, 26]. Research indicates that translational regulation (i.e., the regulation of protein synthesis) has a greater impact on proteins than does the sum of RNA synthesis (i.e., epigenetic and transcriptional regulation), mRNA degradation, and protein degradation [26, 27]. Hence, delving more deeply into the evolution of these posttranscriptional regulations can enhance our understanding of the intricacies and adaptive evolutionary mechanisms of organisms.

In the genomes of both animals and plants, a large number of transposable elements (TEs) exist. Host organisms utilize TEs to fulfill certain cellular functions, a phenomenon considered an evolutionary process that is more common than initially anticipated [28]. The insertion and movement of TEs can lead to gene recombination and chromosomal structural variations, providing the basis for genetic variation in species evolution [29, 30]. TEs also regulate gene transcription levels by introducing transcription factor-binding sites (TFBSs) and enhancers, modulating chromatin 3D structure, and generating noncoding RNAs [31]. Furthermore, TE insertions introduce transcription start sites that drive the transcription of various chimeric RNAs and protein isoforms, thereby generating novel transcripts [32, 33]. At the posttranscriptional level, TE insertions in the 5′-UTR (untranslated region) or 3′-UTR can impact protein expression [34]. TEs are also associated with upstream open reading frames (uORFs) that mediate translational regulation [35]. Miniature inverted-repeat transposable elements (MITEs) located in the 3′-UTR exert regulatory effects through translational suppression mechanisms [36].

Studies have reported that TE insertions in genes may lead to the production of novel protein isoforms through AS, thereby providing possibilities for organisms to adapt to new environments and develop new functions during evolution [37]. Therefore, transposons are regarded as an engine propelling species diversity and evolution. Researchers believe that focusing on TEs is crucial for unraveling the principles of transcriptional regulation and evolution [31].

The mechanisms of evolution are intricate and still harbor many mysteries, necessitating further in-depth research and exploration. Cotton is an important agricultural crop with multiple genome types and extensive geographical distribution and shows rich and diverse phenotypic characteristics, which can provide valuable materials for evolutionary studies. The 45 diploid cotton species mainly include eight genome types, A–G and K. Moreover, there are significant differences in genome sizes, with the largest K genome being more than three times larger than the smallest D genome [38, 39]. Based on phylogeny, diploid cotton species can be divided into three branches: the New World, the African-Asian (A, B, F), and the Australian clades (C, G, K) [40]. In previous work, our laboratory constructed a pangenome and pan-3D genome map covering all diploid cotton genome types and explored the evolution of regulatory elements and orthologous gene networks, involving some genetic variations and expression-related regulatory mechanisms involved in cotton evolution [41]. However, there have been no studies at the posttranscriptional regulatory level. Building upon this work, we investigated the impact of TE-mediated posttranscriptional regulation on diploid cotton species from eight different genome types, providing a new perspective and theoretical foundation for unraveling evolutionary mechanisms.

Results

Effect of TE expansion on gene evolution in cotton

Previous studies have indicated that there is a more than threefold difference in the size of cotton genomes (Additional file 1: Table S1) [41]. Notably, TEs accounted for a significant proportion of these genomes, ranging from 57% (D_5) to 81% (K_2), with the majority being class I LTR retrotransposons, primarily Gypsy elements (Additional file 1: Table S2). Using a TE library of 6180 consensus sequences, we re-annotated TEs across all cotton genomes to investigate the differential expansion of TEs.

Among the 6180 TE entries, 698 entries expanded in all cotton lineages, 582 entries specifically expanded in D_5 (cluster 6), 1148 entries specifically expanded in E_1 (cluster 1), 1401 entries specifically expanded in the African-Asian cottons (A_2, B_1, F_1) (cluster 2), and 2351 entries specifically expanded in the Australian cottons (C_1, G_1, K_2) (cluster 4). There are substantial numbers of lineage-specific TEs within the African-Asian and Australian clades, indicating a potential association between cotton lineage divergence and TEs differential expansion (Fig. 1a, Additional file 2: Fig. S1). To further investigate the impact of TEs on gene evolution after speciation, we categorized genes from eight cotton species into 18,676 single-copy orthologous genes, 12,501 (D_5)–13,768 (C_1) multicopy gene families, 6035 (K_2)–8938 (C_1) variable conserved genes, and 471 (C_1)–1990 (D_5) species-specific genes present in only one genome (Fig. 1b, Additional file 2: Fig. S2, Additional file 1: Table S3). To confirm the accuracy of their classification as species-specific, we compared these genes with orthologous genes from other cotton species.

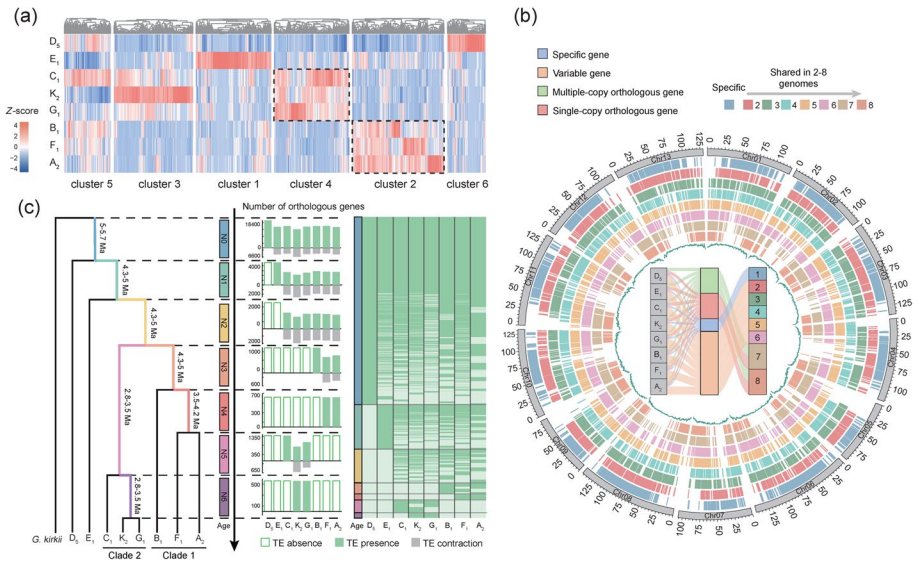

Fig. 1 Identification and insertion characteristics of transposable elements (TEs) in the genomes of eight diploid cotton species. **a** A clustered heatmap of *Z*-score normalized copy numbers of conserved TE families inserted in the genomes of eight cotton species, divided into 6 clusters. Cluster 5 represents conserved TEs that are amplified in all cotton species, cluster 6 shows amplification specific to D_5, cluster 1 exhibits amplification specific to E_1, cluster 2 indicates species-specific amplification in the African-Asian cottons (A_2, B_1, F_1), and cluster 4 shows species-specific amplification in the Australian cottons (C_1, G_1, K_2). Highlight with a dashed box. The *Z*-score normalization was applied to the copy numbers to account for large differences in copy numbers between species. **b** Circos plot illustrates the landscape of genomic blocks with varying degrees of conservation using A_2 as the reference genome. The outer to inner tracks represent species-specific genomic blocks, conserved genomic blocks among two to eight species, the green lines represent TE density. The chord diagram in the center displays single-copy orthologous gene families, multicopy gene families, variable gene families among 2–6 genomes, and species-specific genes among different cotton species. **c** The phylogenetic tree and conserved TE insertion map of eight cotton species. Divergence time among cotton lineages was estimated based on Ks peak, with a synonymous substitution rate (R) of 3.48×10^{-9}. The heatmap on the right displays the landscape of conserved transposon insertions and absence in eight cotton species at different evolutionary nodes, where green areas indicate the presence of conserved TE insertions and light areas represent the absence of conserved TE insertions. The bar graph in the middle represents the number of genes with or without conserved TE insertions in single-copy orthologous genes at each time node. An empty box indicates genes without conserved TE insertions, green bars represent genes with conserved TE insertions, and gray bars indicate genes with TE contractions

Genome-wide comparison and gene family analysis revealed no homologous counterparts for these genes in other species, confirming that the identified genes are indeed species-specific and not unannotated homologs or annotation errors. By comparing the TE coverage of different gene types, we found a significant enrichment of TEs on species-specific genes ($P < 0.0001$), suggesting that TE activity may contribute to the formation of species-specific genes (Additional file 2: Fig. S3). We also found that the number of conserved TE insertions on single-copy orthologous genes was highly similar among closely related species, such as the Australian and African-Asian cotton clades (Additional file 2: Fig. S4). This aligns with the general understanding of evolutionary biology, where closely related species share more genetic similarities due to their recent common ancestry. Conversely, in relatively distant species, there was significant divergence, consistent with the accumulation of genetic differences over time. These results, along with previous studies indicating that TEs provide a phylogenetic signal [42], further support the role of TEs in the genetic divergence of cotton species.

We focused on the impact of differential TE expansion on 18,676 single-copy orthologous genes (the following as orthologous genes). Based on the presence of conserved TE insertion events in genes, the orthologous gene families were classified into various evolutionary time nodes and a TE insertion map was constructed (Fig. 1c). This map provides a more intuitive display of differential TE expansion among orthologous genes in cotton. The orthologous genes on the N0 branch exhibited conserved TE insertions in all eight species, indicating that TE insertions occurred before species divergence of the *Gossypium* genus. Within this branch, we identified 4103 (F_1) to 6557 (K_2) genes that have undergone TE contraction, which also serve as potential factors leading to structural and functional differences in orthologous genes. There were 1183 gene families in the N3 branch (clade 1) that exhibited conserved TE insertions only in the African-Asian cotton, and 1384 gene families in the N5 branch (clade 2) that exhibited conserved TE insertions only in the Australian cotton (C_1, G_1, K_2) (Fig. 1c). These results further indicate that TEs potentially contributed to the genetic divergence of cotton species.

Differential TE expansion regulates orthologous expression divergence

To explore the impact of TE on posttranscriptional regulation during the divergence of cotton species, we first analyzed gene transcription in eight cotton species to determine their transcription status and any differences in transcription levels. We constructed RNA-seq libraries and quantified gene expression in each species (Additional file 1: Tables S4, S5). Out of 18,676 single-copy orthologous genes (the following as orthologous genes), we categorized 1636 genes as not expressed in any species because their FPKM values were ≤ 0.05. Additionally, we identified 8420 genes that were expressed in only some (1–7) cotton species and 8620 genes that were expressed in all eight species (Fig. 2a middle). Among the genes expressed in only some cotton species, 32.30% of orthologous genes were not expressed in a single cotton species, and 7.69% of orthologous genes were detected to be expressed in only one cotton species (Fig. 2a right). Genes expressed in some species (FPKM > 0.05) but not in others suggest a regulatory mechanism controlling their transcription presence or absence. This differs from differentially expressed genes, which are expressed across all species but show varying expression levels, implying differential regulation or adaptation. For example, among the genes expressed in all species, significant differences in expression levels were observed. In a comparison between K_2 and A_2, we found 1076 genes upregulated and 1037 genes downregulated in A_2 relative to K_2 (Fig. 2a left). This distinction helps us better understand the regulatory dynamics and potential functional significance of gene expression across different cotton species.

