



OPEN ACCESS

EDITED BY

Hushuai Nie,
Inner Mongolia Agricultural University,
China

REVIEWED BY

Haodong Chen,
Hunan Academy of Agricultural Sciences,
China
Ting-Ting Zhang,
Gansu Agricultural University, China

*CORRESPONDENCE

Junduo Wang
✉ 13579975299@163.com
Yajun Liang
✉ 13999966149@163.com

[†]These authors have contributed equally
to this work

RECEIVED 29 March 2026
REVISED 20 April 2026
ACCEPTED 21 April 2026
PUBLISHED 08 May 2026
CORRECTED 12 June 2026

CITATION

Chen H, Liu X, Yang N, Gong Z, Sun F,
Geng S, Zheng J, Qian S, Wang J and
Liang Y (2026) Evolutionary origin and
asymmetric subgenomic retention of the
lncRNA *pGhFAD2-1* that
regulates cotton lipid metabolism.
Front. Plant Sci. 17:1841757.
doi: 10.3389/fpls.2026.1841757

COPYRIGHT

© 2026 Chen, Liu, Yang, Gong, Sun,
Geng, Zheng, Qian, Wang and Liang. This
is an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Evolutionary origin and asymmetric subgenomic retention of the lncRNA *pGhFAD2-1* that regulates cotton lipid metabolism

Haihong Chen^{1†}, Xuan Liu^{2†}, Ni Yang^{1†}, Zhaolong Gong¹,
Fenglei Sun¹, Shiwei Geng¹, Juyun Zheng¹, Shuaishuai Qian¹,
Junduo Wang^{1*} and Yajun Liang^{1*}

¹National Cotton Engineering Technology Research Center, Cotton Research Institute of Xinjiang
Uyghur Autonomous Region Academy of Agricultural Sciences, Wulumuqi, Xinjiang, China,

²Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences,
Shenzhen, China

The origin and evolutionary dynamics of long non-coding RNAs (lncRNAs) represent a pivotal frontier in plant functional genomics. Here, we investigated *pseudo-GhFAD2-1* (*pGhFAD2-1*), a functionally validated lncRNA regulating cottonseed fatty acid metabolism in *Gossypium hirsutum*, delineating its evolutionary origin and structural dynamics through a comprehensive cross-genomic comparative strategy. Homologous sequence identification, phylogenetic reconstruction, and molecular evolutionary analyses were conducted across 27 *Gossypium* species representing the A, B, D, E, F, G, K, and AD genomes. Our findings reveal that *pGhFAD2-1* originated from a tandem duplication event of *GhFAD2-1D* in the D-genome ancestor approximately 5 million years ago (MYA). This duplicated locus was subsequently disrupted by a 1,221-bp exogenous sequence insertion, which abolished its protein-coding capacity and catalyzed the emergence of this novel lncRNA transcript. Subjected to strong purifying selection, this reorganized locus retained its core functional domain and exhibited high-fidelity inheritance (sequence identity > 85%) across D-genome diploids and AD-genome allotetraploids. Conversely, only fragmented remnants or complete deletions of this locus were detected in non-D genome *Gossypium* species. This study systematically characterizes the evolutionary trajectory and subgenomic conservation of *pGhFAD2-1*, elucidating its adaptive significance as a D-lineage-specific lncRNA. These findings provide novel insights into how lineage-specific structural variations rewire epigenetic regulatory networks to drive adaptive diversification in allopolyploid crops, establishing an evolution-informed molecular target for the synergistic genetic improvement of cotton fiber and seed oil traits.

KEYWORDS

cottonseed oil, evolutionary dynamics, *Gossypium*, *pGhFAD2-1*, structural variation

1 Introduction

Lineage-specific evolution of long non-coding RNAs (lncRNAs) and the subsequent rewiring of epigenetic regulatory networks are crucial drivers of adaptive phenotypic diversification in polyploid plants. In upland cotton (*Gossypium hirsutum* L), a premier dual-purpose crop for natural fiber and edible oil, optimizing cottonseed lipid metabolism—particularly the oleic acid (C18:1) to linoleic acid (C18:2) ratio—has become a primary breeding objective to improve oil oxidative stability and nutritional value (Wu et al., 2022). The conversion of C18:1 to C18:2 in cotton is predominantly catalyzed by the microsomal Δ^{12} fatty acid desaturase 2 (*FAD2*) family (Liu et al., 2002; Ma et al., 2021). In higher plants, the *FAD2* gene family represents a classic model of evolutionary subfunctionalization. Comparative genomic studies across diverse taxa, including *Arabidopsis* (Okuley et al., 1994), soybean (Heppard et al., 1996; Schlueter et al., 2007; Lakhssassi et al., 2017), and peanut (Chi et al., 2011), underscore that *FAD2* genes frequently originate through tandem and whole-genome duplications, subsequently undergoing functional divergence to satisfy distinct physiological requirements (Dar et al., 2017). Rather than conforming to rigid tissue-specific categories, these duplicates typically exhibit a continuous spectrum of spatiotemporal expression profiles. Through this evolutionary partitioning, specific members maintain basal transcript abundance throughout vegetative organs such as roots, stems, and leaves to ensure cellular membrane homeostasis, while others are preferentially recruited to drive the specialized accumulation of storage lipids during seed maturation (Heppard et al., 1996; Dar et al., 2017). Recent genomic analyses have revealed that the *GhFAD2* gene family has significantly expanded in allotetraploid cotton, comprising nine homologous members distributed across the At and Dt subgenomes (Paterson et al., 2012). While these duplicated members exhibit broad functional divergence across various tissues, the *GhFAD2-1* orthologous pair (comprising *GhFAD2-1A* and *GhFAD2-1D*) acts as the master regulator during cottonseed development (Liu et al., 2019; Ma et al., 2021).

Notably, an additional unique locus—originally annotated as the protein-coding gene *Gh_D13G2237*—is located tandemly adjacent to *GhFAD2-1D* on chromosome D13. Our recent functional characterization demonstrated that this locus lacks the conserved histidine boxes essential for typical *FAD2* enzymatic activity. Instead, it functions as a lineage-specific lncRNA, subsequently designated as *pseudo-GhFAD2-1* (*pGhFAD2-1*), which epigenetically suppresses cottonseed lipid accumulation by recruiting the histone deacetylase GhHDT1 to remodel the local chromatin landscape (Chen et al., 2025). Despite its well-documented biological function in lipid metabolism, the evolutionary trajectory and the genomic rearrangements underlying the genesis of *pGhFAD2-1* across the *Gossypium* phylogeny remain largely elusive (Huang et al., 2021). Elucidating this phylogenetic trajectory is essential to determine whether it represents an ancestral genomic feature shared among early *Gossypium* lineages or a recent innovation arising from subgenome-specific structural variations (Glaszmann et al., 2010). Understanding this origin provides broader insights into the *de novo* birth of regulatory lncRNAs and the genetic architecture underlying adaptive trait divergence in polyploid crops (Wu et al., 2017).

To address this, we conducted a comprehensive phylogenomic survey across 27 representative *Gossypium* species, encompassing the major diploid genomes (A, B, D, E, F, G, and K) and the allotetraploid genomes (AD). By integrating fine-scale structural variation analyses, population evolutionary dynamics, and macroscopic phenotypic evaluations, we aimed to delineate the precise *de novo* origination trajectory of this lineage-specific lncRNA (Li et al., 2023). Our findings illustrate how structural variations drive the “duplication-insertion-neofunctionalization” of a pseudogenized duplicate, providing insights into the evolutionary trade-offs shaping cottonseed lipid metabolism and offering an evolution-based molecular target for synergistic crop improvement.

2 Materials and methods

2.1 Genomic data retrieval of *Gossypium* species

The reference genome assemblies and annotation files of 27 *Gossypium* species representing diverse evolutionary clades were retrieved from the CottonGen database (<https://www.cottongen.org>) and the NCBI GenBank database (<https://www.ncbi.nlm.nih.gov/genbank/>). These datasets included *Gossypium hirsutum* L. acc. TM-1 (AD₁), *Gossypium barbadense* L. acc. Hai7124 (AD₂), *Gossypium herbaceum* L. acc. Zhongcao 1 (A₁), *Gossypium arboreum* L. acc. Zhongya 1 (A₂), *Gossypium raimondii* Ulbr. acc. Jianbanmian 4A (D₅), along with 21 other diploid wild accessions (*Gossypium thurberi* Tod. D₁₋₅, *Gossypium armourianum* D₂₋₁, *Gossypium harknessii* D₂₋₂, *Gossypium davidsonii* D_{3d-27}, *Gossypium klotzschianum* Andress. D_{3-k}, *Gossypium aridum* D₄, *Gossypium gossypoides* D₆, *Gossypium lobatum* D₇, *Gossypium trilobum* D₈, *Gossypium laxum* D₉, *Gossypium turneri* D₁₀, *Gossypium schwendimanii* D₁₁, *Gossypium anomalum* Wawr. & Peyr. B₁, *Gossypium stocksii* E₁, *Gossypium longicalyx* F₁, *Gossypium bickii* G₁, *Gossypium australe* G₂, *Gossypium rotundifolium* K₂), wild tetraploids (*Gossypium tomentosum* AD₃, *Gossypium mustelinum* AD₄, *Gossypium darwinii* Watt AD₅), and the closely related distant reference sequence *Gossypoides kirkii* were incorporated (Wang et al., 2012). Detailed information regarding genome assembly versions, accession numbers, and data sources is listed in Table 1.

