Biology Direct

https://doi.org/10.1186/s13062-025-00718-8

Article in Press

Integrative multi-omics analysis widens annotation and functional insights into long non-coding RNAs of *Arabidopsis thaliana*

Received: 13 May 2025

Accepted: 16 December 2025

Published online: 05 January 2026

Cite this article as: Vivek A., Kiran H., Sahu N. *et al.* Integrative multi-omics analysis widens annotation and functional insights into long non-coding RNAs of *Arabidopsis thaliana*. *Biol Direct* (2025). https://doi.org/10.1186/s13062-025-00718-8

AT Vivek, Harikumar Kiran, Namrata Sahu, Garima Kalakoti & Shailesh Kumar

Integrative multi-omics analysis widens annotation and functional insights into long non-coding RNAs of *Arabidopsis thaliana*

AT Vivek[1], Harikumar Kiran[1], Namrata Sahu[1], Garima Kalakoti[1], Shailesh Kumar[1,*]

[1]Bioinformatics Lab, BRIC-National Institute of Plant Genome Research(NIPGR), Aruna Asaf Ali Marg, New Delhi 110067, India

***Corresponding author:** Shailesh Kumar, shailesh@nipgr.ac.in

Abstract

Background: Long non-coding RNAs (lncRNAs) play key roles in regulating plant growth, development, and stress responses. Despite their increasing identification in plant transcriptomes, a systematic characterization of lncRNAs is still lacking, leaving a significant knowledge gap. To address this, we systematically identified and characterized Arabidopsis lncRNAs through integrative analysis of strand-specific RNA sequencing data and multi-omics datasets, revealing their genomic features, regulatory interactions, and evolutionary characteristics.

Results: Using a custom pipeline applied to hundreds of stranded RNA-seq datasets, we assembled a comprehensive catalog of 4,772 intergenic and antisense *Arabidopsis* lncRNAs. In comparing multiple key features of lncRNAs with those of protein-coding genes, we found that intergenic lncRNAs contain high transposable element-derived fragments and display broader TE diversity. Distinct DNA methylation and histone modification signatures further distinguished lncRNAs from protein-coding genes. We additionally uncovered R-loop connections and associations with sRNAs involved in post-transcriptional regulation and RNA-directed DNA methylation, with a minor subset classified as Pol V–transcribed. Of note, our results revealed lncRNAs mediating stress-responsive cis interactions and others linked to trait-associated loci. Probing further, an experimental evidence resource confirmed small peptide production from multiple lncRNA loci. Extending our investigation, comparative analyses across Brassicaceae species revealed syntenic lncRNAs enriched for shared sequence motifs despite substantial sequence divergence.

Conclusions: This study provides a valuable and extensively annotated catalog of *A. thaliana* lncRNAs, revealing their diverse genomic features, regulatory interactions, and evolutionary characteristics. Altogether, our work advocates for multi-omics integrative analysis as a potent strategy to efficiently enhance lncRNA annotation, providing insights into functionality and addressing annotation limitations. Our comprehensive bioinformatic analyses of *Arabidopsis* lncRNAs pave the way for future functional characterization of these transcripts.

Keywords: Long non-coding RNAs; Strand-specific RNA-seq; Multi-omics; Bioinformatics; *Arabidopsis thaliana*; Epigenetics; RdDM pathway; lncRNA-sRNA interaction; Evolution

Background

LncRNAs are key regulators in plants, integral to growth, development, and stress responses via epigenetic, transcriptional, and post-transcriptional mechanisms [1–4]. Although many lncRNA loci have been identified in plant transcriptomes, efforts to characterize them remain limited, resulting in a substantial knowledge gap. Currently, fewer than 1% have been

experimentally characterized [5]. Functional elucidation requires high-quality transcriptome annotations integrated with complementary genomic approaches, yet these resources remain rudimentary. As a result, there is uncertainty in the total number of lncRNAs and incomplete transcript structures for many of those identified.