To investigate the impact of TEs on the transcriptional regulation of orthologous genes, the differential TE insertions within 2 kb upstream of orthologous genes across different cotton species were examined (Additional file 1: Table S6). Subsequently, we analyzed the orthologous genes with these differential TE insertions and the differences in their transcription. The results revealed that among genes expressed only in some species, the predominant TE type with differential insertions was DNA transposons, accounting for 45.6% (Fig. 2b left). However, the total length of DNA transposons only accounts for 6% of all TEs. A chi-square test confirmed that the proportion of DNA transposons in these genes is significantly higher than expected based on

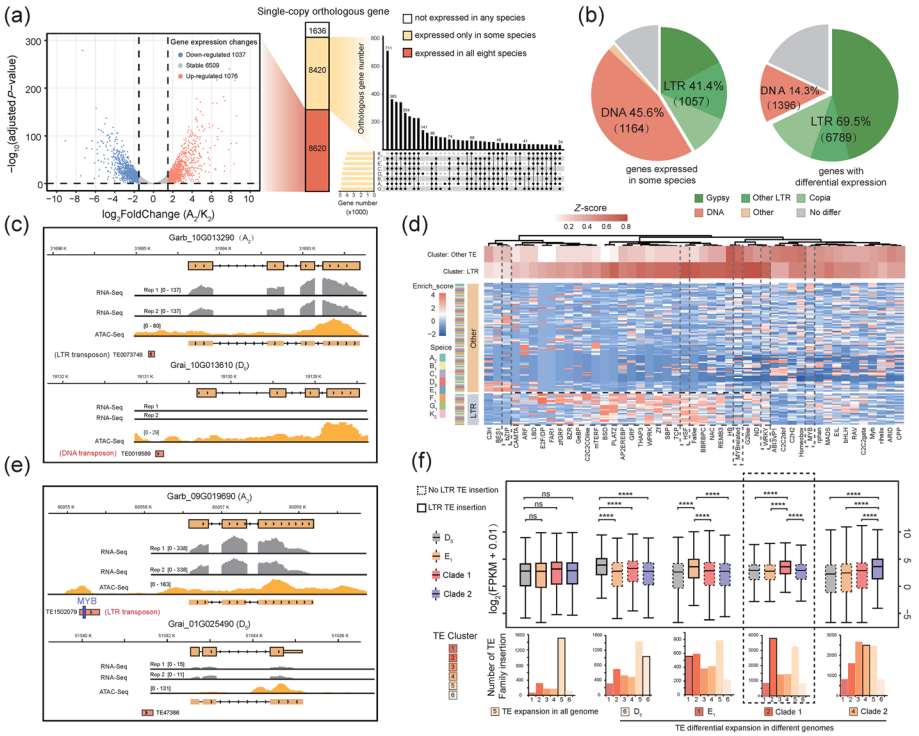

Fig. 2 The regulatory effects of different TE families on gene transcription. **a** The expression patterns of single-copy orthologous genes in eight cotton species. The central bar plot shows the number of orthologous genes that are not expressed in any of the eight cotton species (white), expressed in some cotton species (yellow), and expressed in all eight cotton species (red). The volcano plot on the left illustrates the expression differences among the orthologous genes expressed in all eight cotton species, comparing A_2 and K_2 as an example The X-axis represents the $\log_2$-fold change in expression levels (FPKM) between A_2 and K_2. Red points represent genes upregulated in A_2, while blue points indicate genes downregulated in A_2. The Upset plot on the right summarizes the types and numbers of expressed genes in some cotton species. **b** The percentage of TE families in single-copy orthologous genes expressed in some cotton species (left) and in orthologous genes with differential expression levels (right). The number of TE families is indicated in parentheses. **c** Genome browser view displaying the expression levels of the *Garb_10G013290* (A_2) and *Grai_10G1013610* (D_5) genes and TE insertion events in the promoter regions. The values in square brackets represent the peak of RNA-seq expression signals (gray peak), with higher peaks indicating higher transcription levels. The yellow peaks represent the signals of ATAC-seq, with lower peaks indicating more closed and silent regions. DNA transposon TE0019589 insertion in the upstream 2 kb region of *Grai_10G1013610*. **d** The heatmap illustrates the enrichment of TF motifs in LTR transposons and other TE families across eight cotton species. The top clustered heatmaps provide a more intuitive comparison of TF motifs enrichment between LTR transposons and other TE families. **e** The expression levels of gene *Garb_09G019690* (A_2) and *Grai_01G025490* (D_5) are visualized along with TE insertions in their promoter regions. The values in square brackets represent the peak of RNA-seq expression signals (gray peak), with higher peaks indicating higher transcription levels. The yellow peaks represent the signals of ATAC-seq, with higher peaks indicating more open and active regions. **f** The box plot above shows the comparison of gene expression levels in the promoter regions of orthologous genes in cotton branches with and without specific LTR transposon insertions (Wilcoxon rank sum test, "ns" indicates not significant, $^{***}P < 2.2 \times 10^{-16}$). Dashed boxes represent no specific LTR transposon insertions, while solid boxes represent genes with specific LTR transposon insertions. The bar chart below shows the number of TE clusters in different categories. Cotton branches with TE-specific amplification are indicated at the bottom

their overall genome proportion ($P < 0.01$). This significant enrichment of DNA transposons in genes expressed only in some cotton species implies that they play a crucial role in regulating whether genes are transcribed. For example, in the gene family G0020242 encoding a peroxidase-like protein that influences plant development and

stress resistance, no expression signal was detected in *Grai_10G013610* (D_5) with DNA transposon insertion within 2 kb upstream. In contrast, the orthologous gene *Garb_10G013290* in A_2 exhibited a high expression level (Fig. 2c). This result has been validated through qRT-PCR (Additional file 2: Fig. S5a, Additional file 1: Table S7). Additionally, the ATAC-seq analysis revealed that the data peak at the upstream position of *Grai_10G013610* (D_5) where the DNA transposon TE0019589 was identified was remarkably low (Fig. 2c), suggesting that this region might be in a state of tight chromatin closure and silence. We may speculate that the DNA transposon insertion in the upstream 2 kb region potentially disrupted the promoter of *Grai_10G013610* (D_5) or possibly replaced its transcription start site (TSS), thereby impeding proper recognition and binding by RNA polymerase. This observation suggests a potential impact of the DNA transposon insertion on the expression of orthologous genes.

Moreover, the main type of TE with differential insertions in the differentially expressed orthologous genes is LTR retrotransposons, constituting 69.5% (Fig. 2b right). This indicates that LTR retrotransposons may play a role in regulating differences in gene transcription levels. It appears that different types of TE insertions have varying impacts on the expression of orthologous genes. Promoter regions are known to contain numerous regulatory elements, and TE insertions into these regions can influence gene expression. To further explore the impact of TE insertions on promoters, we identified TE-associated transcription factor-binding sites (TFBSs) in cotton. The heatmap illustrates the *Z*-scores of TFBS enrichments across various TE families, with LTR retrotransposons exhibiting significantly higher enrichment scores compared to other TE clusters (Fig. 2d). The LTR subfamily Gypsy exhibits the highest proportion of TFBSs among the LTR families, followed by Copia (Additional file 2: Fig. S6). The clustering analysis clearly differentiates between LTR retrotransposons and other TEs based on their TFBS enrichment profiles, highlighting the greater impact of LTR retrotransposons on gene regulation. The results indicate that LTR retrotransposons exhibit a higher enrichment of TFBSs compared to other types of TEs. Specifically, well-known regulators of gene expression such as MYB and WRKY are associated with LTR retrotransposons (Fig. 2d). Within the gene family OG0000794, which is involved in plant cell wall growth, the LTR retrotransposon TE1502079 inserted upstream of the *Garb_09G019690* (A_2) contains a MYB TFBS. The expression of *Garb_09G019690* was significantly upregulated compared to that of the orthologous gene in D_5 (Fig. 2e). The same results were obtained after qRT-PCR validation (Additional file 2: Fig. S5b, Additional file 1: Table S7). Additionally, the ATAC-seq peak at the upstream position of *Garb_09G019690* (A_2) where the LTR transposon TE1502079 is located exhibits a very high intensity (Fig. 2e), suggesting that this region may be in an open and accessible state. Therefore, LTR retrotransposons may regulate gene expression by influencing the binding of transcription factors.

Subsequently, we investigated the specific LTR retrotransposon insertions of orthologous genes on four major cotton branches (D_5, E_1, clade 1, and clade 2). When there were no specific LTR insertions on genes in each branch, their expression did not significantly differ; however, branches with specific LTR insertions exhibited significantly greater expression levels than did the other branches (Fig. 2f). Interestingly, in branches with specific LTR insertions, their genes contain more TEs that have undergone specific expansion in that branch. For instance, in clade 1 with specific LTR insertions, the most

abundant TE type in their genes is cluster 2 TE, which was also the TE family that has undergone specific expansion in clade 1 (Figs. 2f and 1b). This result suggests that specifically expanded TEs can regulate lineage divergence at the transcriptional level of orthologous genes by introducing specific LTR insertions, thereby affecting the evolution of cotton species. Similarly, comparing the expression levels of orthologous genes with and without specific DNA TE insertions on the four branches revealed that branches with specific DNA TE insertions had significantly lower expression levels (Additional file 2: Fig. S7). In conclusion, these results reveal the regulatory role of differential TE expansion at the transcriptional level in gene function and the divergence of cotton species, providing new evidence for our understanding of species evolution.

TEs regulate transcript isoforms of orthologous genes by affecting AS

Following the analysis of transcriptional level differences among genes in various cotton species, Nanopore direct RNA sequencing (DRS) was utilized to investigate the variations in the transcript structures of the aforementioned transcribed orthologous genes (Additional file 1: Table S8). DRS does not require reverse transcription and PCR amplification, and directly sequencing RNA strands, enabling the acquisition of more comprehensive and precise transcriptomic information, thereby unveiling the intricate structures and splicing patterns within transcripts. We generated full-length DRS data and identified transcript isoforms in eight species. A total of 45,075 (F_1)–57,274 (D_5) isoforms were transcribed from 22,538 (D_5) to 23,950 (F_1) genes across the species analyzed (Additional file 2: Fig. S8a). Unexpectedly, D_5 showed slightly more isoforms, but additional evidence should be provided to demonstrate whether this is related to the relatively high content of TEs in the gene regulatory regions (upstream and downstream 2 Kb) and gene body regions (Additional file 2: Fig. S9). Among these genes, 11,640 (K_2)–12,574 (G_1) orthologous genes were transcribed, with almost half of them transcribing multiple (≥ 2) isoforms in each species (Fig. 3a, Additional file 1: Table S9). Furthermore, there is a significant disparity in the number of isoforms among orthologous genes. Approximately 76% of these genes exhibit a coefficient of variation (CV) in isoform number exceeding 0.3 (Additional file 2: Fig. S8b). This indicates that there are differences in splicing patterns among orthologous genes, leading to the generation of diverse mRNA isoforms.