2.2 Identification and extraction of *pGhFAD2-1* homologous sequences

To comprehensively trace the insertion and retention status of the *pGhFAD2-1* locus across the *Gossypium* phylogeny, local BLASTN searches were performed using TBtools v1.120 (Chen et al., 2020). The full-length exonic genomic DNA sequence of *pGhFAD2-1* from the *Gossypium hirsutum* TM-1 genome served as the query against the local genomic databases of the other 26 species and the distant reference sequence. The screening criteria for homologous sequences were rigorously set with the following thresholds: E-value $\leq 1e-10$, sequence identity $\geq 70\%$, alignment coverage $\geq 50\%$, word size = 11, match score = 2, and mismatch

TABLE 1 *pGhFAD2-1* homologous sequences in cotton species.

Taxa	Genome	Locality/ cultivar	Year first described	Length (bp)	Sequence identity to reference <i>pGhFAD2-1</i> (%)	Insertion region integrity	Accession no.
Cultivated species							
<i>Gossypium herbaceum</i> L.	A ₁	Africa-Asia	1753	202	12.52	Partial	CP106688.1
<i>Gossypium arboreum</i> L.	A ₂	Asia	1753	202	12.45	Partial	CM002691.1
<i>Gossypium hirsutum</i> Linn.	AD ₁	Southern Mexico	1763	1477	100	Complete	AY321158.1
<i>Gossypium barbadense</i> Linn.	AD ₂	NW South America	1753	1477	96.15	Complete	CM018227.1
<i>Gossypium tomentosum</i>	AD ₃	Hawaiian Islands	1865	1477	95.89	Complete	CM017635.1
<i>Gossypium mustelinum</i>	AD ₄	NE Brazil	1907	1478	95.89	Complete	CM017661.1
<i>Gossypium darwinii</i> Watt	AD ₅	Galapagos Islands	1907	1477	96.09	Complete	CM017713.1
Wild species							
<i>Gossypium anomalum</i> Wawr. & Peyr.	B ₁	Africa	1860	202	12.58	Partial	CM033496.1
<i>Gossypium thurberi</i> Tod.	D ₁₋₅	Mexico and SW U.S.	1854	1544	87.29	Complete	CM013390.1
<i>Gossypium armourianum</i>	D ₂₋₁	Mexico	1933	1471	88.20	Complete	CM025244.1
<i>Gossypium harknessii</i>	D ₂₋₂	Mexico	1889	1435	86.90	Complete	CM025231.1
<i>Gossypium davidsonii</i>	D _{3d-27}	Mexico	1873	1476	89.50	Complete	CM025218.1
<i>Gossypium klotzschianum</i> Andress.	D _{3-k}	Galapagos Islands	1853	1476	89.50	Complete	CM025205.1
<i>Gossypium aridum</i>	D ₄	Mexico	1911	1468	87.55	Complete	CM025192.1
<i>Gossypium raimondii</i> Ulbr.	D ₅	Peru	1932	1480	95.37	Complete	CM001752.1
<i>Gossypium gossypoides</i>	D ₆	Mexico	1913	1449	87.22	Complete	CM025153.1
<i>Gossypium lobatum</i>	D ₇	Mexico	1956	1520	88.57	Complete	CM025283.1
<i>Gossypium trilobum</i>	D ₈	Mexico	1824	1478	92.05	Complete	CM025166.1
<i>Gossypium laxum</i>	D ₉	Mexico	1972	1478	85.31	Complete	CM025270.1
<i>Gossypium turneri</i>	D ₁₀	Mexico Northwest	1978	1476	88.98	Complete	CP032583.1
<i>Gossypium schwendimanii</i>	D ₁₁	Mexico Michoacan	1987	1476	89.31	Complete	CM025257.1
<i>Gossypium stocksii</i>	E ₁	Arabia	1874	88	5.8	Partial	CM035632.1
<i>Gossypium longicalyx</i>	F ₁	Arabia	1958	228	12.45	Partial	CM047007.1
<i>Gossypium bickii</i>	G ₁	Australia	1910	50	3.13	Partial	CP106709.1
<i>Gossypium australe</i>	G ₂	Australia	1858	50	3.06	Partial	SMMG02000001.1
<i>Gossypium rotundifolium</i>	K ₂	Western Australia	1992	235	10.23	Partial	CP106734.1
<i>Gossypoides kirkii</i>	other	Southeast Africa	1947	425	13.62	Partial	CP032255.1

penalty = -3. To unambiguously distinguish true orthologous sequences from other paralogous copies within the expanded *FAD2* gene family, the highly conserved flanking sequences (2,000 bp upstream and downstream of the *pGhFAD2-1* locus in TM-1) were utilized as syntenic anchors to accurately demarcate the conserved genomic blocks. Only sequences located within these conserved syntenic blocks were retained as bona fide orthologs, with paralogous copies completely excluded (Charif and Lobry, 2007). Subsequently, the strictly orthologous sequences, including the regions exhibiting presence/absence variations (PAVs) of the large insertion fragment, were precisely extracted based on their genomic coordinates for downstream analyses.

2.3 Sequence alignment and structural motif analysis of *pGhFAD2-1* homologous sequences

Multiple sequence alignments of the *pGhFAD2-1* homologous sequences were initially generated using DNAMAN v9.0 software (Wang, 2016) and subjected to rigorous manual refinement to eliminate ambiguous regions and uninformative alignment gaps. The resulting alignments were subsequently formatted and visualized utilizing the ESPript 3.0 server (Gouet et al., 2003). Conserved motifs across the 27 homologous sequences were predicted using the MEME Suite v5.5.4 (Bailey et al., 2009). For *cis*-acting regulatory

element analysis, the exonic genomic DNA sequences of *pGhFAD2-1* orthologs were submitted to the New PLACE database (Higo et al., 1999) for element prediction. The integrated visualization of these structural features was performed using TBtools v1.120 (Chen et al., 2020).

2.4 Molecular evolutionary analysis

To comprehensively evaluate the evolutionary constraints and substitution patterns of the *pGhFAD2-1* locus across divergent *Gossypium* species, a series of molecular evolutionary analyses were conducted based on the refined multiple sequence alignments manually curated in the previous section. The closely related species *Gossypioides kirkii* was utilized as the distant reference sequence only for the verification of the lineage-specific distribution of this locus, and was not used for phylogenetic rooting or divergence time estimation, as it does not meet the core criterion for a valid outgroup (possessing intact orthologous sequences with the inner group) defined by classic phylogenetic systematics studies (Hillis et al., 1996; Nixon and Carpenter, 1993). Prior to phylogenetic reconstruction, the optimal nucleotide substitution model (GTR+G+I) was automatically recommended by the ModelFinder algorithm implemented in the IQ-TREE software (Kalyaanamoorthy et al., 2017; Nguyen et al., 2015). This algorithm identifies the best-fit substitution model by calculating and comparing the Bayesian Information Criterion (BIC) scores among various candidate models according to the principle of statistical optimality. The GTR+G+I model is a highly versatile and classic model for nucleotide phylogenetic analysis. It explicitly accounts for unequal base frequencies, differential base substitution rates, among-site rate heterogeneity (G, Gamma distribution), and the proportion of invariable sites (I), which perfectly matches the data characteristics of our sequence-based evolutionary inferences. Divergence times were estimated using the RelTime-ML method implemented in MEGA 11 v11.0.13 (Tamura et al., 2021) under a relaxed molecular clock model. Established divergence nodes (~10 MYA for *Gossypium-Gossypioides* divergence and ~5 MYA for A-D genome divergence) were utilized as calibration points (Pathak et al., 2023; Wang et al., 2022). The subsequent allopolyploidization event denoted as WGD in the motif analysis was phylogenetically demarcated at the specific divergence nodes distinguishing the allotetraploid At and Dt subgenomes from their respective diploid ancestors. Furthermore, to characterize the structural features of the exogenous insertion, transposable elements (TEs) and repetitive sequences within the 1,221-bp fragment were identified using RepeatMasker (Smit et al., 2013–2015) against the Repbase library.

To assess the equality of evolutionary rates among different subgenomic lineages (D/AD genome lineages harboring the intact *pGhFAD2-1* locus vs. non-D genome lineages with only fragmented homologous sequences), Tajima's relative rate test was performed (Tajima, 1993), with $p < 0.05$ set as the threshold for significant rate heterogeneity. The interspecific evolutionary divergence (base differences per site) was estimated using the p-distance method with 1,000 bootstrap replicates (Liu et al., 2001). In addition, the homogeneity of substitution patterns between the two lineage groups was strictly evaluated using a Monte Carlo test (1,000 replicates) to estimate P -values, with $p < 0.05$ set as the significance threshold (Kumar and Gadagkar, 2001).

Finally, to infer the evolutionary dynamics and stability of the *pGhFAD2-1* locus within the D-genome lineage, the mismatch distribution (pairwise sequence differences) was calculated using DnaSP v5.10.01 (Librado and Rozas, 2009). Following the established evolutionary framework for plant lineage-specific loci, a multimodal distribution of mismatch frequencies indicates that the locus has remained evolutionarily stable under consistent selective constraints over time, whereas a unimodal distribution signifies recent rapid sequence diversification. The outputs from these evolutionary analyses were subsequently visualized as heatmaps utilizing TBtools v1.120 (Chen et al., 2020) and annotated using the *ggtree* package in R (Yu et al., 2017).

2.5 Histological staining and transmission electron microscopy

Histological localization of lipid droplets was conducted following our previously established protocols (Chen et al., 2025; Liu et al., 2019). Briefly, fresh cottonseed specimens were fixed in FAA solution at 4 °C for 2 h, followed by standard dehydration and paraffin embedding. Sections (5 µm thickness) were deparaffinized and co-stained with the neutral lipid-specific fluorescent dye Nile Red (0.16 µg/mL) and toluidine blue (0.5% w/v) for 20 min in the dark. Imaging and subsequent morphometric quantification were performed using a confocal microscope equipped with CaseViewer software. For TEM observation, fresh tissue blocks (~1 mm³) were fixed in 1% osmium tetroxide for 7 h, dehydrated through a graded ethanol series, and embedded in an 812-acetone resin mixture. Ultrathin sections (60–80 nm) were double-stained with uranyl acetate and lead citrate prior to ultrastructural observation and image acquisition under a transmission electron microscope.