Extensive RNA-sequencing data, generated through a range of sequencing methodologies, have been archived in public databases. Indeed, numerous projects using varied approaches have significantly contributed to the expansion of both the quantity and scope of available lncRNA annotations. Despite the substantial acceleration in discovering novel lncRNAs, facilitated by massively parallel RNA sequencing technologies, efforts in functional annotation are lagging [6]. Primarily, this is due to the ongoing challenge of balancing throughput and quality in annotation endeavors. Consequently, there is a critical need to integrate newly reported lncRNAs and comprehensively characterize them across multi-omics levels within diverse biological contexts.

With rising interest in lncRNAs, identifying effective algorithms for their detection is crucial, enabling robust method comparisons and adherence to best practices. LncRNAs have unique features that make accurate classification challenging. Therefore, RNA-seq pipelines require specialized considerations [7, 8]. For instance, antisense lncRNAs pose challenges for alignment-based quantification, where stranded protocols and pseudoalignment methods offer improvements [9, 10]. Most RNA-seq samples use non-standard library preparation protocols, so pseudo-alignment methods provide a cost-effective approach to improving lncRNA annotation. However, they require many samples to accurately determine strand information.

In this study, we apply large-scale transcriptomics to identify and analyze lncRNAs in *A. thaliana* using a pseudoalignment-guided approach. We characterize the key properties of the identified lncRNAs, expanding their annotations and elucidating their potential functional roles through the integration of multi-omic data. In addition, we investigate the evolutionary relationships of lncRNAs with other *Brassicaceae* species through comparative genomics. Overall, this work provides a comprehensive view of the functional and evolutionary landscape of *Arabidopsis* lncRNAs, providing extensive annotations that will guide future functional analyses into this “dark matter” of the transcriptome.

Methods

Datasets, preprocessing, mapping, and assembly

To comprehensively characterize *A. thaliana* lncRNAs, we utilized publicly available RNA-seq datasets from the NCBI Sequence Read Archive (SRA) (Supplementary Table S1). We retrieved libraries using fastq-dl (v1.2.0) (https://github.com/rpetit3/fastq-dl) and performed quality trimming with TrimGalore (v0.6.10) (https://github.com/FelixKrueger/TrimGalore). Strand-specificity of the libraries was assessed through transcriptome extraction using RSEM prepare-reference (v1.3) [11] and determination of strandedness with Salmon (v0.8.12) [12]. Non-strand-specific libraries were excluded to prevent potential confounding of lncRNA expression by reads corresponding to other genomic features located on the opposite DNA strand. The refined set of libraries underwent read mapping to the reference genome TAIR10 (assembly 22) [13] using HISAT2 (v2.1.1) [14].

Computational prediction of *Arabidopsis* lncRNAs

We applied a genome-guided transcriptome assembly approach using Stringtie (v1.3.3) [15] to reconstruct transcriptomes from individual RNA-seq data. In parallel, we conducted a comparative analysis between these assemblies and de novo transcriptome reconstructions generated by Trinity (v2.8.5) [16]. To assess assembly accuracy, we used the TAIR10 reference annotations as benchmarks. Subsequently, we integrated transcriptome assemblies from all samples within each species by aggregating individual StringTie-generated GTF files using StringTie merge with the parameter *-g 50*. We then compared these merged transcriptomes with the native genome annotations of each species using gffcompare (v0.11.2) [17] to identify novel transcripts. Specifically, we focused on novel intergenic ("i") and antisense ("a") transcripts longer than 200 bp, corresponding to the "u" and "x" class codes, respectively, in the gffcompare output. In cases where multiple isoforms were present, we retained only the longest isoform using CGAT gtf2gtf (v0.3.2) [18]. We further evaluated the coding potential of predicted transcripts using the CPC2 (v0.9) [19] and FEELnc (v0.1.1) [19] software packages. CPC2 analysis was performed against TAIR10-annotated protein sequences, and transcripts classified as noncoding were retained for further analysis. For FEELnc, protein-coding gene sequences processed in shuffle

mode served as training datasets for the Random Forest algorithm, enabling the determination of coding potential thresholds through tenfold cross-validation.