We next explored the putative role of TEs in regulating gene splicing. We conducted a correlation analysis between isoform number and TE enrichment score for 7196 orthologous genes with differential isoform number across 8 cotton species. We identified 291 genes with significant correlation (FDR < 0.05), including 183 positively correlated genes (Fig. 3b) and 108 negatively correlated genes (Fig. 3d). For instance, the gene family OG0015491, which regulates rRNA generation, transcribed 5 isoforms in *Garb_07G013780* (A_2) with a TE enrichment score of 0.93 and 3 isoforms in *Glon_07G013840* (F_1) with a score of 0.70 (Fig. 3c). Compared with F_1, the TE1548733 introduced a splicing site C in A_2, resulting in the splicing of a new isoform T64325, while the introduction of splicing sites C and G generated a new isoform T64323 (Fig. 3c). This demonstrates that certain TEs can introduce new splicing sites, leading to the production of a greater number of transcripts through AS. However, the gene family OG0014419, which regulates fatty acid synthesis, transcribed 4

isoforms in *Gsto_06G009650* (E_1) without TE insertions, while its orthologous gene *Garb_06G018610* in A_2 has a TE enrichment score of 0.52 but only transcribed a single short isoform (Fig. 3e). This is probably due to the TE insertion introducing a premature transcription termination site, resulting in the premature termination of transcription and the production of only one isoform. These results collectively illustrate the regulatory role of TEs on the number of transcripts isoforms.

Apart from affecting the number of transcripts, how do TEs influence the structure of transcripts? The AS events within transcripts are directly linked to the diversity

(See figure on next page.)

Fig. 3 TE mediates transcript isoform variation in cotton species. **a** Bar graph displaying the number of transcribed genes and the number of genes transcribing different numbers of isoforms across eight cotton species. Darker colors indicate a greater number of isoforms. **b** The Sankey diagram illustrates a positive correlation between the number of transcript isoforms and the TE enrichment scores of orthologous genes across eight cotton species. Arrange the values of the number of transcript isoforms and the TE enrichment scores of orthologous genes in ascending order (1–8) within the eight cotton species. Parallel lines in the Sankey diagram indicate that orthologous genes with higher numbers of transcript isoforms across the eight cotton species also exhibit higher TE enrichment scores. This suggests a significant positive correlation between the number of transcript isoforms and the TE enrichment scores of genes. The line graphs on the right display the correlation coefficients (cor_r > 0) and *q* value (FDR correction, Benjamini-Hochberg, $q < 0.05$), respectively. **c** The network diagram on the left displays all transcript isoforms transcribed from the gene family OG0015491 across 8 cotton species, including the number and names of the isoforms. The size of the isoform circles corresponds to the percent spliced in (PSI). The largest circle represents the major isoform, with the name highlighted in blue. The red numbers in the middle indicate the TE enrichment scores of this gene family in each cotton species. The higher the TE enrichment scores of orthologous genes, the greater the number of transcript isoforms transcribed. On the right illustrate the transcript isoforms of *Garb_07G013780* (A_2) and *Glon_07G013840* (F_1) two orthologous genes, along with the TE insertion sites. The major isoforms are shown in blue. The red dashed lines indicate the positions where TEs introduce new splicing sites. Circle the two transcripts that A_2 transcribed more than F_1 with a box. **d** The Sankey diagram illustrates genes that exhibit a negative correlation between the number of transcript isoforms and the TE enrichment scores. By sorting the values of the number of transcript isoforms and the TE enrichment scores of orthologous genes across the eight cotton species in ascending order (1–8), the intersecting lines in the Sankey diagram indicate that for orthologous genes with larger numbers of transcript isoforms across the eight cotton species, their TE enrichment scores are lower. This implies a significant negative correlation between the number of transcript isoforms and the TE enrichment scores of genes. The line graphs on the right display the correlation coefficients (cor_r < 0) and *q* value (FDR correction, Benjamini-Hochberg, $q < 0.05$), respectively. **e** The transcript isoforms of the orthologous genes *Garb_06G018610* (A_2) and *Gsto_06G009650* (E_1) were visualized using IGV (Integrative Genomics Viewer), along with the TE positions (red line segment) in *Garb_06G018610* (A_2). There is no TE inserted in *Gsto_06G009650* (E_1). **f** The types and numbers of alternative splicing (AS) events present in eight cotton species. The size of the circles represents the number of AS events, with larger circles representing a greater quantity. The schematic diagram in the center shows 7 types of splicing events: ES (exon skipping), A3 (alternative 3′ splice site), A5 (alternative 5′ splice site), IR (intron retention), ME (mutually exclusive exons), AP (alternate promoter), and AT (alternate terminator). **g** Violin plot comparing the distribution of AS event counts with and without TE insertions within each cotton species. Without TE (blue), TE insertion (red) (intraspecific comparison, Wilcoxon rank-sum test, with the *P* values annotated on the plot). **h** Distribution of AS event counts with and without TE insertions in orthologous single-copy genes among cotton species (interspecific comparison, Kruskal–Wallis rank sum test, with the *P* values annotated on the plot). **i** Identification of 32,752 sets of conserved isoforms across all genes in cotton species. The proportion of conserved isoforms from different numbers of cotton species are represented by different colored blocks. **j** Distribution of AS in the major isoforms of 8620 orthologous genes transcribed in all cotton species and their conservation levels across cotton species. The bottom bars 1–8 represent the numbers of conserved genes in 1–8 cotton species for the major isoforms. "AS" indicates the presence of alternative splicing events in the isoforms. The length of the bars is proportional to the number of genes. The boxplot on the right compares the TE enrichment scores between isoforms with and without alternative splicing events (paired *t*-test, $P = 2.1 \times 10^{-6}$). **k** The number of non-conserved major isoforms of orthologous genes with specific TE insertions in branches D_5, E_1, clade 1, and clade 2 compared to those in other cotton species. The solid line represents cotton species with specific TE insertions, while the dashed line represents cotton species without specific TE insertions. The number of non-conserved major isoforms is indicated within the corresponding figure

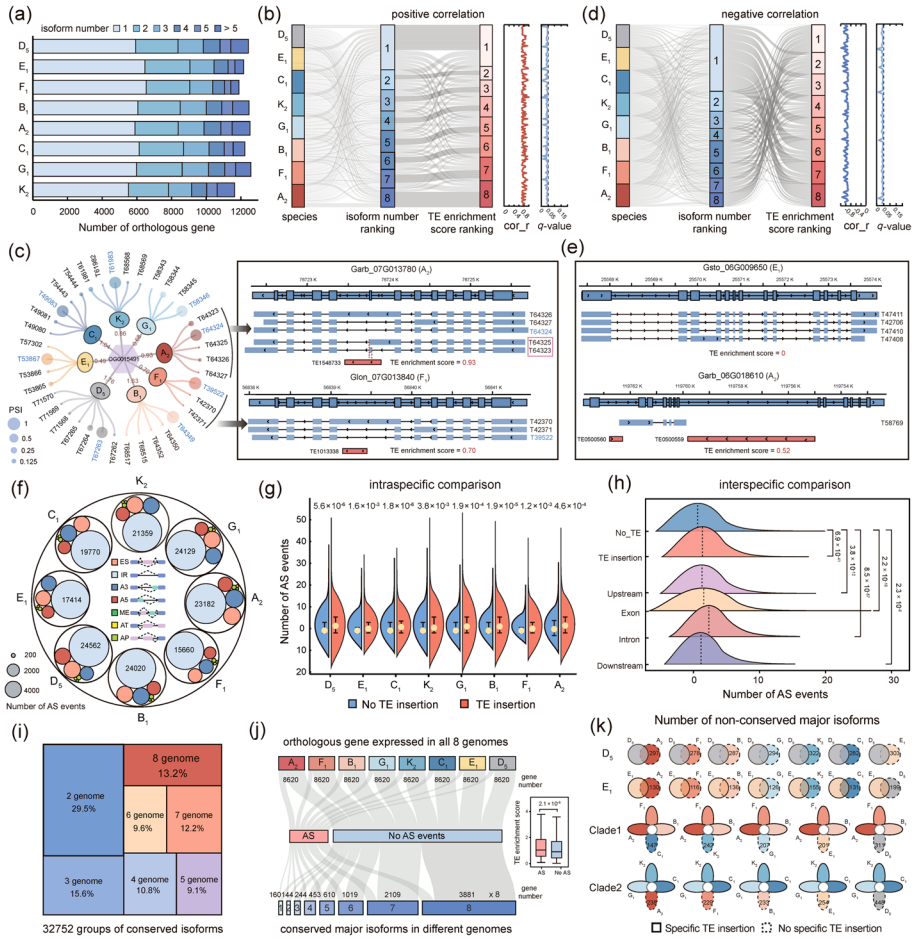

Fig. 3 (See legend on previous page.)

of transcript structure. We proceeded to identify AS events across 8 species, totaling 22,568 (F_1)–36,716 (A_2) AS events (Additional file 1: Table S10). Among these events, intron retention (IR) is the most abundant, accounting for 63.1% (A_2) to 71.5% (G_1) of all events (Fig. 3f, Additional file 1: Table S10), consistent with other species such as maize [43]. A comparative analysis of genes with and without TE insertions in each species revealed that genes with TE insertions had significantly more AS events ($P < 0.0016$) (Fig. 3g). Furthermore, among orthologous genes across cotton species, genes with TE insertions had significantly more AS events ($P = 6.9 \times 10^{-41}$) compared to those without TE insertions (Fig. 3h). Interestingly, by subdividing genes with TE insertions based on insertion positions, we found that genes with TE insertions in introns contained more AS events than those with insertions in upstream/downstream regions and exons (Fig. 3h).

We analyzed the level of transcript conservation among cotton species. Based on isoform sequence similarity, a total of 32,752 sets of conserved isoforms were identified across all genes. Among these, 4323 sets of isoforms were conserved in all eight species, accounting for only 13.2% (Fig. 3i). This indicates significant differences in the splicing patterns of genes. The conservation level of major isoforms among orthologous genes was slightly greater. Among the 8620 orthologous gene families transcribed

in all cotton species, 3881 (45%) had major isoforms conserved in all eight species, and these major isoforms mostly did not involve splicing events. However, the major isoforms with AS are conserved only in some (2–7) cotton species (Fig. 3j). This indicates that splicing events contribute to structural differences in major isoforms. Importantly, isoforms containing splicing events exhibited significantly greater TE enrichment scores ($P = 2.1 \times 10^{-6}$) compared to the major isoforms without splicing events (Fig. 3j). This further demonstrates that TE can mediate change of alternative splicing event, leading to structural differences in major isoforms of orthologous genes. Additionally, we identified the potential non-conserved major isoforms caused by TE difference expansion on four branches to investigate the impact of TE-mediated alternative splicing on the lineage differentiation of genes. In branches with specific TE insertions (D_5, E_1, clade 1, and clade 2), there were 278 (F_1)–322 (K_2), 116 (F_1)–199 (D_5), 147 (C_1)–311 (D_5), and 229 (F_1)–448 (D_5) genes with different major isoform structures (Fig. 3k). In summary, these results suggest that TEs can alter splice sites or splicing regulatory sequences, leading to changes in gene splicing patterns and subsequently influencing the transcript structural differentiation of orthologous genes. These observations contribute to further elucidating the posttranscriptional regulatory mechanisms underlying cotton species divergence.