2.6 Cottonseeds oil and fatty acid composition analysis

Cottonseed sampling at different developmental stages was performed as previously described (Ma et al., 2021). Seed oil was extracted and quantified according to Liu et al. (2019) and Ma et al. (2021). For fatty acid composition analysis, fatty acid methylation was performed using the KOH-methanol method, and qualitative and quantitative analyses were conducted using a GCMS-QP2020 system. The relevant detailed procedures were implemented according to Liu et al. (2019) and Ma et al. (2021).

2.7 Statistical analysis

Data processing and statistical analyses were performed using Microsoft Excel 2019 and SPSS 18.0 (SPSS, Chicago, USA). Phenotypic differences including total oil content and fatty acid components were analyzed using independent samples t-tests for pairwise comparisons, and one-way ANOVA followed by Tukey's *post-hoc* test for multiple comparisons. The means of different species were compared at the 0.05 probability level, with exact P -values provided for specific pairwise phenotypic comparisons (Harms and Thornton, 2013).

3 Results

3.1 Comprehensive phylogenetic analysis and divergence time estimation of *pGhFAD2-1*

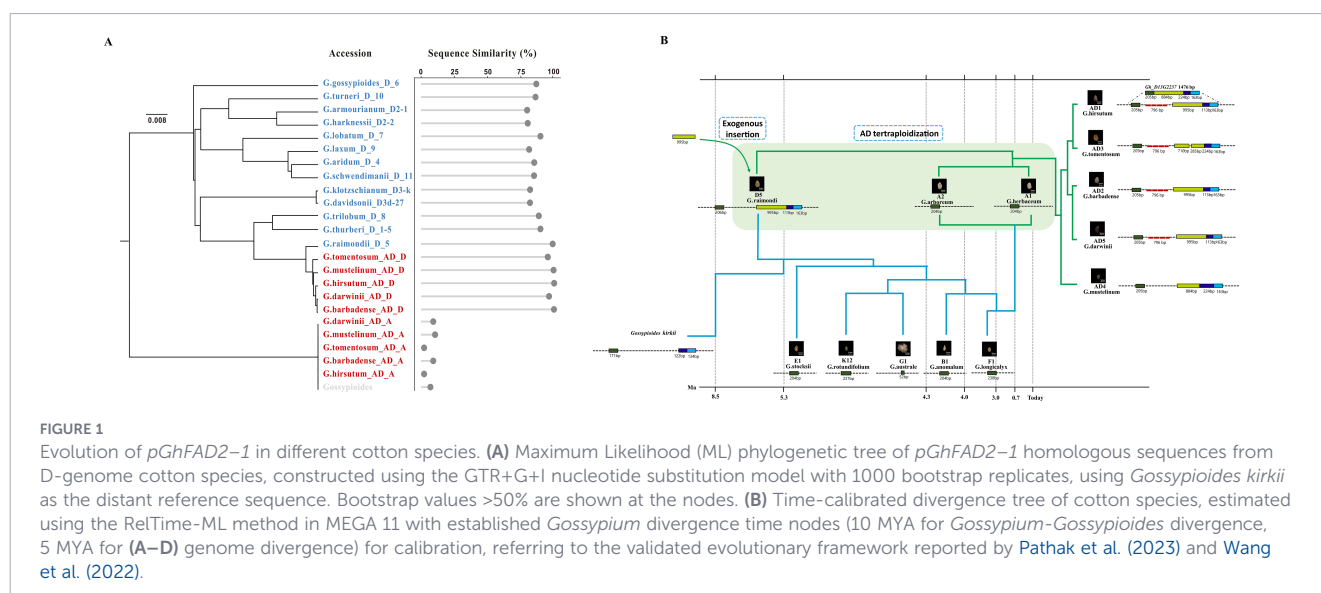
As previously established, the functional lncRNA *pGhFAD2-1* is specifically retained in D-genome lineages. To trace the precise evolutionary origin of this lineage-specific locus, we conducted a comprehensive phylogenomic analysis. Using the closely related species *Gossypioides kirkii*—a widely recognized distant reference sequence for *Gossypium* phylogenetic studies—we reconstructed a Maximum Likelihood (ML) phylogenetic tree of *pGhFAD2-1* homologous sequences under the optimal GTR+G+I nucleotide substitution model, with 1,000 bootstrap replicates. The reliability of the tree topology was supported by high bootstrap values (>80%) at all major nodes, indicating a strong phylogenetic signal during the evolutionary inference. This phylogenetic tree was constructed only to display the sequence homology and lineage-specific distribution characteristics of *pGhFAD2-1* homologous sequences across *Gossypium* species, not for resolving the deep phylogenetic relationships of different cotton genome groups.

Sequence comparisons revealed that the sequence identity of the *pGhFAD2-1* locus in the examined D-genome species (encompassing D-genome diploids and the Dt subgenome of allotetraploids) consistently exceeded 85%, utilizing the *Gossypium hirsutum* TM-1 sequence as the reference. This high sequence conservation aligned with the tree topology, where all intact D-lineage homologous sequences clustered together within a core evolutionary clade. In contrast, the At subgenome of AD-genome cotton species retained only highly fragmented remnant sequences. Due to this severe sequence degradation, the At-subgenome remnants exhibited considerable genetic distance from the D-lineage clade and clustered into a distinct, distant branch (Figure 1A). This topology corroborates that the intact *pGhFAD2-1* locus is exclusively a lineage-specific evolutionary innovation of the D-genome.

To delineate the evolutionary timeframe of this locus, we integrated a relaxed molecular clock model for divergence time estimation. Previous comparative genomics studies have established that the major evolutionary divergence between the *Gossypium* lineages containing the D-genome ancestor and other major lineages was initiated approximately 5 million years ago (~5 MYA) (Wang et al., 2022; Pathak et al., 2023). Our time-calibrated phylogeny precisely coincided with this major historical node. These phylogenomic data suggest that the *pGhFAD2-1* locus is not an ancient retention from the common ancestor of *Gossypium*, but originated *de novo* during the early divergence stage of these two major *Gossypium* lineages approximately 5 MYA (Figure 1B). This estimation provides a definitive evolutionary timeframe for the origination of this functional lncRNA.

3.2 Structural evolution reconstructed by tandem duplication and exogenous insertion

Using the full-length exonic genomic DNA sequence of *pGhFAD2-1* as a query, its homologous sequences were comprehensively identified across 27 cotton species. Genomic localizations revealed that *pGhFAD2-1* is tandemly adjacent to the functional *GhFAD2-1D* gene on the chromosome, representing a direct product of an *in situ* tandem duplication of its parental protein-coding gene *GhFAD2-1D*. Fine-scale structural alignments further demonstrated that the full-length *pGhFAD2-1* mature RNA transcript (1,476 nt) is transcribed and reorganized from four distinct genomic DNA blocks (203 bp, 997 bp, 113 bp, and 163 bp in length, respectively), whereas the adjacent *GhFAD2-1D* locus primarily comprises three coding exons (801 bp, 238 bp, and 113 bp) (Figure 2; Supplementary Figure 1). As summarized in Table 1, the full-length intact *pGhFAD2-1* locus was exclusively detected in D-genome diploid and AD-genome allotetraploid species, whereas homologous sequences in other wild species (A, B, E, F, G, and K genomes) exhibited structural fragmentations or complete deletions.



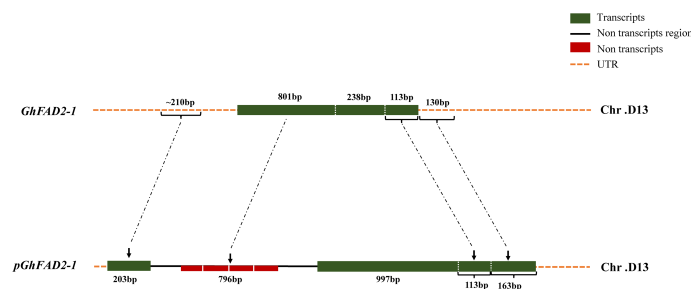


FIGURE 2

Structural alignment of *pGhFAD2-1* and its parental gene *GhFAD2-1D*. Schematic diagram showing the sequence structure comparison and homologous correspondence (dashed arrows) between the parental protein-coding gene *GhFAD2-1D* (upper panel) and its derived lncRNA *pGhFAD2-1* (lower panel) on *Gossypium* chromosome D13. Sequence annotations: green blocks = mature transcript regions; black solid lines = intergenic non-transcribed regions; red blocks = duplicated sequences from the transcribed region of *GhFAD2-1D* that lost transcriptional activity in *pGhFAD2-1*; orange dashed lines = UTRs. The length of each core structural domain is labeled above the corresponding region.