Our classification framework integrated evidence from transcriptional expression and computational noncoding potential. Noncoding potential was assessed using CPC2 and FEELnc, with transcripts classified as “CPC + FEELnc” when both tools concurred or as “single-tool supported” when supported by only one. Expression support was determined using TPM values derived from StringTie quantification across all RNA-seq samples, with TPM > 0 indicating expressed transcripts. Each lncRNA was subsequently assigned to a confidence category: “High (Stringent)” for those with dual noncoding support and detectable expression; “High (Stringent, No Expression)” for those with dual noncoding support but no detectable expression; and “Medium (Single-tool + Expression Supported)” for those with single-tool support and detectable expression. Transcripts lacking evidence from both criteria were excluded from the final set.

lncRNA distribution and comparison to other established catalogs

We used chromPlot (v1.14.0) [20] to visualize the distribution of lncRNAs across the *Arabidopsis* genome. Comparisons were conducted between the lncRNAs identified in this study and those cataloged in CANTATAdb 2.0 [21], PLncDB v2.0 [22], EVLncRNAs 2.0 [5], and TAIR 10 (Araport 11 annotated) [23] databases. We also compared our findings with lncRNA collections from [24] and [25]. Locus coordinates were evaluated using the intersectBed function (f=0.5) within the BEDTools (v2.30.0) [26].

Expression estimation

To assess the overall expression levels of lncRNAs and compare them with protein-coding genes, we used featureCounts (v1.6.4) [27] to compute read counts for both transcript categories across all analyzed samples. The count data were then normalized by transcript length and library size to generate transcripts per million (TPM) values. Transcriptomes were categorized into broad tissue types based on their sample origin, while accounting for cases where certain tissues, such as whole plants and seedlings, did not fit standard classifications. To identify tissue-specific

lncRNAs, we calculated the tau metric [28] using normalized TPM expression values across tissues.

Analysis of sequence variation

To assess and compare the variability of lncRNA sequences with that of protein-coding genes and intergenic regions, we used variant-calling data from the 1001 Genomes Project [28, 29]. Using the BEDTools *intersect* function, we determined the number of variants present in both classes of lncRNAs, as well as in protein-coding genes and intergenic regions. Intergenic regions were identified using the BEDTools *complement* function applied to genome annotation files generated with gffcompare.

Analysis of genomic repeats and TE elements

The *Arabidopsis* genome was analyzed for repeat elements using perSVade (v0.10) pipeline [30] with default settings. We examined various repeat categories, including simple repeats, low-complexity repeats, LTRs, LINEs/SINEs, and those classified as unknown. To assess overlap, we required at least 50% of the repeat sequence to intersect with either lncRNAs or protein-coding genes. To identify transposable element (TE) fragments within lncRNAs and protein-coding genes, we compared TAIR10-annotated TE sequences with individual loci using BLASTN (v2.10.0) [31]. We applied specific parameters, including a word size of 10, evaluation of both strand orientations, and an e-value threshold of $1e-7$, ensuring more than 80% sequence identity. The resulting hits were grouped into seven TE superfamilies.

Analysis of DNA methylation, histone modification, and R-loop formation

We analyzed the epigenetic profiles of lncRNAs and protein-coding genes by reprocessing publicly available datasets, including whole-genome bisulfite sequencing (WGBS), chromatin immunoprecipitation sequencing (ChIP-seq), RNA immunoprecipitation sequencing (RIP-seq), global run-on sequencing (GRO-seq), and single-strand DNA ligation from DNA:RNA hybrid immunoprecipitation sequencing (ssDRIP-seq) (Supplementary Table S2), to investigate DNA methylation, histone modifications, RdDM pathway associations, and R-loop connections, respectively.