TE-mediated uORF regulates translational differences of orthologous genes

Given that transcriptional differences do not fully elucidate the diversity and differentiation mechanisms among cotton species, the gene translation levels are more indicative of the actual protein levels. Therefore, we focused on differences at the translation level among orthologous genes. We constructed polysome profiling libraries for eight cotton species and quantified their translation levels (Additional file 1: Table S11). The correlation between the transcriptome and translatome data in each cotton species ranged from 0.73 to 0.90 (Fig. 4a, Additional file 2: Fig. S10), indicating a certain discrepancy between the transcriptional and translational levels. The coefficient of variation (CV) for transcriptional and translational levels within each cotton species (intraspecific comparison) showed that the CV for translational levels (CV2) was greater than that for transcriptional levels (CV1) (Fig. 4b). To ensure comparability between the two layers, we used FPKM-based expression values normalized across samples and species. Only coding regions were included in the analysis to avoid annotation bias [44]. This indicates that while the transcriptome and translatome show a global good correlation, there are some discrepancies between the transcriptional and translational levels within cotton species. We compared the differences (Δ) in transcriptional and translational level changes among orthologous genes across cotton species (interspecific comparison) and found that the translational level changes were greater ($\Delta > 0$) (Fig. 4c). A Δ value of 0 indicates equal evolutionary rates at both expression layers; Δ greater than 0 indicates a higher evolutionary rate at the translatome layer; and Δ less than 0 indicates a lower evolutionary rate at the translatome layer [44]. This finding also indicates that greater differences exist at the translational level between orthologous genes.

To investigate the post-transcriptional changes of orthologous gene regulation, we simultaneously compared the transcriptional and translational levels among orthologous genes. For example, compared to E_1, there were 1804 genes in D_5 with no difference

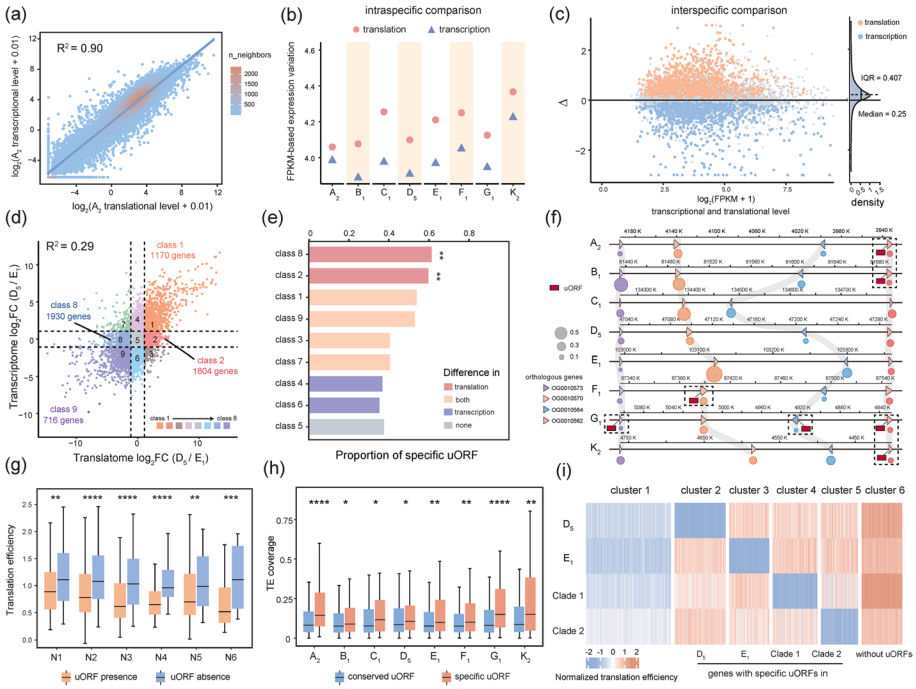

Fig. 4 TE regulates translation differences of orthologous genes in cotton. **a** Pairwise correlation (Spearman's ρ) between the transcriptional and translational levels of 18,676 orthologous genes in A_2. **b** Comparison of gene transcriptional level changes (coefficient of variation, CV1) and translational level changes (CV2) within each cotton species. **c** Differential changes in transcriptional and translational levels among orthologous genes across eight cotton species. The x-axis represents the expression levels (fragments per kilobase of transcript per million mapped reads, FPKM) of orthologous genes at the transcription and translation, which have been $\log_2$-transformed. Density distribution, median Δ, IQR (interquartile range) of Δ, and density of Δ significantly higher or lower than zero are shown on the right. $\Delta = 0$: equal evolutionary rates at both the transcription and translation layers; $\Delta > 0$: higher evolutionary rate at the translation (translatome) layer; $\Delta < 0$: lower evolutionary rate at the translation layer. **d** A quadrant plot displays the differences in transcription and translation between D_5 and E_1. The dashed lines correspond to the classification of genes into nine responsive groups based on both fold change ($|\log_2(\text{fold change})| \geq 1$) and q value ≤ 0.01. Class 1 (orange) represents genes in D_5 with both transcription and translation upregulation. Class 2 (red) represents genes in D_5 with no transcription difference but translation upregulation. Class 8 (dark blue) represents genes in D_5 with no transcription difference but translation downregulation. Class 9 (purple) represents genes in D_5 with both transcription and translation downregulation. **e** A bar graph shows the proportion of genes with specific uORFs in each gene category (classes 1–9). Classes 2 and 8 have the highest numerical values (Wilcoxon rank sum test, $^{**}P < 0.01$). **f** The translation efficiency based on the presence or absence of uORFs in four gene families among eight cotton species. Each color represents a different gene family, with larger circles indicating higher translation efficiency. **g** The box plot shows the comparison of the translation efficiency of orthologous genes with and without uORFs at different evolutionary nodes (Wilcoxon rank sum test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$). **h** Comparison of TE coverage between orthologous genes containing conserved uORFs and species-specific uORFs among cotton species (Wilcoxon rank sum test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$). **i** The clustered heatmap shows that genes with specific uORFs in the D_5, E_1, clade 1, and clade 2 branches exhibit lower translation efficiency. Cluster 6 represents genes that do not contain specific uORFs in any of the four branches. Clusters 2–5 correspond to genes with specific uORFs in D_5, E_1, clade 1, and clade 2, respectively

in transcription but upregulated translation (class 2) and 1930 genes with no difference in transcription but downregulated translation (class 8), which were the two most abundant types of differentially expressed genes (DEGs) (Fig. 4d). Similarly, in comparison with those in other species, class 2 and class 8 genes were also the most abundant (Additional file 2: Fig. S11, Additional file 1: Tables S12, S13). The presence of large

numbers of class 2 and class 8 genes suggests significant variability in gene expression levels after transcription, with some post-transcriptional regulatory factors playing crucial roles. Further analysis revealed that genes with differential translation but no difference in transcription levels contained more specific uORFs (upstream open reading frames) at the 5′ end (Fig. 4e, Additional file 1: Table S14). The conservation of uORFs among orthologous genes is very low, with only 7611 sets (28.5%) of uORF clusters being conserved across all 8 cotton species, while the remaining 19,090 sets (71.5%) of uORFs are not conserved with at least one species (Additional file 2: Fig. S12). During translation, uORFs may compete with mORFs (major open reading frame) for ribosome binding, thereby reducing the translation efficiency of mORFs [45, 46]. Therefore, we speculate that these specific uORFs regulate the differences in translation levels of orthologous genes after transcription. Investigations into four genes neighboring the orthologous gene family OG0010564, which encodes a WRKY transcription factor, demonstrated that genes containing specific uORFs exhibited lower gene translation efficiencies (Fig. 4f). For instance, OG0010573 in G_1, OG0010570 in F_1, OG0010564 in G_1, and OG0010562 in A_2, B_1, G_1, and K_2 contained specific uORFs, and their translation efficiencies were significantly lower than those without uORFs (Fig. 4f). Additionally, the translation efficiencies of orthologous genes with and without uORFs at different evolutionary nodes (N1–N6) were compared and found that genes with uORFs had significantly lower translation efficiencies (Fig. 4g). These results demonstrate that uORFs may regulate the translation efficiency differences of orthologous genes, thereby impacting cotton species evolution.

To explore whether differential TE expansion leads to uORF variations, the TE coverage in the 5′ regions of transcripts containing uORFs was calculated. The results revealed that genes with specific uORFs have greater TE coverage in their 5′ regions than conserved uORFs present in all cotton species (Fig. 4h). We found that TIR transposons contributed significantly more to uORF coverage compared to the Gypsy and Copia families (Additional file 2: Fig. S13). We hypothesized that differential TE expansion may regulate the translation efficiency of orthologous genes by introducing specific uORFs, thereby mediating species divergence. For example, the gene *Garb_03G025400* (A_2) that responds to light stimulation contains a specific uORF introduced by a TE at the 5′ end, leading to significantly reduced translation levels compared those of to the orthologous gene *Grai_03G004290* in D_5 (Additional file 2: Fig. S14). Subsequently, we compared the translational efficiencies of genes with and without specific uORFs introduced by differential TE expansion across the four branches (D_5, E_1, clade 1, clade 2). The clustered heatmap revealed that orthologous genes with uORFs in all four branches (cluster 1) showed lower translation efficiencies. The translation efficiencies of genes containing specific uORFs in D_5 (cluster 2), E_1 (cluster 3), clade 1 (cluster 4), and clade 2 (cluster 5) were significantly lower than in the other three branches. Conversely, genes in all four branches lacking uORFs (cluster 6) exhibited higher translation efficiencies. In summary, genes in branches with specific uORFs had significantly lower translation efficiencies than those in other branches (Fig. 4i). This result indicates that differential TE expansion introduced specific uORFs and contributed to differences in gene translation levels and lineage divergence in cotton. GO enrichment analysis of genes with differential translation in the four branches showed that the genes in D_5 were mainly

enriched in pathways such as cellular response to stimulus and cellular nitrogen compound metabolic processes. The genes in clade 1 were enriched in macromolecule catabolic processes and protein transport pathways, and genes in clade 2 were enriched in intracellular signal transduction and cytoskeletal protein binding (Additional file 2: Fig. S15) [41]. These findings provide insights into understanding the posttranscriptional regulation evolution of gene expression among cotton species.

TE-derived miRNA regulates translation differences in transcripts

Structural variations in transcript isoforms may affect the effective binding of regulatory factors, thereby influencing translation efficiency [47–49]. Analysis of translational efficiency differences at the transcript isoforms level can provide deeper insights into the detailed mechanisms of posttranscriptional regulation. The actions of small RNAs on mRNA degradation and translation inhibition are well-known posttranscriptional regulatory mechanisms [50, 51]. To investigate the regulatory roles of small RNAs in transcript translation, we performed small RNA sequencing for each cotton species (Additional file 1: Table S15). A total of 315,936–947,549 small RNA loci were identified (Additional file 2: Fig. S16). It is found that some miRNAs and siRNAs are shared among different cotton species (Additional file 2: Fig. S17). Given the large number of siRNAs making analysis cumbersome, we focused on miRNAs, which have 193 (K_2)–249 (G_1) loci (Additional file 1: Table S16). We predicted their targets and constructed a network of miRNA-regulated target genes (Additional file 2: Fig. S18a). We observed that miRNA targets were not conserved in some orthologous genes, with some cotton species showing an absence of miRNA target sites. For example, the target of MIR477 is present only in the Australian cotton but is absent in other cotton species (Additional file 2: Fig. S18b). This implies that miRNAs may contribute to the expression differences among orthologous genes in cotton species. Subsequently, we counted the numbers of conserved and non-conserved targets at each evolutionary node (Fig. 5a). The translation efficiencies of miRNA target and nontarget genes were compared, revealing significantly lower translation efficiency of miRNA target genes within each species and between orthologous genes of different species (Fig. 5b, Additional file 2: Fig. S19). These results indicate that miRNAs indeed exert translational inhibition on cotton target isoforms. As expected, the conservation level of these miRNA target sites among orthologous genes was relatively low. For example, the ancient superfamily miR482 and miR2118 exhibit target site specificity of 63.8% and 53.4% in the orthologous genes, respectively (Fig. 5c, Additional file 1: Tables S17, S18). This finding suggests that miRNAs can mediate the expression differences by regulating the translation efficiency of genes.