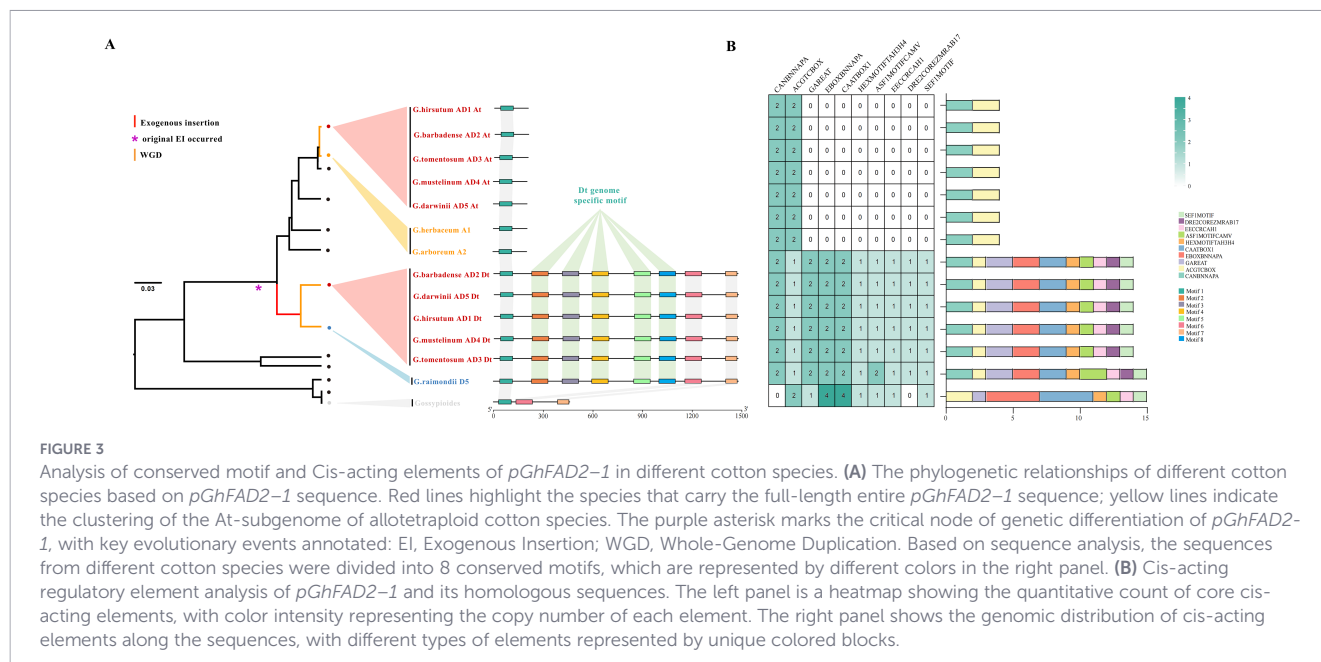
Based on these structural divergences, we further elucidated the mechanistic origin of this lncRNA. Concomitant with the *in situ* tandem duplication, the *pGhFAD2-1* locus underwent structural rearrangements, most notably a large exogenous sequence insertion. Specifically, a 1,221-bp exogenous DNA sequence was inserted into the duplicated coding region. To elucidate the exact origin of this inserted fragment, we performed a comprehensive genome-wide sequence homology search and structural analysis. Strikingly, our results revealed that this 1,221-bp exogenous sequence shares 86.69% sequence identity with a specific uncharacterized intergenic region located on chromosome D11 within the D-subgenome of *Gossypium hirsutum* (AD1) (physical coordinates: ChrD11:43912970-43913997; [Supplementary Figure 2](#)). The accumulated sequence divergence (~13%) between the inserted fragment and its donor locus is consistent with independent sequence evolution following the ancient insertion event (~5 MYA). Furthermore, sequence characterization confirmed that this fragment lacks typical transposable element (TE) features. This crucial evidence strongly indicates that the insertion did not emerge via canonical transposon mobilization, but rather originated from an inter-chromosomal ectopic recombination or genomic sequence capture event originating from the D11 chromosome, which has been stably conserved throughout the allotetraploidization process. During RNA splicing and maturation, a specific splicing event occurred within this inserted sequence: a 224-nt intronic sequence was excised, while the remaining 997-nt segment was retained as the core scaffold of the mature lncRNA. This insertion directly disrupted the original 238-bp core coding segment inherited from the ancestral *GhFAD2-1D* on the genome, leading to the pseudogenization of the duplicated coding sequence, as it lost the third conserved histidine box essential for *FAD2* enzymatic activity. Further sequence comparisons revealed that the 203-bp genomic segment in *pGhFAD2-1* corresponds to a highly conserved region located upstream of the *GhFAD2* locus across all *Gossypium* species. Intriguingly, while the 801-bp 5' coding region of *GhFAD2-1D* is physically present at the *pGhFAD2-1* genomic locus, it has lost transcriptional activity and degenerated into a silent region. In contrast, the 113-bp terminal segment remains highly conserved and maintains robust transcriptional activity within *pGhFAD2-1*. In addition, a 130-bp non-transcribed downstream sequence of

GhFAD2-1D evolved into a 163-nt transcriptional component in *pGhFAD2-1*. Collectively, these structural variations redefined precise transcriptional boundaries (initiating exclusively at an 'A' site and terminating at a 'G' site), ultimately driving the transformation of a duplicated coding gene into a novel non-coding transcript.

3.3 Alteration of *cis*-regulatory elements driven by structural variation

To clarify the reshaping effect of the large exogenous insertion on the regulatory potential of *pGhFAD2-1*, the 1,476-bp exonic genomic DNA sequences corresponding to the mature transcript from diverse cotton species were analyzed for *cis*-acting elements and conserved motifs. Among the identified 8 conserved motifs and 10 core *cis*-acting elements, several critical regulatory modules were specifically enriched within the D-lineage-specific inserted sequence. Notably, the 997-bp transcribed genomic segment exclusively introduced the SEF1 (soybean embryo factor 1) binding motif, the ACGTCBOX (bZIP transcription activator RITA-1 binding motif), and the HEXMOTIF (a well-characterized *cis*-element associated with histone gene expression and chromatin state modification). Consequently, in wild cotton species lacking the D-genome (B, E, F, G, and K genomes), the absence of this large insertion naturally abolishes these key conserved motifs and *cis*-regulatory elements ([Figure 3](#)). Furthermore, as annotated in this phylogenetic framework, whole-genome duplication (WGD) event exclusively represents the recent allopolyploidization of the AD genome lineages. The stable retention of the inserted sequence across the Dt subgenomes confirms that the regulatory architecture of the *pGhFAD2-1* locus, acquired prior to this WGD event, was highly conserved throughout the allotetraploidization process.

Remarkably, the *de novo* acquisition of this histone-related DNA motif precisely aligns with the epigenetic regulatory nature of the *pGhFAD2-1* locus. As demonstrated in our previous functional characterization ([Chen et al., 2025](#)), the mature *pGhFAD2-1* RNA transcript—specifically harboring this transcribed lineage-specific insertion—serves as a molecular scaffold to directly recruit the histone deacetylase GhHDT1. Therefore, this structural innovation exerted dual regulatory effects: it not only introduced novel DNA *cis*-regulatory elements to modulate its own transcription at



the genomic level, but also provided an essential RNA-protein interaction interface for epigenetic silencing at the transcript level. This coordinated regulatory divergence illustrates how lineage-specific structural variation can rewire epigenetic layers to drive adaptive phenotypic diversification in allopolyploid crops.

3.4 Evolutionary dynamics and purifying selection of *pGhFAD2-1*

To evaluate the evolutionary constraints and selective pressures acting on *pGhFAD2-1* during the lineage divergence of *Gossypium*, we analyzed the evolutionary rates and substitution patterns of this sequence across different cotton species. The results indicated that the relative evolutionary rates of *pGhFAD2-1* sequences in the AD and D genomes were significantly different from those in the A, B, E, F, G, and K genomes ($p < 0.05$) (Figure 4A). Compared to the AD and D-genome lineages, other wild cotton species lacking the D-genome exhibited a significantly higher accumulation of base substitutions at each site, indicating substantial interspecific sequence divergence between lineages with and without the intact *pGhFAD2-1* locus (Figure 4B). In addition, significant differences were observed in the homogeneity of substitution patterns between these two lineage groups (Figure 4C). This significant difference in evolutionary rates further confirms that the intact *pGhFAD2-1* locus in D/AD genomes is subject to strong selective constraints, while the fragmented sequences in non-D genomes evolve under neutral selection without functional constraints."

To comprehensively evaluate the evolutionary stability of *pGhFAD2-1* following its origination in D-genome lineages, we performed a mismatch distribution analysis. The mismatch frequencies of *pGhFAD2-1* homologous sequences among D-genome species exhibited a typical multimodal distribution pattern (Figure 5). Consistent with established evolutionary paradigms for plant lineage-specific functional loci, this stable multimodal distribution pattern, combined with the >85% sequence identity across

D-genome lineages, demonstrates that the *pGhFAD2-1* locus has been subjected to strong purifying selection to maintain its core epigenetic regulatory function, and remained highly conserved within D-genome lineages after its *de novo* origination.

3.5 Phenotypic divergence of lipid metabolism across representative *Gossypium* species

To investigate whether the molecular evolutionary trajectory of the *pGhFAD2-1* locus translates into phenotypic divergence, we selected five typical species representing the major evolutionary clades of *Gossypium* (diploid ancestors and allotetraploids) as research materials. These included *Gossypium herbaceum* L. (A₁), *Gossypium arboreum* L. (A₂), *Gossypium raimondii* Ulbr. (D₅), *Gossypium hirsutum* L. (AD₁), and *Gossypium barbadense* L. (AD₂). The lipid distribution and oil content of their mature cottonseeds were phenotypically evaluated.

Transmission electron microscopy (TEM) observations revealed that the distribution of lipid bodies (LBs) and protein bodies (PBs) within the A₁ genome cottonseeds was relatively loose, whereas those in the A₂, AD₁, AD₂, and D₅ genome cottonseeds were distributed more compactly and densely (Figure 6A). Further Nile Red fluorescence staining indicated that, despite obvious variations in overall organ volume among different genome cottonseeds, the fluorescence intensity and spatial distribution density of LBs were fully consistent with the subsequent quantitative oil content measurements (Figure 6B).