Raw paired-end FASTQ files underwent quality trimming using fastp (v0.20.0) [32] to remove adapter sequences and low-quality reads. Reads from ChIP-seq libraries were then aligned to the reference genome assembly using the BWA-MEM algorithm (v0.7.17) [33], followed by filtering of mapped reads with a MAPQ quality score below 30 and removal of PCR duplicates using samtools (v1.16) [34] to ensure high-quality alignments. WGBS reads were mapped with BatMeth2 [35] under default parameters, and DNA methylation levels were called from uniquely mapped reads, requiring a minimum coverage of three at each cytosine. Additionally, GRO-seq, RIP-seq, and ssDRIP-seq reads were aligned to the TAIR10 genome using BWA with default settings. Duplicates, unmapped reads, reads with more than three mismatches, and non-uniquely mapped reads were filtered out using samtools. Finally, indexed BAM files were converted to BigWig format using bamCoverage, and metaplot profiles were generated using plotProfile from deepTools [36]. Pairwise correlations between histone modification ChIP-seq datasets were calculated using Pearson correlation coefficients (PCCs) derived from normalized ChIP signal intensities (BigWig coverage) across ±3 kb regions flanking TSS and TES of each categorized transcript class.

Characterization of polyadenylation status and RNA polymerase dependence

Publicly available RNA-seq datasets were re-analyzed to determine the polyadenylation status and RNA Polymerase V (Pol V) dependency of lncRNAs in *A. thaliana* seedlings (Supplementary Table S3). To profile polyadenylated (poly(A)+) transcripts, Oxford Nanopore long-read RNA-seq data from 17 samples encompassing seedlings and floral buds aged 6–35 days were obtained from multiple studies (Zhang et al., 2020; Parker et al., 2020; Qin et al., 2022; Xu et al., 2022). To capture non-polyadenylated (non-poly(A)) RNAs, rRNA-depleted Illumina short-read RNA-seq data from eight samples representing 7-day- and 47-day-old seedling tissues were used (Philips et al., 2020). In addition, Illumina total RNA-seq data from 2-week-old mock-treated Col-0 (wild type, WT) and *nrpe1-11* mutant seedlings (three biological replicates each; Yuan et al., 2024) were analyzed to identify Pol V-dependent lncRNAs. Adapter sequences in Nanopore reads were trimmed using Porechop [37], and all datasets were aligned to the *Arabidopsis* TAIR10 reference genome using HISAT2 (v2.2.1) with default parameters. Read counts per lncRNA were quantified using featureCounts in stranded mode and normalized to TPM.

The polyadenylation ratio for each lncRNA was calculated as: poly(A) ratio = mean TPM (poly(A)+) / [mean TPM (poly(A)+) + mean TPM (non-poly(A)) + 1×10^{-6}]. Based on this ratio, lncRNAs were classified as poly(A)+ (ratio ≥ 0.7), non-poly(A) (ratio ≤ 0.3), or bimodal/intermediate (0.3 < ratio < 0.7). To identify polymerase-dependent lncRNAs, differential expression analysis was performed using DESeq2 (v1.38.0) on raw count data from WT and *nrpe1*. LncRNAs with an adjusted p-value (padj < 0.05) and a $\log_2$ fold change < –1 were considered significantly downregulated. Pol V–dependent lncRNAs were stringently defined as transcripts that were both non-polyadenylated and significantly downregulated in the *nrpe1* mutant. We also examined overlap with previously annotated Pol V-transcribed regions [38], taking into account both polyadenylation status and differential expression.

Identification of lncRNA and sRNA association

Multiple methods were employed to identify lncRNA loci involved in sRNA generation or regulation in *Arabidopsis*, with a focus on links to sRNA precursors. To identify putative miRNA binding sites using miRNAs from Plant Small RNA Genes (PSRG) database [39], lncRNAs and mRNAs were scanned with psRNATarget [40] (Schema V2, 2017 Release) using default parameters, except that the number of top targets was set to 5. LncRNAs and mRNAs with putative miRNA binding motifs were compared, and the target genes were functionally annotated using MapMan BIN ontology via Mercator (v4) [41]. Pathway enrichment was assessed with ShinyGO (v0.8) [42], and pathways with an adjusted p-value < 0.05 were considered significant. We then compared these sRNA-producing loci with existing annotations by examining the genomic locations of Pol IV/V–codependent sites and Pol IV–independent Pol V sites identified via Pol V ChIP–seq and GRO–seq in *nrpd1* mutant in previous studies [43]. Additionally, we considered Pol V transcripts annotated by [38] and identified sRNAs those bound by AGO4 using AGO4 RIP sRNA-seq data processed with ShortStack (v3.8.5) [44]. The PSRG small RNA data were used as loci files, and sRNA-producing loci with >1 FPKM were annotated as AGO4-bound.