To investigate the role of TEs in this context, we quantified the contributions of different TE families to the evolution of miRNA target sites. We found a significant enrichment of miRNA target sites within LTR retrotransposons, particularly at the insertion sites of *Gypsy* transposons (Fig. 5d, Additional file 1: Table S19). This observation indicates that TE expansion could be one of the factors contributing to the increase in miRNA target sites, thereby influencing the regulation of gene translation. Interestingly, the translation efficiencies of genes with miRNA target sites decreased, especially as TE enrichment increased, in comparison to those of nontarget genes (Fig. 5e). This

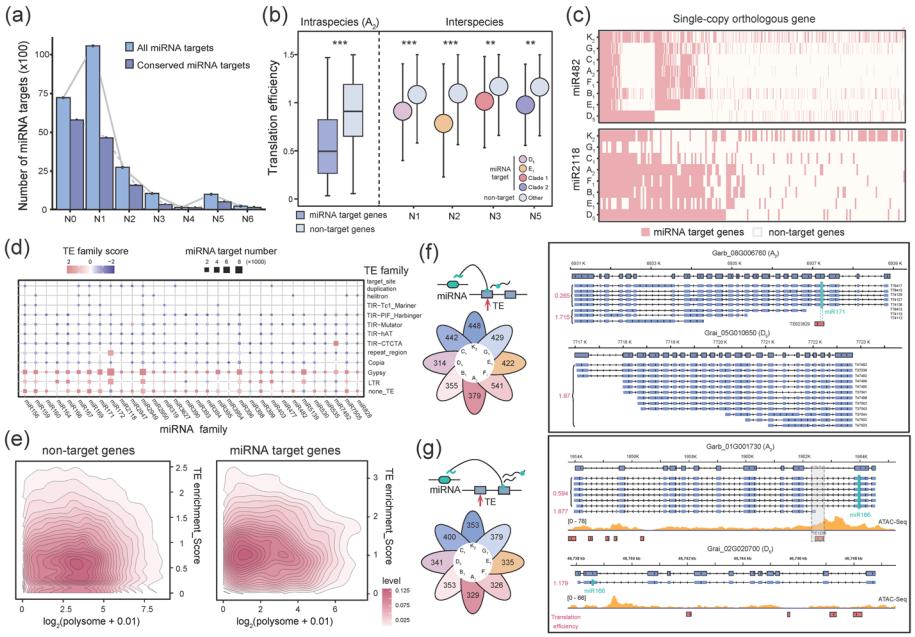

Fig. 5 The mechanism of small RNA regulation on the translation differences of transcripts. **a** The figure shows the number of total and conserved miRNA target genes across seven evolutionary nodes. **b** Comparison of the translation efficiency of genes with and without miRNA targets is displayed. The boxplot on the left compares various genes within A_2, while the bubble plot on the right compares orthologous genes across 4 branches of different cotton species. Statistical significance was determined using a two-sided Wilcoxon rank-sum test, with precise *P* values shown (***$P < 2.2 \times 10^{-16}$). **c** The heatmap illustrates the distribution of targets for the miR482 and miR2118 miRNA families in single-copy orthologous genes across 8 cotton species, where red indicates the presence of targets and white indicates the absence of targets. **d** Enrichment scores (*z*-scores) of different TE families in miRNA target genes. The depth of color indicates the magnitude of the enrichment score, while the size of the squares represents the number of target genes for different miRNA families. **e** The relationship between the enrichment level of TE and translational abundance for non-target genes (left) and target genes (right) within the cotton lineage. **f** MiRNAs derived from TEs reduce the translation efficiency of transcripts containing target sites. In *Garb_08G006760* (A_2), the bottom three transcripts contain miR171 targets, and their translation efficiency is lower than that of non-target transcripts. The miR171 site is located on TE623829. **g** TEs lead to the production of transcript isoforms through alternative splicing that lack miRNA targets, resulting in increased translation efficiency. For instance, in *Garb_01G001730* (A_2), TE1229 leads to generate an isoform lacking miR166 targets, which exhibits significantly higher translation efficiency compared to other target isoforms. The yellow peaks represent the signals of ATAC-seq, with higher peaks indicating more open and active regions

further demonstrates that TEs can regulate gene translation through miRNA-mediated mechanisms.

Upon closer observation, it was found that TEs primarily affect the regulation of miRNA-mediated isoform translation differences through two mechanisms. First, TE-derived miRNAs can reduce the translation efficiencies of isoforms containing target sites, while nontarget isoforms undergo normal translation. For example, in *Garb_08G006760* (A_2), three nontarget isoforms exhibited greater translation efficiencies, whereas the isoform containing the TE-derived miR171 target site had a translation efficiency of only 0.265, resulting in a lower translation efficiency of the entire *Garb_08G006760* gene compared to its orthologous gene in D_5 (Fig. 5f). The second mechanism involves TEs introducing new splicing sites in miRNA target genes, which leads to the emergence of isoforms that evade miRNA regulation due to the absence of

miRNA target sites. This leads to an increase in the translation efficiency of that isoform. For example, the splicing sites on *Garb_01G001730* (A_2) originating from TE1229 demonstrate a higher level of chromatin accessibility (Fig. 5g, Additional file 2: Fig. S20a). Additionally, the consensus splice motifs of TE1229-derived splice sites align with the canonical splice site motifs (Additional file 2: Fig. S20b). Therefore, we speculate that TE1229 might introduce a new splicing site in *Garb_01G001730* in A_2, potentially leading to the transcription of an isoform lacking the miR166 target site. This isoform exhibited significantly greater translation efficiency than other target isoforms (Fig. 5g). Notably, both of these mechanisms result in nontarget transcripts becoming new major transcripts, leading to possible structural and functional alterations in the proteins encoded by orthologous genes. We identified the number of genes regulated by these two mechanisms in each cotton species, ranging from 314 (D_5) to 541 (F_1) and 326 (F_1) to 400 (C_1), respectively (Fig. 5f, g). We further investigated the impact of TE-derived miRNAs on cotton lineage divergence. By comparing the translation efficiencies of genes with and without TE-derived miRNA targets across four branches, we found that branches with miRNA targets exhibited significantly lower translation efficiencies (Additional file 2: Fig. S21). These results indicate that TE-derived miRNAs mediate the lineage divergence of gene translation in cotton, and provide clues for exploring the role of TEs in driving the evolution.

Phylogenetic insight into posttranscriptional evolution in the *Gossypium* genus

To delve deeper into gene expression regulation during cotton species divergence, we generated expression profiles of orthologous genes from four lineages (D_5, E_1, clade 1, clade 2) at three levels (transcription, splicing, translation) and found numerous lineage-specifically expressed genes through K-means clustering. It is noteworthy that there were variations in lineage-specific expression of orthologous genes across the levels of transcription, splicing, and translation (Fig. 6a). Specifically, there were distinct regulatory divergences of orthologous genes among different lineages. For instance, the orthologous gene OG0012439 exhibited clade 1 lineage specificity at the transcriptional level, while showing clade 2 lineage specificity at the translational level (Fig. 6b). A total of 68 orthologous genes display this pattern. These findings indicate a dynamic trend in gene expression regulation during cotton divergence, involving a substantial number of genes. Among the genes exhibiting clade 1 lineage-specific expression at the transcriptional level, 72.1% showed changes in lineage-specific expression at both the splicing and translation levels, while only 2.3% remained unchanged (Additional file 2: Fig. S22). These results reveal the high flexibility and plasticity of gene expression regulation during evolution, providing a fundamental guarantee for organisms to adapt to diverse environments and growth requirements.

To comprehensively explore the expression divergence of orthologous genes during the evolution of cotton species, we focused on key evolutionary time points in the cotton evolutionary process. Based on phylogenetic analysis, the evolution of cotton species was divided into four stages: an undifferentiated ancestral stage of eight cotton species (stage 1), the D_5 differentiation stage (stage 2), the E_1 differentiation stage (stage 3), and the clade 1 and clade 2 differentiation stage (stage 4) (Fig. 6c left). We meticulously

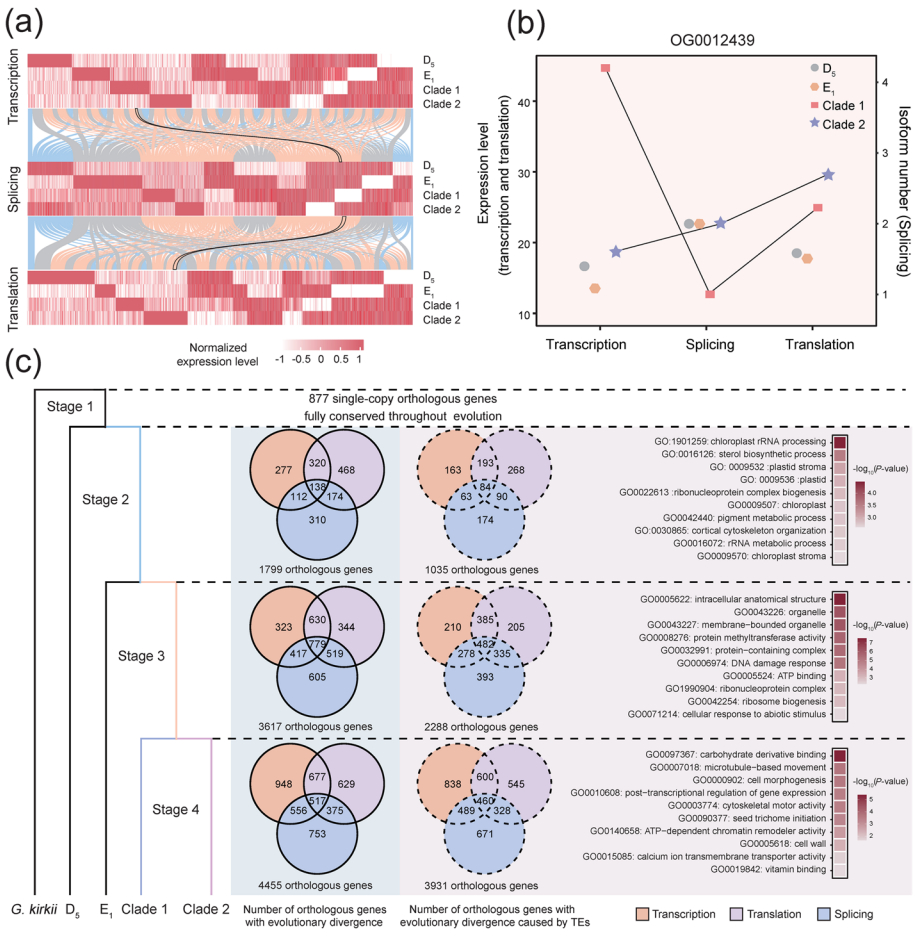

Fig. 6 Expression divergence of orthologous genes in cotton evolution. **a** Heatmap illustrates the expression levels of orthologous genes in different lineages (D_5, E_1, clade 1, clade 2) at three levels (transcription, splicing, and translation). The expression level is represented by color intensity. The lineage-specific expression dynamics of the example gene family OG0012439 across three levels are depicted with black lines. **b** The scatter plot illustrates the expression levels of the gene family OG0012439 across three levels in four lineages. Transcription and translation expression levels are quantified using FPKM values, while splicing levels are represented by the number of isoforms. Black lines are used to highlight the expression trends across three levels in clade 1 and clade 2. **c** Venn diagrams and GO enrichment analysis illustrating the conservation and divergence of orthologous genes across four key evolutionary stages within *Gossypium*. The diagrams categorize genes into universally conserved orthologs (stage 1) and those showing divergence specific to evolutionary stages post-D_5 (stage 2) and post-E_1 (stage 3) differentiation, as well as between clade 1 and clade 2 (stage 4). The right panel highlights significant GO terms associated with differentially expressed genes at each stage, suggesting adaptive functional shifts corresponding to the evolutionary history of the cotton genus

identified DEGs at the three levels present in these four evolutionary stages. Specifically, there were 877 conserved gene families across all eight cotton species (stage 1), 1799 gene families with lineage-specific expression in the D_5 (stage 2), 3617 gene families with lineage-specific expression in the E_1 (stage 3), and 4455 gene families exhibiting significant expression differences between clade 1 and clade 2 (stage 4) (Fig. 6c middle). These DEGs are key factors driving the divergence of cotton lineages. We further quantified the regulatory role of TEs on the DEGs and documented the potential number of genes with TE-mediated differential expression in the four evolutionary stages. The expression divergences of 1035 genes (stage 2), 2288 genes (stage 3), and 3931 genes (stage 4) were

found to be caused by differential TE insertions, accounting for 57.5%, 63.3%, and 76.1% of the DEGs, respectively (Fig. 6c middle).