Subsequently, we quantitatively determined the absolute oil content and core fatty acid compositions of the examined wild and cultivated cotton species (Table 2). The results demonstrated that the relative oil content of the D-genome wild cotton (D₅) was 19.85%, which was essentially equivalent to that of the AD₁ genome species. However, compared to the A₁ and A₂ genome species lacking the intact *pGhFAD2-1* locus, the oil content of D₅ was lower by 3.4

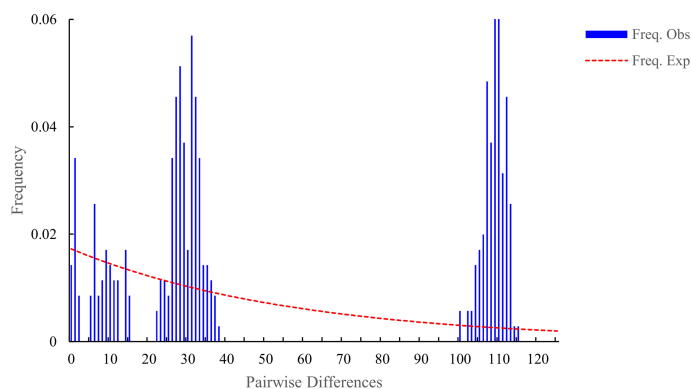
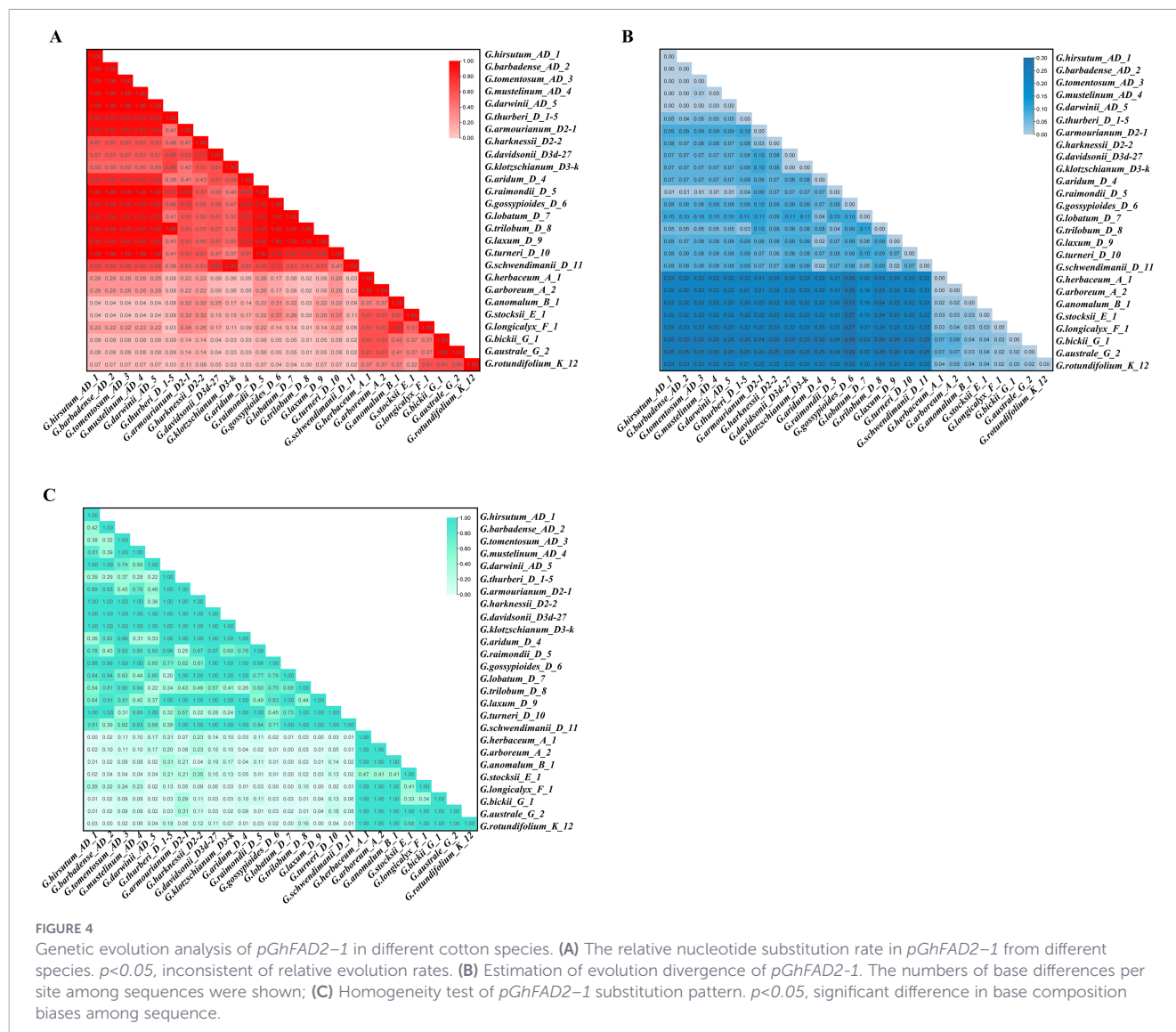
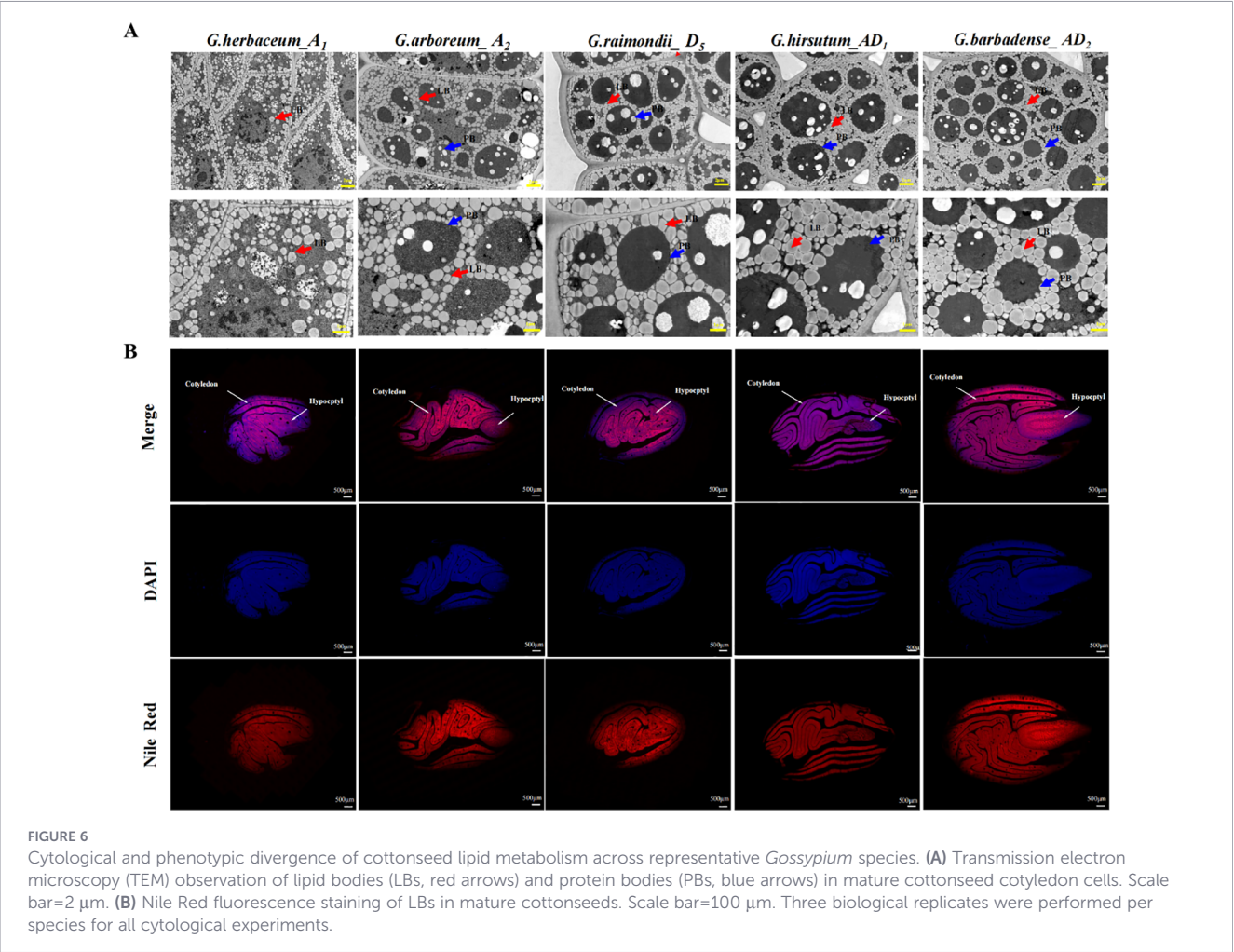


FIGURE 5

The pairwise mismatch distribution of homologous *pGhFAD2-1* in D-genome species. The multimodal pattern of the mismatch distribution indicates a relatively constant population size of D-genome species without recent severe demographic expansion.



percentage points ($p = 0.087$) and significantly lower by 9.3 percentage points ($p = 0.003$), respectively. Regarding fatty acid composition, the relative palmitic acid (C16:0) proportion in the A₁ and A₂ genomes was significantly lower than that in the D₅, AD₁, and AD₂ genomes. Specifically, the C16:0 proportion of A₁ was 14.23% ($p = 0.004$), 12.63% ($p < 0.001$), and 15.64% ($p < 0.001$) lower than

that of D₅, AD₁, and AD₂, respectively. Conversely, the oleic acid (C18:1) proportion in the A₁ genome species was significantly higher than that in the D₅ and AD₁ genomes, showing an increase of 14.92% ($p = 0.021$) and 10.45% ($p = 0.031$), respectively.