lncRNA-chromatin interactions

We applied the PATO [45] tool to screen for potential DNA:DNA:RNA triplex sites, aiming to detect lncRNA-DNA triplexes. To further investigate lncRNA-mediated RNA–chromatin interactions, we referenced genome-wide RNA–DNA interaction data obtained from GRID-seq [46]. Additionally, we examined the identified lncRNA-chromatin interactions for intersections with transcription factor binding motifs (TFBMs) within Accessible Chromatin Regions (ACRs) data from PlantCADB [47].

Prediction of sORFs and small peptides

We identified small open reading frames (sORF) (18-300 nt) within lncRNAs and mRNA transcripts using Orfipy (v0.0.4) [48] selecting for those capable of encoding micropeptides (6-100 aa). Start codons considered were ATG, TTG, GTG, CTG, AAG, AGG, ACG, ATA, ATT, and ATC, and stop codons were TAG, TAA, and TGA, evaluated across all three reading frames [49]. Additionally, we performed BLASTp searches against the Arabidopsis PeptideAtlas [50] to identify experimentally validated peptides corresponding to the predicted sORFs.

Classification of lncRNA genes into families in Brassicacea

To explore the evolutionary connections among predicted lncRNAs in Brassicaceae species, we applied multiple complementary approaches, including reciprocal BLAST hits, secondary structure similarity assessments, and synteny analysis. Considering the potential influence of overlapping features in antisense lncRNAs, we restricted our analyses to intergenic genes.

To identify the best reciprocal matches between all possible pairs of species, we used BLASTn by constructing a custom database for each species-specific lncRNA set. Each set of lncRNAs was aligned against its respective database with an e-value cutoff of $<1e-3$, and parameters "-max_hsps 1" and "-max_target_seqs 1" were applied to select the best alignment for each query–sequence pair, from which the best reciprocal hits were chosen. To examine secondary structure similarities among intergenic lncRNAs, RNAfold (v2.4.10) [51] was used to predict secondary structures for all intergenic lncRNAs, followed by pairwise

structural alignments within and between species using Beagle (v0.276) [52] in local alignment mode. Like the BLASTn reciprocal hit strategy, we selected hits with the highest zScores (minimum zScore >3) and p-values <0.01. Furthermore, intergenic lncRNAs were categorized into syntenic transcripts following the method described in [53]. Orthologs among protein-coding genes across five Brassicaceae species were identified using OrthoFinder (v2.5.4) [54], and pairwise syntenic relationships among lncRNAs were determined using the synteny analysis pipeline from [53]. To this, we set “3 3 1” parameter requiring 3 protein-coding genes on either side of a given intergenic lncRNA, a minimum of 3 shared genes for pairwise comparisons between Brassicacea species, and at least 1 shared gene on each side of an intergenic lncRNA. Finally, results from BLAST, secondary structure alignments, and synteny-based comparisons were integrated to cluster the best reciprocal hits into lncRNA families across Brassicaceae species.

Finding sequence motifs in lncRNA syntenic families

We applied two methods to identify shared sequence motifs among syntenic lncRNA families. First, significant motifs (e-value < 0.05) were identified using MEME [55]. Second, conserved motifs (p- and e-values < 0.05) were detected using lncLOOM [56]. A syntenic family was considered to share a motif if every member contained at least one of these significant motifs.