Finally, we conducted GO enrichment analysis on these divergent genes and found that the functional enrichment patterns of DEGs affected by TE varied significantly across different evolutionary stages. In stage 2, genes with D_5-specific divergence were predominantly enriched in processes related to pigment metabolic process and sterol biosynthetic process, reflecting the differentiation of secondary metabolic pathways. In stage 3, the enrichment of E_1 lineage-specific divergent genes in cellular response to abiotic stimulus and protein methyltransferase activity implies a differentiation towards stress resistance. Genes with divergence between clade 1 and clade 2 at stage 4 were enriched in seed trichome initiation, cell wall, and cell morphogenesis (Fig. 6c right). Importantly, A_2 (*G. arboreum*) in clade 1 can produce spinnable cotton fibers. In contrast, K_2 (*G. rotundifolium*) and C_1 (*G. sturtianum*) in clade 2 lack fibers or have very short fibers [52], possibly implying differentiation in cotton fiber evolution. In conclusion, the results of this study provide valuable insights into deciphering posttranscriptional regulatory mechanisms in species evolution.

Discussion

With the advancement of functional studies on TEs, an increasing body of evidence has shown that TEs have profound impacts on the regulatory networks of both animals [53] and plants [54, 55]. Research on transposon polymorphisms in tomato domestication highlights that transposon insertion polymorphisms (TIPs) are significant sources of plant phenotypic variation. It reports that transposable elements inserted into gene regions have a substantial impact on gene transcription, leading to the production of multiple transcripts [56]. In wheat, it has been found that a significant number of distal regulatory elements derived from transposable elements influence the transcriptional regulation of subgenomes, thereby leading to phenotypic diversity such as spike morphology [57]. In studies of mouse and *Drosophila*, it has also been reported that TE influences transcription by generating new promoters, enhancers, or alternative splicing sites, thereby promoting species evolution and diversity formation [53, 58, 59]. These are consistent with our study on TE-mediated transcriptional regulation promoting phenotypic diversity in cotton.

Previous studies have assembled eight high-quality reference genomes of the genus *Gossypium*, laying the foundation for an in-depth exploration of lineage-specific TE expansion and its differential regulatory effects on orthologous genes between cotton species, as well as its role in species divergence and adaptive evolution [39, 41]. Here, leveraging these high-quality genome assemblies, we integrated transcriptomic, translatomic, and small RNA sequencing data from eight cotton species to systematically characterize the TE insertions and assess how their activity and regulation affect the evolution of the *Gossypium* genome. Phylogenetic approaches revealed that over 80% of TE families exhibited lineage and species-specific expansion. Concurrently, we discovered that different transposon families exhibited diverse gene expression regulatory mechanisms. LTR transposons may facilitate gene expression by increasing transcription factor-binding sites, while DNA transposon insertions may disrupt cis-regulatory

elements, affecting the normal expression of genes. Furthermore, the differential expansion of TEs significantly impacts the translation of orthologous genes. Lineage-specific TE insertions can regulate the translation levels of orthologous genes by introducing new uORFs or affecting mRNA stability and translation efficiency. Despite the turnover of TEs among genes, the relative distance between certain TE families and genes surprisingly remains constant, suggesting that some TE families may have insertion preferences relative to genes [60]. Lineage-specifically amplified TE insertions are significantly enriched near genes that are lineage-specifically expressed. The impact of TEs on gene-level transcription and translation further emphasizes their role in gene expression regulation, highlighting the key role of TEs in promoting species divergence and adaptive evolution in cotton.

Thanks to the continuous advancement of sequencing technologies, third-generation full-length transcriptome sequencing technology has provided us with the opportunity to better characterize the transcriptomes of cotton species, despite the inherent limitations of long-read technologies. Transcript-level analyses have revealed how differential TE expansion affects the expression patterns and functions of orthologous genes between cotton species through various aspects, including mRNA processing, stability, and translation regulation. The location and type of TE insertions have significant effects on orthologous gene alternative splicing and mRNA stability, possibly leading to cotton species-specific transcript expression patterns. Moreover, studies have shown that small RNA targeting is one of the important silencing mechanisms against TEs [61], and through affecting the targeting of miRNAs, TE expansion may intervene in posttranscriptional regulatory processes, further achieving fine control of gene expression. The combined effects of these mechanisms provide new molecular insights into cotton species divergence. These findings deepen our understanding of the regulatory mechanisms of orthologous gene expression in *Gossypium* and offer an important molecular basis for future crop genetic improvement and conservation. By conducting a comprehensive analysis of the regulatory mechanisms of orthologous gene expression, this study demonstrates the multidimensional role of TE expansion in plant adaptive evolution and interspecies differentiation, providing valuable insights for further studies on the dynamic changes and evolutionary processes of plant genomes.

Conclusions

Leveraging multi-omics data such as direct RNA sequencing, polysome profiling-seq, and small RNA-seq, the role of transposable element-mediated post-transcriptional regulation in driving the divergence of eight diploid cotton species was analyzed. It was found that transposable element amplification would lead to changes in splicing sites and regulatory sequences, thereby altering the alternative splicing patterns and expression levels of orthologous genes; Regulatory elements such as uORF and small RNA derived from transposable elements mediated the differences in the translation levels of orthologous genes. We identified genes exhibiting lineage-specific divergence at the levels of transcription, splicing, and translation, and further investigated the expression regulation divergence of orthologous genes during the evolution of cotton species. This

study provides insights into the evolutionary mechanisms of post-transcriptional regulation and biodiversity formation of cotton species.

Methods

Plant materials

The eight diploid cotton species used in this study included *G. arboreum* ("A_2," accession Shixiya-1), *G. anomalum* ("B_1," accession GPB1lz), *G. sturtianum* ("C_1," accession GPC1lz), *G. raimondii* ("D_5," accession GPD5lz), *G. stocksii* ("E_1," accession GPE1lz), *G. longicalyx* ("F_1," accession GPF1lz), *G. bickii* ("G_1," accession GPG1lz), and *G. rotundifolium* ("K_2," accession GPK201). These germplasm resources were obtained from the National Wild Cotton Nursery in Hainan and were cultivated in the greenhouse of Huazhong Agricultural University in Wuhan, China. The young leaves were immediately placed in liquid nitrogen for rapid freezing and stored in an ultralow temperature freezer at − 80 °C for later use.

TE re-annotation and analysis

Genomes used in this study were assembled in our previous work [41]. In prior research, preliminary annotation of the genomes was performed using RepeatMasker [62]. However, the complex TE activity in cotton species posed challenges for analyzing transposable element activity across species. To address this issue, we first merged the species-specific TE libraries obtained. Since RepeatMasker can provide overlapping annotations, we used bedtools merge to combine overlapping annotations, generating chimeric sequences. To reduce redundancy in the merged TE library, we employed CD-HIT2 for two rounds of clustering. In the first round, CD-HIT2 was used to group redundant sequences, while in the second round, representative sequences were manually selected based on sequence length, self-identity, and the presence of full-length TE insertions. After two rounds of clustering following the above steps, we removed redundant sequences using the "cleanup_nested.pl" script provided by Extensive de novo TE Annotator (EDTA v2.0.1) [63], resulting in a non-redundant TE library specific to cotton species. Based on the generated TE library, structural and fragmented TEs in each genome were annotated using EDTA. By integrating homology-based annotations (RepeatMasker) and structure-based annotations (EDTA), a comprehensive TE annotation for each genome was created. Family classification of TEs identified in structural annotations was based on the 80–80-80 rule, where TE sequences were considered to belong to the same family if they had at least 80% similarity over 80% of the sequence length, as first described by Wicker et al. [64].

Non-reference genome alignment

The nonredundant TE library constructed in this study was input into RepeatMasker to soft-mask all the genomes, followed by non-reference genome alignment using cactus (v2.6.0) [65]. MAF alignment files were organized using hal2maf with specific parameters. Low-quality alignment regions were filtered using trimAl (v1.4. rev 22) (https://github.com/inab/trimal).

Phylogenetic analysis and gene family evolution

A phylogenetic tree comprising the 12 diploid *Gossypium* species and *Gossypioides kirkii* (outgroup) was constructed using RAxML with the maximum likelihood method [66]. The Ks values of orthologous gene pairs were calculated via MCScanX downstream analyses. Orthologous gene pairs between *G. kirkii*/*Gossypium* and *Gossypium* were inferred by one-to-one alignment with a BLASTP E value cutoff of 1×10^{-10}. The whole-genome duplication and evolutionary time of species speciation were calculated using the formula T = Ks/2r, where r is the synonymous mutation rate for *Gossypium* species (3.48×10^{-9}) as described previously [67]. To identify orthogroups among the eight diploid cotton species, we used OrthoFinder (v 2.3.8) [68]. Octad genes, representing one-to-one corresponding orthologs across the eight genomes, were classified into two categories: single-copy orthologous genes and orthologous genes with variable copy numbers in some genomes. For the latter, we selected the gene with the longest sequence on the same chromosome.

Cross-species TEs and enrichment score calculation

To identify conserved transposable elements (TEs) among cotton species, we first utilized previously annotated TEs and conducted a comparison with homologous blocks between cotton species. Using the software bedtools, we aligned the coordinates of the TEs to the homologous sequence blocks. If TEs from different cotton species were located within the same homologous block and belonged to the same TE family, we classified this as a conserved TE insertion event. Based on this method, we constructed a map of conserved TE insertions among cotton species. To evaluate the enrichment of TEs within specific genes, we calculated the TE enrichment score. This score is determined by normalizing the transposon density of a gene by the average transposon density across all genes. The transposon density of a gene is calculated as the number of transposon insertions divided by the length of the gene. The average transposon density is calculated by averaging the transposon densities of all genes. The TE enrichment score is then given by the following formula:

$$\text{TE Enrichment Score} = \frac{\left(\text{transposon count}/\text{gene length}\right)}{\left(\sum_{i=1}^{n} \text{transposon density}_i\right)/n}$$

RNA-seq and polysome profiling-seq experimental processing

Around 0.1 g of leaves was rapidly ground into powder in liquid nitrogen. RNA was extracted using a rep Pure Plant Plus Kit (Polysaccharides & Polyphenolics-rich) (DP441, TIANGEN, Beijing, China), with two biological replicates per assay. Subsequently, libraries were constructed using an Illumina TruSeq Stranded RNA Kit (Illumina, San Diego, CA, USA), and sequencing was performed on the MGI2000 system.