In summary, the representative cotton species containing the D-genome (AD and D genomes) generally exhibited higher C16:0

TABLE 2 Total oil content and main fatty acid composition of mature cottonseeds from representative *Gossypium* species.

Taxa	Genome	Oil content	Relative content (% , mean $\pm$ SD)				
			Palmitic acid C16:0	Stearate acid C18:0	Oleate acid C18:1	Linoleate acid C18:2	Linolenic acid C18:3
<i>Gossypium hirsutum</i> L.	(AD) ₁	19.81 $\pm$ 0.35c	24.14 $\pm$ 0.36b	3.35 $\pm$ 0.23b	19.03 $\pm$ 0.71c	50.94 $\pm$ 0.4a	0.24 $\pm$ 0.06a
<i>Gossypium barbadense</i> L.	(AD) ₂	21.66 $\pm$ 0.63a	25 $\pm$ 0.31a	3.33 $\pm$ 0.31b	21.65 $\pm$ 0.82b	47.06 $\pm$ 0.71b	0.12 $\pm$ 0.03b
<i>Gossypium herbaceum</i> L.	A ₁	20.55 $\pm$ 0.31b	21.09 $\pm$ 0.2e	4.27 $\pm$ 0.19a	21.02 $\pm$ 0.78b	50.43 $\pm$ 0.82a	0.1 $\pm$ 0.03b
<i>Gossypium arboreum</i> L.	A ₂	21.89 $\pm$ 0.32a	22.3 $\pm$ 0.62c	3.08 $\pm$ 0.13c	30.68 $\pm$ 0.8a	41.48 $\pm$ 0.67c	0.12 $\pm$ 0.03b
<i>Gossypium raimondii</i> Ulbr.	D ₅	19.85 $\pm$ 0.44bc	24.59 $\pm$ 0.97ab	4.11 $\pm$ 0.06a	18.29 $\pm$ 1.02c	50.56 $\pm$ 1.35a	0.08 $\pm$ 0.01b

Data are from three biological replicates (n=3). Different lowercase letters within a column denote significant differences ($p < 0.05$, one-way ANOVA followed by Tukey's post-hoc test). Data are presented as mean $\pm$ SD of three biological replicates.

proportions but significantly lower overall oil content and C18:1 proportions compared to the A-genome species. These results indicate that cotton species situated in divergent evolutionary clades (with and without the functional *pGhFAD2-1* locus) exhibit significant divergence in cottonseed lipid traits. This phenotypic divergence closely aligns with the genomic evolutionary trajectories revealed previously, providing phenotypic evidence that the lineage-specific retention of *pGhFAD2-1* is tightly associated with the adaptive divergence of cottonseed lipid metabolism, which is fully consistent with its previously validated repressive function in lipid accumulation.

4 Discussion

As a premier global economic crop, upland cotton (*Gossypium hirsutum* L.) serves dually as a core raw material for the textile industry and a primary source of edible oil. In this context, *FAD2* (Δ^{12} -fatty acid desaturase) acts as the rate-limiting enzyme catalyzing the conversion of oleic acid (C18:1) to polyunsaturated linoleic acid (C18:2), thus dictating the final ratio of unsaturated fatty acids. In our previous study, we identified that a tandemly duplicated sequence of *GhFAD2-1D* on chromosome D13 lost its protein-coding capacity and instead functions as a lineage-specific long non-coding RNA (lncRNA), designated as *pseudo-GhFAD2-1* (*pGhFAD2-1*). This lncRNA epigenetically regulates cottonseed lipid metabolism by recruiting the histone deacetylase GhHDT1 (Chen et al., 2025). However, the evolutionary origin and adaptive significance of this lineage-specific regulatory lncRNA have long remained unresolved. The present study systematically dissects the *de novo* origination of *pGhFAD2-1* approximately 5 MYA within the D-genome lineage. Our findings reveal that a single lineage-specific structural reorganization simultaneously endowed this locus with DNA-level transcriptional responsiveness and an RNA-level epigenetic interaction interface. Furthermore, the strict asymmetric retention of this locus in the Dt subgenome extends the paradigm of polyploid asymmetric evolution to non-coding regulatory elements, highlighting the evolutionary trade-offs that shape cottonseed lipid metabolism.

4.1 Structural variation and the *de novo* origination of *pGhFAD2-1*

Lineage-specific structural variations and tandem duplications are widely established as fundamental evolutionary engines driving the *de novo* origination of functional lncRNAs (Ulitsky, 2016; Zhao et al., 2022). Our phylogenomic data across 27 *Gossypium* species reveal that the genesis of *pGhFAD2-1* was initiated by an *in situ* tandem duplication of *GhFAD2-1D*, followed by a D-lineage-specific sequence capture (Grover et al., 2022). Crucially, our genomic tracing explicitly links this exogenous insertion to an inter-chromosomal ectopic recombination event originating from an uncharacterized intergenic region on chromosome D11. Indeed, the inter-chromosomal capture of genomic fragments is emerging

as a potent, yet underappreciated, mechanism for regulatory innovation (Kaessmann, 2010).

This structural reorganization was not merely a disruptive event. Traditionally, duplicated genes that incur severe structural disruptions and frame-shifts are considered evolutionary “dead ends” destined for rapid pseudogenization and elimination from the genome (Lynch and Conery, 2000). However, emerging evolutionary paradigms demonstrate that such pseudogenized loci can occasionally be repurposed into regulatory non-coding RNAs (Cheetham et al., 2020). Our findings provide striking empirical evidence for this pseudogene-to-lncRNA transition in polyploid crops. Unlike the gradual accumulation of neutral point mutations that drives the slow functional divergence of paralogs under the canonical neofunctionalization model (Ohno, 1970; Conant and Wolfe, 2008), this single large-scale structural rearrangement event acted as an “evolutionary leapfrog.” It simultaneously pseudogenized the coding sequence, introduced novel *cis*-regulatory elements natively inherited from the captured D11 sequence, and repurposed the duplicate into an epigenetic regulatory non-coding node. By bypassing the traditional slow divergence trajectory of duplicated genes, this D11-derived inter-chromosomal sequence capture provides a compelling evolutionary model for how gene duplication coupled with ectopic genomic rearrangements can instantaneously orchestrate the *de novo* birth of complex regulatory lncRNAs within the dynamic genomes of allopolyploid crops.

4.2 Evolution of *pGhFAD2-1* into a lineage-specific epigenetic regulatory node

Following the lineage-specific structural innovation, a critical evolutionary question arises regarding how a pseudogenized locus rapidly acquired the capacity to function as a sophisticated epigenetic regulator. The functionalization of *de novo* originated lncRNAs typically necessitates a protracted evolutionary period, relying on the gradual accumulation of point mutations to generate functional transcriptional promoters and RNA interaction domains (Schlötterer, 2015; Statello et al., 2021). However, the molecular architecture of *pGhFAD2-1* illustrates an alternative and highly efficient trajectory. The interchromosomal sequence capture event essentially donated an intact regulatory module, conferring dual functionalities simultaneously and thereby bypassing the slow divergence trajectory of paralogs.

At the transcriptional level, lipid accumulation during seed development is finely coordinated by a complex transcriptional regulatory network (Ruuska et al., 2002; Zhao et al., 2018). Within this network, basic leucine zipper (bZIP) transcription factors recognize specific ACGTCBOX motifs to fine-tune metabolic flux (Armstrong et al., 1992; Mendes et al., 2013; Yang et al., 2023). The captured D11 sequence natively harbored this bZIP-responsive module, directly endowing the pseudogenized locus with DNA-level transcriptional competence during the critical window of seed maturation. Because the A-genome species naturally lack this ectopic insertion, the recruitment of transcription factors such as bZIPs to this genetic node is strictly confined to the D-lineage. This structural presence and absence variation elegantly explains the

lineage-specific transcriptional activation of this regulatory module without the necessity to invoke independent *cis*-element evolution.

Furthermore, the functional significance of this sequence capture extends beyond typical transcriptional activation and extends directly to epigenetic remodeling. The co-captured sequence inherently contains the HEXMOTIF, which is a conserved *cis*-element linked to chromatin state modification. As validated by our previous functional experiments (Chen et al., 2025), the *pGhFAD2-1* RNA transcript strictly requires this specific exonic region to serve as a functional scaffold for recruiting the histone deacetylase GhHDT1. The concurrent acquisition of these functional motifs implies that the structural insertion event endowed the locus with an integrated and immediately active interaction platform (Ariel et al., 2014).

Collectively, this structural reorganization catalyzed a synergistic dual functionalization process. By simultaneously acquiring DNA-level responsiveness for spatiotemporal transcription and an RNA-level interface for epigenetic protein recruitment, the *pGhFAD2-1* locus was instantaneously equipped with the molecular capacity to remodel the local histone acetylation landscape and suppress cottonseed lipid biosynthesis. This integrated mechanism not only elucidates the genetic basis for the phenotypic divergence in fatty acid composition between D and A genome species but also highlights how single large-scale genomic rearrangements can rapidly assemble complex epigenetic nodes in allopolyploid crops.

4.3 Purifying selection and asymmetric subgenome retention of *pGhFAD2-1*

While the vast majority of recently emerged lncRNAs exhibit rapid sequence turnover and are continuously purged from the genome via neutral drift (Hezroni et al., 2015), our molecular evolutionary assessments confirm that the *pGhFAD2-1* locus has been subjected to strong purifying selection and has remained highly conserved within the D-genome lineage (Ponjavic et al., 2007). This robust conservation indicates that the structural variation and the subsequent acquisition of epigenetic regulatory capacity conferred a distinct adaptive advantage. This evolutionary trajectory perfectly exemplifies how newly emerged non-coding transcripts that become physically integrated into essential epigenetic complexes face stringent functional constraints, which actively prevent their sequence degradation (Ulitsky and Bartel, 2013). Consequently, these selective pressures dictated the conservative fixation of this locus in D-genome diploids and its subsequent vertical transmission to the Dt subgenome of allotetraploid cottons following allopolyploidization.