Results

Inference and characterization of *Arabidopsis* lncRNA catalog

To characterize the lncRNA repertoire of *A. thaliana*, we assembled a comprehensive dataset of 727 samples from publicly available RNA-seq libraries. We processed these samples using lncFETCHER pipeline, a systematic framework that integrated multiple RNA-seq datasets with a combination of robust, established methodologies [10, 57–60]. The pipeline also incorporated a pseudoalignment approach paired with full transcriptome annotation, as recommended by [9]. The specific parameters used in this analysis were detailed in Supplementary Table S4. This pipeline included data integration, trimming, strand detection, and read mapping. These steps were followed by merging the assembled transcripts and cross-referencing them with existing annotations. We then assessed their coding potential to identify putative lncRNAs. These procedures ultimately produced a consolidated catalog of predicted lncRNAs through genome-guided transcriptome assembly, as illustrated in Fig. 1A. The

classification of these genes followed strict criteria, categorizing them as antisense, which overlap coding genes or other features on the opposite DNA strand, or intergenic, which do not intersect with any protein-coding genes or features.

Fig. 1. Identification and characterization of lncRNAs in *A. thaliana*. (**a**) Schematic overview of the pipeline used for lncRNA prediction and analysis. (**b**) Sensitivity and specificity of samples used in the final transcriptome assemblies (n=727). (**c**) Comparison of StringTie and Trinity assemblies for total transcripts, unique transcripts, and the percentage of unique transcripts. (**d**) Genomic distribution of antisense (blue, left) and intergenic (grey, right) lncRNAs. Black vertical lines denote chromosomes, and black circles mark centromeres. Each bar shows the number of lncRNAs within 50-kb genomic windows. (**e**) Upset plot showing overlap among major lncRNA annotation databases using a cumulative stepwise intersection method. Vertical magenta bars represent intersection sizes, and horizontal bars indicate total gene counts per source. Black dots denote shared or unique gene sets. (**f**) Venn diagram showing overlap between the current lncRNA catalog and those reported in recent studies ([25] and [24]). SR, short-read; LR, long-read–supported.

Before examining the lncRNA repertoires, we evaluated the effectiveness of transcript assembly methods for identifying lncRNAs by comparing the genome-guided approach using StringTie with the de novo approach using Trinity. We assessed their sensitivity, specificity, transcript coverage, redundancy, and structural complexity. Previous studies in other species [59, 61] had documented differences between these assemblers, which motivated us to examine whether similar effects also applied to *Arabidopsis* samples. Both methods produced consistent results; however, StringTie demonstrated superior specificity, sensitivity, and lower fragmentation rates (Fig. 1B–C), making it more precise in reconstructing transcripts, even in the absence of lncRNAs in the benchmark annotations. We inferred that the advantage of StringTie stemmed from its ability to leverage genomic information effectively. In contrast, Trinity exhibited higher sensitivity, capturing a broader range of potential lncRNAs, but at the cost of increased false positives and greater redundancy. Further analysis of assembly characteristics (Fig. 1C) showed that Trinity achieved higher total transcript coverage, suggesting more comprehensive reconstruction, while StringTie produced assemblies with reduced redundancy and a higher

proportion of multi-exonic lncRNAs, indicating a more refined and biologically relevant output. Given StringTie's overall efficiency in reconstructing reference transcripts with greater precision and lower fragmentation, we opted for the genome-guided assembly approach for downstream analyses.

Using this optimized framework, we identified 7,895 candidate lncRNAs categorized based on concordant noncoding potential and expression support. Of these, 4,772 transcripts (4,746 with expression support and 26 lacking detectable expression but classified as noncoding by both CPC2 and FEELnc) were classified as high confidence, constituting the final lncRNA catalog used for all downstream analyses (Supplementary Table S1). This catalog comprised 1,735 antisense and 3,037 intergenic transcripts.