Approximately 0.5–1 g of leaves was rapidly ground into powder in liquid nitrogen. Ribosome complexes were extracted using the method as previously described [15], with two biological replicates per assay. Sucrose solution containing ribosome-RNA complexes (monosome and polysome) were collected. RNA was extracted from each sucrose solution. A Ribo-off rRNA depletion kit (N409, Vazyme, Nanjing, China) was used to remove rRNA (5S, 18S, and 25S rRNA) from the total RNA, preserving the mRNA and

other noncoding RNAs. Subsequently, libraries were constructed using the VAHTS Universal V8 RNA-seq Library Prep Kit for Illumina (NR605, Vazyme, Nanjing, China), and sequencing was performed on the MGI2000 system.

RNA-seq and polysome profiling-seq analysis

The Illumina RNA-seq and polysome profiling-seq data from eight cotton species were subjected to quality control to remove low-quality reads and adapters using Trimmomatic (v0.36) with the parameters set to ILLUMINACLIP:TruSeq3-SE:2:30:10 to remove sequencing adapters and low-quality reads [69]. For all samples, reference genomes assembled from previous studies were used for the preliminary mapping of trimmed reads [41]. The filtered reads were then aligned separately to their respective genomes using HISAT2 (v2.2.1) with default parameters [70]. Gene transcription and translation expression levels were quantified using StringTie (v2.1.4) based on uniquely mapped reads [71]. Transcript transcription and translation expression levels were estimated using Salmon (v1.5.2) with reference to previously assembled transcript annotation files [72]. To aggregate the expression levels from the transcript to the gene level, we used the R tximport package to process the TPM (fragments per kilobase of transcript per million mapped reads) values obtained from Salmon. Only genes with transcription and translation level TPM values ≥ 1 were considered for calculating pairwise Pearson correlation coefficients (R^2) between replicates. Translation efficiency was determined by the ratio of TPM (translation level) to TPM (transcript level) for genes with TPM ≥ 1 at both levels. The $\log_2$-transformed translation efficiency values were used to visualize the distribution of the translation efficiency of the genes.

Direct RNA sequencing (DRS) experimental processing and data analysis

Approximately 0.1 g of leaves was rapidly ground into powder in liquid nitrogen, and an RNAprep Pure Plant Plus Kit (Polysaccharides & Polyphenolics-rich) (DP441, TIANGEN, Beijing, China) was used for RNA extraction. Libraries were constructed using a Direct RNA Sequencing Kit (SQK-RNA002, Oxford Nanopore Technologies, Oxford, UK), and sequencing was performed on the PromethION 48 (P48) system.

The FAST5 files of the raw reads were basecalled with ONT Guppy v3.1.5, and the basecalled reads were saved in FASTQ format. Postbasecalling quality control was performed with NanoFilt (v2.8.0) [73] to verify the consistency of the sequencing runs. Clean reads were aligned to the reference genomes of the corresponding species using minimap2 [74] in spliced alignment mode using a kmer size of 14 and a maximum intron size of 10,000 nt. Sequence Alignment/Map (SAM) and BAM file manipulations were performed using samtools version 1.9 [75]. FLAIR correction (v1.7) was utilized to refine the splice site boundaries of the reads [76]. The validity of all splice sites was evaluated based on reference genome annotations and Illumina RNA-seq data support. Splice junctions were extracted from long-read alignment data using the junctions_from_sam script in FLAIR, and only those supported by at least three uniquely mapped reads were considered valid.

Transcriptome assembly

Transcriptome assembly for each cotton species was conducted by integrating previously assembled genomes and short-read sequencing data with Nanopore direct RNA sequencing data. FLAIR was run with default settings to obtain transcript annotations. The identified transcripts were compared with reference genome-annotated transcripts using gffcompare (https://github.com/gpertea/gffcompare) to determine the relationship between each FLAIR transcript and its most similar reference transcript. Only transcripts supported by at least three reads were retained as final transcripts. Transcript sequences in FASTA format were extracted using Gffread (https://github.com/gpertea/gffread), and the FASTA files generated for each species, including novel and known transcripts, were used as transcript references for further quantification.

Alternative splicing (AS) event calling, filtering, and PSI calculation

AS event type analysis was performed using SUPPA2 [77] on the merged GTF annotation files obtained. SUPPA2 can generate seven types of AS events: A5, A3, IR, ES, ME, AT, and AP. For each AS event within genes describing any event type, ioe format files were created. Specifically, ioe files provide the transcripts that contribute to the numerator (one form of the event) and the denominator (both forms of the event) for PSI calculation. The pool genes function of SUPPA2 was used to cluster overlapping transcripts that share a substantial amount of sequence, thus considering relative splicing events. PSI calculation was performed based on TPM values and each event in the ioe files. Bar plots and circular stacked bar plots depicting the number of different event types were created using the R tidyverse and ggplot2 packages (https://www.tidyverse.org/).

Identification of conserved AS/isoform through chain file

To identify conserved AS events/Isoform among cotton species, we established pairwise chain files between each pair of cotton species to identify conserved exons. The chain files were generated by modifying the workflow from the UCSC Genome Browser (https://github.com/ENCODE-DCC/kentUtils/blob/master/src/hg/utils/automation/doBlastzChainNet.pl). The parameters used for lastz in this workflow were BLASTZ_H = 3000, BLASTZ_M = 254, BLASTZ_O = 400, BLASTZ_E = 30, BLASTZ_K = 3000, BLASTZ_Y = 3000, and BLASTZ_T = 1. The Lift-Over tool [78] was used to select conserved AS events/Isoform by identifying 1:1 overlaps between the pairwise chain files.

Lineage divergence of orthologous genes

The transcription, splicing, and translation specificity among orthologous genes in the lineages of the four cotton species were determined, considering all sample replicates. K-means clustering was performed to group similar orthologous genes based on transcriptional expression levels, PSI values, and translational expression levels. Specific lineage orthologous gene clusters were identified through correlation analysis using the R cor function. The *P* values for the correlations were calculated using the R cor.test function. The R ComplexHeatmap package was used to cluster all specific orthologous gene clusters together to create a heatmap, with *Z*-score transformation

performed prior to heatmap plotting. To determine the functions of these genes, we annotated and enriched the corresponding genes through Gene Ontology (GO) annotation and enrichment analysis of transcripts. GO annotations were obtained from previous studies. The R clusterProfiler package was used for GO enrichment of each category of orthologous genes.

Promoter region motif analysis

In each genome, the region upstream of the transcription start site (TSS) within 2 kilobases (2 kb) was used to predict motifs using the findMotifs.pl program from HOMER (v5.0) software, with the parameters "-len 8,10,12 -size 200." The predicted motifs were filtered based on assumed cutoff values of known motif enrichment ($P \leq 0.01$) and de novo motif prediction ($P \leq 1 \times 10^{-10}$) [79]. The motif file was obtained from ChIP-seq or DAP-seq experiments in *Arabidopsis*, rice, and other plants. Only TFBSs with motif scores above 10 were retained. A total of 506 plant TF motifs are contained in the HOMER database, and only 262 TF motifs were related to TFBS after removing unknown motifs.

Identification of putative canonical uORFs

Based on the GFF3 files and cDNA sequences of the eight cotton species, we identified potential uORFs within annotated 5′ untranslated regions (UTRs) of protein-coding genes using a custom Python script. These uORFs started with an AUG codon and ended with a stop codon (UAA/UAG/UGA). uORFs in which the start codon overlapped with the coding sequence (CDS) of other transcripts were excluded from the analysis. Only uORFs supported by polysome data were retained for subsequent analysis [83–88].

Small RNA-seq and data analysis

Around 0.1 g of leaves was rapidly ground into powder in liquid nitrogen, and total RNA extraction was essentially carried out as described previously using the guanidinium isothiocyanate method [80]. Libraries were constructed using an MGIEasy Small RNA Library Prep Kit (1000005269, MGI, Shenzhen, China), and sequencing was performed on the BGISEQ-500 system. Quality-controlled small RNA sequencing data were subjected to the removal of rRNA, tRNA, and other noncoding RNA sequences. The remaining clean reads were analyzed using sRNAminer [81] to construct a comprehensive small RNA atlas. miRNAs were identified based on their size (21–22 nt), precursor structures (hairpin formation), and homology to known miRNA sequences, as determined through sRNAminer's integrated filtering and annotation functions. In contrast, siRNAs, including heterochromatic siRNAs (hc-siRNAs), were identified based on their sequence length (21–24 nt) and repetitive, transposon-derived genomic loci. These siRNAs were distinguished from miRNAs through mapping to repetitive regions and using coverage analysis of the loci to determine their origins from double-stranded RNA precursors. MiRNA target genes were identified using psRobot and sRNAminer with a sequence matching threshold of 3.0 [82]. The miRNA-mRNA interaction network was visualized using Cytoscape software [83].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03534-5.

Additional file 1. Supplementary tables S1–S19.

Additional file 2. Supplementary figures S1–S22.

Acknowledgements
We thank the National Wild Cotton Nursery (Sanya, China) for providing seeds of wild diploid cotton species. We thank the high-performance computing platform at the National Key Laboratory of Crop Genetic Improvement in Huazhong Agricultural University. Special thanks to Professor Lin Li from Huazhong Agricultural University for assisting with the extraction of ribosome complexes.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
M.W. conceived and designed the project. F.L. provided the materials. X.T. extracted RNA and performed the polysome profiling-seq experiment. X.T., R.W., Z.L., S.L., and X.C. analyzed the data. Z.Z., H.L., X.Z. contributed to project discussion. X.T. and R.W. wrote the manuscript draft, and M.W. and X.Z. revised it. All authors read and approved the final manuscript.

Funding
This study was supported by the National Natural Science Foundation of China (32170645) and the National Key Research and Development Program of China (2021YFF1000900). This study was also supported by the Foundation of Hubei Hongshan Laboratory (2021hszd014) and the Tianchi Talent Introduction Plan of Xinjiang Uyghur Autonomous Region.