This asymmetric presence and absence pattern is a typical molecular imprint of the well-documented asymmetric subgenome evolution in allotetraploid cotton. Multiple pan-genomic studies have established that the Dt subgenome retained significantly more selection signatures associated with agronomic traits than the At subgenome during domestication (Zhang et al., 2015; Wang et al., 2017; Ma et al., 2018). However, these foundational models have predominantly concentrated on the asymmetric expression and retention of protein-coding genes. By elucidating the evolutionary trajectory of *pGhFAD2-1*, we extend this classic paradigm to lineage-specific non-coding regulatory elements. The exclusive retention of this locus serves as a striking example of how the Dt subgenome uniquely preserves epigenetic

repressors to fine-tune highly conserved metabolic pathways inherited from diploid ancestors. This extension underscores the critical, yet previously underappreciated, role of the non-coding genome in driving the subgenomic coordination and lineage-specific adaptation of allopolyploid crops.

4.4 Evolutionary trade-offs in carbon allocation drive the adaptive retention of *pGhFAD2-1*

Building on the signature of strong purifying selection and asymmetric subgenomic retention of *pGhFAD2-1* characterized above, the stringent evolutionary constraints acting on this locus translate directly into macroscopic phenotypic divergence in cottonseed lipid metabolism. Our phenotypic evaluations demonstrate that total seed oil content exhibits no significant difference between the diploid *Gossypium raimondii* (D5) and the allotetraploid *Gossypium hirsutum* (AD1), both of which harbor the functional *pGhFAD2-1* module. Conversely, the overall oil content and oleic acid levels in the A-genome diploid species, which naturally lack this epigenetic repressor, are significantly higher. This clear phenotypic divergence pattern strictly corroborates the negative regulatory role of the *pGhFAD2-1-GhHDT1* module. This lineage-specific repression aligns with the Dt subgenome-biased genomic landscape of cotton domestication, where modern resequencing studies have identified profound selection signatures at loci associated with agronomic traits and metabolic adaptation in the Dt subgenome (Su et al., 2016; Ma et al., 2018). Furthermore, the exclusive retention of this repressive module in the Dt subgenome is highly consistent with the extensive *cis*- and *trans*-regulatory evolution observed during the domestication of allotetraploid cotton (Bao et al., 2019).

The stable retention of a lipid-suppressing locus within the D-genome lineage presents an apparent evolutionary paradox: energy-dense triacylglycerol reserves are widely recognized to confer adaptive advantages for seed germination and seedling establishment, so why has this repressive locus been maintained by strong purifying selection? This counterintuitive pattern is best interpreted through the classic framework of carbon allocation trade-offs during plant adaptation (Ding and Chen, 2018). The biosynthesis of storage lipids is a metabolically expensive process that competes for finite photosynthetic carbon precursors with alternative sink tissues. In oilseed models such as *Brassica napus*, suppressing seed lipid accumulation has been shown to inversely divert carbon flux toward competing sink components (Chapman et al., 2008). Within the evolutionary context of *Gossypium*, we propose that the *pGhFAD2-1* locus functions as an adaptive metabolic checkpoint. By restricting excessive carbon drainage into storage oils, this epigenetic module potentially facilitates a more favorable metabolic background for the development of alternative carbon-demanding structures, specifically the highly specialized seed coat trichomes (cotton fibers). While the direct quantitative contribution of this specific locus to fiber elongation remains to be precisely validated through future genetic experiments, our macro-evolutionary framework suggests that the Dt subgenome may have co-opted this ancestral metabolic constraint that arose ~5 million years ago in the D-genome ancestor to accommodate the immense carbon demands of fiber cell development.

The characterization of this lineage-specific epigenetic repressor provides a clear theoretical pathway for breaking the long-standing negative genetic correlation between cottonseed lipid metabolism and fiber development. By utilizing targeted genome editing or epigenetic perturbation to precisely decouple this metabolic checkpoint, future molecular breeding strategies could optimize carbon partitioning, facilitating the synergistic improvement of both cottonseed oil nutritional value and fiber yield in dual-purpose elite cotton varieties.

5 Conclusion

By integrating comprehensive evolutionary and phenotypic analyses across 27 representative *Gossypium* species, this study systematically resolves the *de novo* origination and adaptive trajectory of the lineage-specific lncRNA *pGhFAD2-1*. We demonstrate that this locus emerged approximately 5 million years ago (MYA) through an *in situ* tandem duplication of *GhFAD2-1D*, followed rapidly by an interchromosomal sequence capture event unique to the D-genome lineage. Rather than succumbing to pseudogenization, this structural innovation catalyzed an instantaneous dual functionalization. By simultaneously acquiring cis-regulatory elements at the DNA level and a *GhHDT1*-binding scaffold at the RNA level, the locus was elegantly transformed into a critical epigenetic repressor. Subsequently, this newly assembled regulatory module experienced strong purifying selection, dictating its conserved fixation in D-genome diploids and its asymmetric retention within the Dt subgenome of allotetraploid cottons. This exclusive subgenomic retention reflects a macroscopic evolutionary trade-off in carbon resource allocation, effectively restricting excessive lipid biosynthesis to prioritize the metabolic demands of fiber development. Ultimately, these findings establish a comprehensive evolutionary paradigm for the *de novo* origination of functional lncRNAs in polyploid crops. Furthermore, they highlight the *pGhFAD2-1* metabolic checkpoint as a premier molecular target for decoupling lipid and fiber developmental networks, providing a robust theoretical foundation for breaking the long-standing negative genetic correlation between fiber yield and seed oil content, and enabling the synergistic genetic improvement of dual-purpose elite cotton varieties.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#). Further inquiries can be directed to the corresponding authors.

Author contributions

HC: Writing – original draft, Writing – review & editing. XL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources,

Software, Supervision, Validation, Visualization, Writing – original draft. NY: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. ZG: Writing – review & editing. FS: Writing – review & editing. SG: Writing – review & editing. JZ: Writing – review & editing. SQ: Writing – review & editing. JW: Writing – review & editing. YL: Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the Stable Support Project of Xinjiang Academy of Agricultural Sciences (xjnkywdzc-2026002-2-2); the Xinjiang Uygur Autonomous Region Tianshan Talent Training Program (Youth Support) (2024TSYCQNTJ0036); the Project Tianshan Talent Cultivation Plan of Xinjiang Uygur Autonomous Region (2023TSYCTD0003); and the Talent Introduction Project of Key Innovation Platforms (Bases) in the Science and Technology Field of Xinjiang Leading Talent Introduction Plan (XJRC-2025-KJ-YJ-CXPT-031).

Acknowledgments

We thank the Topic Editors and reviewers for their constructive comments and valuable suggestions on this manuscript. We also thank all members of our laboratories for their assistance with the experiments, data collection and bioinformatic analysis during this study.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer HC declared a past co-authorship with the author(s) JZ to the handling editor.

Correction note

A correction has been made to this article. Details can be found at: [10.3389/fpls.2026.1878870](https://doi.org/10.3389/fpls.2026.1878870).

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. The author(s) declare the use of