We analyzed the genome-wide distribution of lncRNAs across the genome of *A. thaliana* (Fig. 1D). Antisense lncRNAs exhibited a relatively uniform distribution with a slight preference for chromosome arms, whereas intergenic lncRNAs displayed an uneven distribution with distinct hotspots near centromeres. These observations aligned with previous studies [25, 62], suggesting preferential genomic localization. Antisense lncRNAs were enriched in gene-rich regions, particularly near protein-coding loci, supporting their potential roles in cis-regulation, chromatin remodeling, and transcript stability. In contrast, intergenic lncRNAs were more widely dispersed, reflecting their diverse genomic origins. Chromosome-specific variations in lncRNA density revealed regions with concentrated intergenic and antisense lncRNAs, suggesting potential regulatory hotspots. These findings indicated a structured and functionally diverse genomic landscape of *Arabidopsis* lncRNAs.

We then focused on antisense and intergenic lncRNAs, comparing them with existing lncRNA databases to evaluate classification consistency and coverage. We identified a substantial number of lncRNAs (Fig. 1E), with the most notable differences observed among antisense transcripts, likely due to variations in annotation methodologies, sequencing depth, and dataset selection compared to other available databases. Some previously annotated lncRNAs also exhibited length discrepancies or biotype reclassifications, reflecting differences in transcript assembly strategies. We detected overlaps with well-characterized lncRNAs such as TAR-224, npc48, At5NC055270, and npc83, previously cataloged in EVLncRNAs v2.0. However, biotype inconsistencies were present; for instance, npc48, classified as antisense in prior studies, was annotated as intergenic in our dataset, indicating the complexity of lncRNA classification and the need for standardization.

To further assess the coverage of our catalog, we compared it with two previous annotation efforts [24, 25] (Fig. 1F). Approximately 25% of our identified lncRNAs overlapped with [25], supporting their presence. Additionally, we examined lncRNAs identified using both short-read and long-read sequencing from [24]. Among the 229 short-read–derived lncRNAs, most were intergenic, aligning with our findings, while 45 of the 53 long-read–supported (Oxford Nanopore sequencing) lncRNAs also belonged to the intergenic category, further supporting the reliability of the approach (Supplementary Table S5). However, despite these overlaps, substantial discrepancies remained between different catalogs, largely because differences in transcript assembly strategy, sequencing depth, and dataset selection continued to influence lncRNA annotation.

Overall, these analyses generated an updated and well-supported lncRNA catalog for *A. thaliana*, providing improved accuracy and broad genomic coverage. At the same time, the remaining discrepancies among catalogs showed that methodological differences continue to influence lncRNA annotation.

Characteristics of *Arabidopsis* lncRNAs

We analyzed the key characteristics of lncRNAs to identify features that could enhance future lncRNA detection efforts and to verify consistency with prior studies. First, we conducted a systematic comparative analysis of their transcript length and GC content, juxtaposed with those of protein-coding genes. The analysis revealed that both types of lncRNAs were shorter in length (Fig. 2A) and possessed lower GC content (though higher than that of intergenic regions) (Fig. 2B). Subsequently, we examined lncRNA sequence variability and found that protein-coding genes showed lower SNP density, whereas lncRNAs exhibited a higher mutation frequency overall. Among lncRNAs, intergenic lncRNAs had the highest accumulation of variants, whereas antisense lncRNAs displayed lower SNP density, like protein-coding genes (Fig. 2C). In addition, lncRNAs clearly exhibited lower expression levels compared to protein-coding genes (Fig. 2D).

Fig. 2. Genomic features, expression profiles, and repeat/TE associations of *Arabidopsis* lncRNAs. (a–c) Comparisons of transcript length, GC content, and sequence variation. (**d**) Comparison of overall expression and tissue-specific expression across six tissues (flower, fruit,