Data availability
For wild diploid cotton species, we utilized genome assembly data publicly available from the NCBI database under BioProject accession number PRJNA788082 [84]. Meanwhile, genome assemblies and annotations were downloaded from the Figshare repository (https://figshare.com/projects/Gossypium_pan-genomes/128336). To construct a comprehensive wild diploid cotton TE library, we processed annotated TEs using a combination of CD-HIT [85, 86] and EDTA [87]. We then used the curated TE library generated in the previous step to re-annotate TEs in the wild diploid cotton genomes. This re-annotation was performed using EDTA [87]. All sequencing data generated in the present study have been deposited into the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/) under the accession number PRJNA1087584 [88]. There were no other scripts and software used other than those mentioned in the Methods section.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 August 2024 Accepted: 7 March 2025
Published online: 17 March 2025

References

1. O'Malley MA, Leger MM, Wideman JG, Ruiz-Trillo I. Concepts of the last eukaryotic common ancestor. Nat Ecol Evol. 2019;3:338–44.
2. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. The physiology and habitat of the last universal common ancestor. Nat Microbiol. 2016;1(9):16116.
3. Woese C. The universal ancestor. PNAS. 1998;95:6854–9.
4. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 2003;1:127–36.
5. Arndt NT, Nisbet EG. Processes on the young earth and the habitats of early life. Annual Rev Earth Planet Sci. 2012;40:521–49.
6. Mojzsis SJ, Harrison TM, Pidgeon RT. Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4,300 Myr ago. Nature. 2001;409:178–81.
7. Tashiro T, Ishida A, Hori M, Igisu M, Koike M, Méjean P, Takahata N, Sano Y, Komiya T. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. Nature. 2017;549:516–8.
8. Rivas-González I, Rousselle M, Li F, Zhou L, Dutheil JY, Munch K, Shao Y, Wu DD, Schierup MH, Zhang GJ. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. Science. 2023;380(6648):eabn4409.

9. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012;13:505–16.
10. LaPotin S, Swartz ME, Luecke DM, Constantinou SJ, Gallant JR, Eberhart JK, Zakon HH. Divergent cis-regulatory evolution underlies the convergent loss of sodium channel expression in electric fish. Sci Adv. 2022;8(22):eabm2970.
11. Bachem K, Li X, Ceolin S, Muhling B, Horl D, Harz H, Leonhardt H, Arnoult L, Weber S, Matarlo B, et al. Regulatory evolution tuning pigmentation intensity quantitatively in *Drosophila*. Sci Adv. 2024;10:eadl2616.
12. Kingsley EP, Hager ER, Lassance J-M, Turner KM, Harringmeyer OS, Kirby C, Neugeboren BI, Hoekstra HE. Adaptive tail-length evolution in deer mice is associated with differential Hoxd13 expression in early development. Nat Ecol Evol. 2024;8(4):791–805.
13. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15:201–6.
14. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16:1297–305.
15. Zhu W, Xu J, Chen S, Chen J, Liang Y, Zhang C, Li Q, Lai J, Li L. Large-scale translatome profiling annotates the functional genome and reveals the key role of genic 3′ untranslated regions in translatomic variation in plants. Plant Commun. 2021;2:100181.
16. Jan CH, Williams CC, Weissman JS. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science. 2014;346:1257521.
17. Zeng H, Huang JH, Ren JY, Wang CK, Tang ZF, Zhou HW, Zhou YM, Shi HL, Aditham A, Sui X, et al. Spatially resolved single-cell translatomics at molecular resolution. Science. 2023;380(6652):eadd3067.
18. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell. 2007;129:1401–14.
19. Lunardon A, Johnson NR, Hagerott E, Phifer T, Polydore S, Coruh C, Axtell MJ. Integrated annotations and analyses of small RNA-producing loci from 47 diverse plants. Genome Res. 2020;30:497–513.
20. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:1260419.
21. Guo TN, Kouvonen P, Koh CC, Gillet LC, Wolski WE, Röst HL, Rosenberger G, Collins BC, Blum LC, Gillessen S, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med. 2015;21:407–13.
22. Mazin PV, Khaitovich P, Cardoso-Moreira M, Kaessmann H. Alternative splicing during mammalian organ development. Nat Genet. 2021;53:925–34.
23. Verta JP, Jacobs A. The role of alternative splicing in adaptation and evolution. Trends Ecol Evol. 2022;37(4):299–308.
24. Wright CJ, Smith CWJ, Jiggins CD. Alternative splicing as a source of phenotypic diversity. Nat Rev Genet. 2022;23:697–710.
25. Liu YS, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016;165:535–50.
26. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011;473:337–42.
27. Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, Rodriguez EH, Fields AP, Schwartz S, Raychowdhury R, et al. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science. 2015;347:1259038.
28. Jangam D, Feschotte C, Betran E. Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet. 2017;33:817–31.
29. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10:691–703.
30. Huang CR, Burns KH, Boeke JD. Active transposition in genomes. Annu Rev Genet. 2012;46:651–75.
31. Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. Nat Rev Mol Cell Biol. 2022;23:481–97.
32. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;21:721–36.
33. Etchegaray E, Naville M, Volff JN, Haftek-Terreau Z. Transposable element-derived sequences in vertebrate development. Mob DNA. 2021;12:1.
34. Drongitis D, Aniello F, Fucci L, Donizetti A. Roles of transposable elements in the different layers of gene expression regulation. Int J Mol Sci. 2019;20(22):5755.
35. Kitano S, Kurasawa H, Aizawa Y. Transposable elements shape the human proteome landscape via formation of cis-acting upstream open reading frames. Genes Cells. 2018;23:274–84.
36. Shen J, Liu J, Xie K, Xing F, Xiong F, Xiao J, Li X, Xiong L. Translational repression by a miniature inverted-repeat transposable element in the 3′ untranslated region. Nat Commun. 2017;8:14651.
37. Abascal F, Tress ML, Valencia A. Alternative splicing and co-option of transposable elements: the case of TMPO/LAP2alpha and ZNF451 in mammals. Bioinformatics. 2015;31:2257–61.
38. Hawkins JS, Proulx SR, Rapp RA, Wendel JF. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. PNAS. 2009;106:17811–6.
39. Wang MJ, Li JY, Wang PC, Liu F, Liu ZP, Zhao GN, Xu ZP, Pei LL, Grover CE, Wendel JF, et al. Comparative genome analyses highlight transposon-mediated genome expansion and the evolutionary architecture of 3D genomic folding in cotton. Mol Biol Evol. 2021;38:3621–36.
40. Wendel. JF, Brubaker. CL, Seelanan. T: The origin and evolution of Gossypium. in: Physiology of cotton. (Eds.) Stewart, J M, Oosterhuis, D M, Heitholt, J J, Mauney, J R, Springer Netherlands. Dordrecht. 2010:1-18.
41. Wang M, Li J, Qi Z, Long Y, Pei L, Huang X, Grover CE, Du X, Xia C, Wang P, et al. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. Nat Genet. 2022;54:1959–71.

42. Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, Moreno-Santillán DD, Crookshanks C, Roberts J, Garcia C, et al. Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science. 2023;380(6643):eabn1430.
43. Wang B, Regulski M, Tseng E, Olson A, Goodwin S, McCombie WR, Ware D. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. Genome Res. 2018;28:921–32.
44. Wang ZY, Leushkin E, Liechti A, Ovchinnikova S, Mössinger K, Brüning T, Rummel C, Grützner F, Cardoso-Moreira M, Janich P, et al. Transcriptome and translatome co-evolution in mammals. Nature. 2020;588(7839):642–7.
45. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. PNAS. 2009;106:7507–12.
46. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. Embo J. 2016;35:706–23.
47. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. PNAS. 2010;107:3645–50.
48. Korostelev AA. The structural dynamics of translation. Annu Rev Biochem. 2022;91:245–67.
49. Li Q, Makri A, Lu Y, Marchand L, Grabs R, Rousseau M, Ounissi-Benkalha H, Pelletier J, Robert F, Harmsen E, et al. Genome-wide search for exonic variants affecting translational efficiency. Nat Commun. 2013;4:2260.
50. Zhan JP, Meyers BC. Plant small RNAs: their biogenesis, regulatory roles, and functions. Annu Rev Plant Biol. 2023;74:21–51.
51. Betti F, Ladera-Carmona MJ, Weits DA, Ferri G, Iacopino S, Novi G, Svezia B, Kunkowska AB, Santaniello A, Piaggesi A, et al. Exogenous miRNAs induce post-transcriptional gene silencing in plants. Nat Plants. 2021;7:1379–88.
52. Applequist WL, Cronn R, Wendel JF. Comparative development of fiber in wild and cultivated cotton. Evol Dev. 2001;3:3–17.
53. Wei KHC, Mai D, Chatla K, Bachtrog D. Dynamics and impacts of transposable element proliferation in the species group radiation. Mol Biol Evol. 2022;39(5):msac080.
54. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65(65):505–30.
55. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18:71–86.
56. Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, Quadrana L. The impact of transposable elements on tomato diversity. Nat Commun. 2020;11(1):4058.
57. Zhang YY, Li ZJ, Liu JY, Zhang Y, Ye LH, Peng Y, Wang HY, Diao H, Ma Y, Wang MY, et al. Transposable elements orchestrate subgenome-convergent and -divergent transcription in common wheat. Nat Commun. 2022;13(1):6940.
58. Judd J, Sanderson H, Feschotte C. Evolution of mouse circadian enhancers from transposable elements. Genome Biol. 2021;22(1):193.
59. Modzelewski AJ, Shao WQ, Chen JQ, Lee A, Qi X, Noon M, Tjokro K, Sales G, Biton A, Anand A, et al. A mouse-specific retrotransposon drives a conserved isoform essential for development. Cell. 2021;184:5541–58.
60. Sigman MJ, Slotkin RK. The first rule of plant transposable element silencing: location, location, location. Plant Cell. 2016;28:304–13.
61. Wei KHC, Chan C, Bachtrog D. Establishment of H3K9me3-dependent heterochromatin during embryogenesis in *Drosophila miranda*. Elife. 2021;10:e55612.
62. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009;25:4.10.1-4.10.14.
63. Ou SJ, Su WJ, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):275.
64. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
65. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng SH, Stiller J, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.
66. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
67. Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, Gore MA, Wendel JF. Re-evaluating the phylogeny of allopolyploid *Gossypium L*. Mol Phylogenet Evol. 2015;92:45–52.
68. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
70. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
71. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
72. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
73. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.
74. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
75. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
76. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. Full-length transcript characterization of mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11(1):1438.

77. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19:40.
78. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. The UCSC genome browser database: 2019 update. Nucleic Acids Res. 2019;47:D853–8.
79. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.
80. Connolly MA, Clausen PA, Lazar JG: Preparation of RNA from plant tissue using guanidinium isothiocyanate/cesium chloride ultracentrifugation. CSH Protoc 2006, 2006 (1):pdb.prot4102.
81. Li G, Chen C, Chen P, Meyers BC, Xia R. sRNAminer: a multifunctional toolkit for next-generation sequencing small RNA data mining in plants. Sci Bull (Beijing). 2024;69:784–91.
82. Wu HJ, Ma YK, Chen T, Wang M, Wang XJ. PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res. 2012;40:W22-28.
83. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
84. Wang M, Li J, Qi Z, Long Y, Pei L, Huang X, Grover CE, Du X, Xia C, Wang P, et al. Genomic innovation and regulatory rewiring during evolution of the cotton genus Gossypium. PRJNA788082. Sequence Read Archive. 2022. https://www.ncbi.nlm.nih.gov/bioproject/788082.
85. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
86. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
87. Ou S, Scheben A, Collins T, Qiu Y, Seetharam AS, Menard CC, Manchanda N, Gent JI, Schatz MC, Anderson SN, et al. Differences in activity and stability drive transposable element variation in tropical and temperate maize. Genome Res. 2024;34:1140–53.
88. Tian X, Wang R, Liu Z, Lu S, Chen X, Zhang Z, Liu F, Li H, Zhang X, Wang M. Widespread impact of transposable elements on the evolution of post-transcriptional regulation in the cotton genus Gossypium. PRJNA1087584. Sequence Read Archive. 2025. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1087584.

Publisher's Note