generative AI (Gemini 3 Pro) solely for correcting grammar and refining the English language of the final manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M., et al. (2014). Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Mol. Cell* 55, 383–396. doi: 10.1016/j.molcel.2014.06.011
- Armstrong, G. A., Weisshaar, B., and Hahlbrock, K. (1992). Homodimeric and heterodimeric leucine zipper proteins and nuclear factors from parsley recognize diverse promoter elements with ACGT cores. *Plant Cell* 4, 525–537. doi: 10.1105/tpc.4.5.525
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335
- Bao, Y., Hu, G., Grover, C. E., Conover, J., Yuan, D., and Wendel, J. F. (2019). Unraveling cis and trans regulatory evolution during cotton domestication. *Nat. Commun.* 10, 5399. doi: 10.1038/s41467-019-13386-w
- Chapman, K. D., Neogi, P. B., Hake, K. D., Stawski, A. A., Speed, T. R., Cotter, M. Q., et al. (2008). Reduced oil accumulation in cottonseeds transformed with a Brassica nonfunctional allele of a delta-12 fatty acid desaturase (FAD2). *Crop Sci.* 48, 1470–1481. doi: 10.2135/cropsci2007.11.0618
- Charif, D., and Lobry, J. R. (2007). “SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis,” in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Eds. U. Bastolla, M. Porto, H. E. Roman and M. Vendruscolo (Springer Berlin Heidelberg, Berlin, Heidelberg), 207–232. doi: 10.1007/978-3-540-35306-5_10
- Cheetham, S. W., Faulkner, G. J., and Dinger, M. E. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* 21, 191–201. doi: 10.1038/s41576-019-0196-1
- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.08.001
- Chen, H., Li, Y., Xiong, X., Liu, X., Xue, F., Sun, J., et al. (2025). pseudo-GhFAD2-1 is a lncRNA involved in regulating cottonseed oil content and seed size by recruiting histone deacetylase 1. *Plant Biotechnol. J.* 23, 5229–5251. doi: 10.1111/pbi.70291
- Chi, X., Yang, Q., Pan, L., Chen, M., He, Y., Yang, Z., et al. (2011). Isolation and characterization of fatty acid desaturase genes from peanut (*Arachis hypogaea* L.). *Plant Cell Rep.* 30, 1393–1404. doi: 10.1007/s00299-011-1048-4
- Conant, G. C., and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* 9, 938–950. doi: 10.1038/nrg2482
- Dar, A. A., Choudhury, A. R., Kancharla, P. K., and Arumugam, N. (2017). The FAD2 gene in plants: occurrence, regulation, and role. *Front. Plant Sci.* 8, 1789. doi: 10.3389/fpls.2017.01789
- Ding, M., and Chen, Z. J. (2018). Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. *Curr. Opin. Plant Biol.* 42, 37–48. doi: 10.1016/j.pbi.2018.02.003
- Glaszmann, J. C., Kilian, B., Upadhyaya, H. D., and Varshney, R. K. (2010). Accessing genetic diversity for crop improvement. *Curr. Opin. Plant Biol.* 13, 167–173. doi: 10.1016/j.pbi.2010.01.004
- Gouet, P., Robert, X., and Courcelle, E. (2003). ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* 31, 3320–3323. doi: 10.1093/nar/gkg556
- Grover, C. E., Arick, M. A., Thrash, A., Sharbrough, J., Hu, G., Yuan, D., et al. (2022). Dual domestication, diversity, and differential introgression in Old World cotton diploids. *Genome Biol. Evol.* 14, evac170. doi: 10.1093/gbe/evac170
- Harms, M. J., and Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* 14, 559–571. doi: 10.1038/nrg3540
- Heppard, E. P., Kinney, A. J., Steitz, K. L., and Miao, G. H. (1996). Developmental and growth temperature regulation of two different microsomal oleate desaturases in soybean. *Plant Physiol.* 110, 311–319. doi: 10.1104/pp.110.1.311
- Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122. doi: 10.1016/j.celrep.2015.04.023
- Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297
- Hillis, D. M., Moritz, C., and Mable, B. K. (1996). *Molecular Systematics*. 2nd ed (Sunderland, Massachusetts: Sinauer Associates).
- Huang, G., Huang, J.-Q., Chen, X.-Y., and Zhu, Y.-X. (2021). Recent advances and future perspectives in cotton research. *Annu. Rev. Plant Biol.* 72, 437–462. doi: 10.1146/annurev-arplant-080720-113241
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Nat. Rev. Genet.* 11, 746–759. doi: 10.1038/nrg2878
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285
- Kumar, S., and Gadagkar, S. R. (2001). Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* 158, 1321–1332. doi: 10.1093/genetics/158.3.1321
- Lakhssassi, N., Colantonio, V., Flowers, N. D., Zhou, Z., Henry, J., Das, S. P., et al. (2017). Characterization of the FAD2 gene family in soybean reveals the limitations of gel-based TILLING in genes with high copy number. *Front. Plant Sci.* 8, 324. doi: 10.3389/fpls.2017.00324
- Li, Y., Si, Z., Wang, G., Shi, Z., Chen, J., Qi, G., et al. (2023). Genomic insights into the genetic basis of cotton breeding in China. *Mol. Plant* 16, 662–677. doi: 10.1016/j.molp.2023.01.012
- Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187
- Liu, F., Ma, L., Wang, Y., Li, Y., Zhang, X., Xue, F., et al. (2019). GhFAD2-3 is required for anther development in *Gossypium hirsutum*. *BMC Plant Biol.* 19, 393. doi: 10.1186/s12870-019-2010-9
- Liu, Q., Brubaker, C. L., Green, A. G., Marshall, D. R., Sharp, P. J., and Singh, S. P. (2001). Evolution of the FAD2-1 fatty acid desaturase 5' UTR intron and the molecular systematics of *Gossypium* (Malvaceae). *Am. J. Bot.* 88, 92–102. doi: 10.2307/2657130
- Liu, Q., Singh, S. P., and Green, A. G. (2002). High-stearic and high-oleic cottonseed oils produced by hairpin RNA-mediated post-transcriptional gene silencing. *Plant Physiol.* 129, 1732–1743. doi: 10.1104/pp.001933

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2026.1841757/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Detailed nucleotide sequence annotation and structural features of the *pGhFAD2-1*. The full-length cloned transcript sequence (1,476 nt) is highlighted in yellow. The precise transcription start site (TSS) and transcription termination site (TTS) are indicated in bold red letters. Sequences sharing homology with the ancestral *GhFAD2-1* gene are underlined. The spliced intronic region is characterized by canonical GT-AG splice donor and acceptor sites. The 1,221-bp exogenous insertion, presumably originating from an ectopic genomic region, is denoted in red text.

SUPPLEMENTARY FIGURE 2

Sequence alignment between the 1,221-bp exogenous insertion of *pGhFAD2-1* and its homologous genomic region.

- Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1156. doi: 10.1126/science.290.5494.1151
- Ma, L., Cheng, X., Wang, C., Zhang, X., Fei, X., Li, Y., et al. (2021). Explore the gene network regulating the composition of fatty acids in cottonseed. *BMC Plant Biol.* 21, 446. doi: 10.1186/s12870-021-02952-4
- Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119-7
- Mendes, A., Kelly, A. A., Van Erp, H., Shaw, E., Powers, S. J., Kurup, S., et al. (2013). bZIP67 regulates the omega-3 fatty acid content of Arabidopsis seed oil by activating FATTY ACID DESATURASE3. *Plant Cell* 25, 3104–3116. doi: 10.1105/tpc.113.116343
- Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300
- Nixon, K. C., and Carpenter, J. M. (1993). On outgroups. *Cladistics* 9, 413–426. doi: 10.1111/j.1096-0031.1993.tb00212.x
- Ohno, S. (1970). *Evolution by gene duplication* (Berlin: Springer-Verlag).
- Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E., and Browse, J. (1994). Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell* 6, 147–158. doi: 10.1105/tpc.6.1.147
- Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., et al. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492, 423–427. doi: 10.1038/nature11798
- Pathak, D., Singh, S., Kumar, H., Grover, G., and Kaur, N. (2023). *Cotton: Some Insights* (Ludhiana: The Crop Improvement Society of India). doi: 10.5281/zenodo.8298304
- Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17, 556–565. doi: 10.1101/gr.6036807
- Ruuska, S. A., Girke, T., Benning, C., and Ohlrogge, J. B. (2002). Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* 14 (6), 1191–1206. doi: 10.1105/tpc.000877
- Schlötterer, C. (2015). Genes from scratch — the evolutionary fate of de novo genes. *Nat. Rev. Genet.* 16, 712–726. doi: 10.1038/nrg3947
- Schlueter, J. A., Vasylenko-Sanders, I. F., Deshpande, S., Yi, J., Siegfried, M., Roe, B. A., et al. (2007). The FAD2 gene family of soybean: insights into the structural and functional divergence of a paleopolyploid genome. *Crop Sci.* 47, S14–S26. doi: 10.2135/cropsci2006.06.0382tpg
- Smit, A. F. A., Hubley, R., and Green, P. (2013–2015). *RepeatMasker Open-4.0* (Seattle, WA, USA: Institute for Systems Biology).
- Statello, C., Guo, C. J., Chen, L. L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* 22, 96–118. doi: 10.1038/s41580-020-00315-9
- Su, J., Li, L., Pang, C., Wei, H., Wang, C., Song, M., et al. (2016). Two genomic regions associated with fiber quality traits in Chinese upland cotton under apparent breeding selection. *Sci. Rep.* 6, 38496. doi: 10.1038/srep38496
- Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135, 599–607. doi: 10.1093/genetics/135.2.599
- Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120
- Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* 17, 601–614. doi: 10.1038/nrg.2016.85
- Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154 (1), 26–46. doi: 10.1016/j.cell.2013.06.020
- Wang, W. (2016). “The molecular detection of *Corynespora cassicola* on cucumber by PCR assay using DNAMAN software and NCBI,” in *Computer and Computing Technologies in Agriculture IX. CCTA 2015. IFIP Advances in Information and Communication Technology*, vol. 479. Eds. D. Li and Z. Li (Springer, Cham). doi: 10.1007/978-3-319-48354-2_26
- Wang, M., Li, J., Qi, Z., Long, Y., Pei, L., Huang, X., et al. (2022). Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. *Nat. Genet.* 54, 1959–1971. doi: 10.1038/s41588-022-01237-2
- Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* 49, 579–587. doi: 10.1038/ng.3807
- Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., et al. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* 44, 1098–1103. doi: 10.1038/ng.2371
- Wu, M., Pei, W., Wedegaertner, T., Zhang, J., and Yu, J. (2022). Genetics, breeding and genetic engineering to improve cottonseed oil and protein: a review. *Front. Plant Sci.* 13, 864850. doi: 10.3389/fpls.2022.864850
- Wu, Z., Yang, Y., Huang, G., Lin, J., Xia, Y., and Zhu, Y. (2017). Cotton functional genomics reveals global insight into genome evolution and fiber development. *J. Genet. Genomics* 44, 511–518. doi: 10.1016/j.jgg.2017.09.009
- Yang, Y., Kong, Q., Tee, W. T., Li, Y., Low, P. M., Patra, B., et al. (2023). Transcription factor bZIP52 modulates Arabidopsis seed oil biosynthesis through interaction with WRINKLED1. *Plant Physiol.* 192, 2628–2639. doi: 10.1093/plphys/kiad270
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628
- Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207
- Zhao, Y., Li, J., Lian, B., Gu, H., Li, L., and Qi, Y. (2022). Global identification and characterization of long non-coding RNAs in plant evolution. *Nat. Plants* 8, 1342–1355. doi: 10.1038/s41598-020-67410-x
- Zhao, Y., Wang, Y., Huang, Y., Cui, Y., and Hua, J. (2018). Gene network of oil accumulation reveals expression profiles in developing embryos and fatty acid composition in Upland cotton. *J. Plant Physiol.* 228, 101–112. doi: 10.1016/j.jplph.2018.06.002