leaf, root, seed, shoot). Panels (a–d) compare three transcript categories: protein-coding genes ("pc"), antisense lncRNAs ("a"), and intergenic lncRNAs ("i"), alongside intergenic genomic regions where applicable. (**e**) Density distribution of tissue-specificity (τ) values for lncRNAs and protein-coding loci. (**f**) Pairwise Jaccard similarity of genomic regions exhibiting expression peaks across tissues for protein-coding genes (left) and lncRNAs (right). (**g**) Percentage of lncRNAs overlapping repeat regions; numbers above bars indicate the percentage, and numbers below bars indicate the count of overlapping lncRNAs (individual lncRNAs may overlap multiple repeat categories). (**h**) Distribution of TE superfamilies identified within intergenic, antisense, and protein-coding loci based on BLAST matches to TAIR10-annotated TEs. Differences between pairs of box plots within each species are statistically significant (Wilcoxon rank-sum test, $p < 0.01$) for panels (a–d), unless indicated otherwise. In panel (**h**), only statistically significant comparisons are indicated with asterisks.

To assess the expression dynamics of lncRNAs, we compared their abundance across representative tissues (Fig. 2D). Among these, 99.47% of the lncRNA genes in our curated list were expressed in at least one tissue, although expression levels varied substantially between organs. Expression patterns differed across tissues, with intergenic lncRNAs showing the widest range of variability. Their expression was generally higher in flowers and reduced in fruit, whereas antisense lncRNAs displayed relatively uniform levels across tissues. The marked reduction of intergenic lncRNAs in fruit likely rendered antisense transcripts comparatively more abundant in this tissue, although the difference was not statistically significant. Together, these observations indicate stable antisense expression and tissue variability among intergenic lncRNAs.

We further calculated a tissue specificity score for each transcript using the Tau (τ) metric, utilizing expression data from broadly categorized tissue samples. This τ metric (ranging from 0 to 1) measures how closely the tissue expression of a transcript matches a perfectly tissue-specific pattern ($\tau = 1$), where expression is limited to one tissue (Fig. 2E). The analysis showed consistent tissue specificity in lncRNA gene expression across various tissues. Intergenic lncRNAs exhibited significantly higher tissue specificity compared to their mRNA counterparts, aligning with previous findings. This suggests regulatory nuances or context-specific functions of intergenic lncRNAs, distinguishing them from antisense lncRNAs and protein-coding genes. Subsequently, we determined the relationship between expression peaks in tissues by performing

a pairwise intersection of genomic regions (Fig. 2F). These results revealed relatively similar expression patterns of protein-coding genes and lncRNAs in leaf and flower. There was high pairwise overlap between expressed protein-coding genes across all tissues, whereas the overlap of expressed lncRNAs showed marked differences, with notably reduced observation in fruit. This analysis revealed significant differences in lncRNA expression across tissue types.

Taken together, Arabidopsis lncRNAs exhibit shorter lengths, lower GC content, and greater transcript diversity. Their lower and tissue-specific expression profiles, as revealed by our tissue expression atlas, further establish the context-specific nature of their regulation.

Intergenic lncRNAs are TE-rich

Previous studies have shown that lncRNAs across diverse species often contain TEs and genomic repeats, which can influence their emergence, structural organization, evolutionary patterns, and regulatory mechanisms. Using two approaches, we therefore sought to assess the extent to which the identified lncRNAs harbor these elements compared to protein-coding genes.

First, our approach involved reannotating TEs and repeats and intersecting them with our established lncRNA catalog and protein-coding genes (Fig. 2G). The results revealed that many lncRNAs, whether intergenic or in antisense orientations, contain repeats, albeit with varying prevalence levels. Simple repeats were the most abundant repeat type across all categories, being particularly prevalent in protein-coding genes. In contrast, antisense and intergenic lncRNAs showed markedly lower repeat content. LINE, SINE, and LTR elements were generally rare, although intergenic lncRNAs displayed a mild presence of LTRs. More than half of the lncRNAs lacked overlap with repetitive elements, suggesting that lncRNAs, especially intergenic, tend to originate from repeat-poor regions or are under selective pressure against repeat insertions.

Second, in a BLAST-based analysis against TAIR10-annotated TEs, approximately half of the intergenic lncRNA loci contained TE-derived sequences, designated as ‘TE pieces’ (Fig. 2H). Intergenic lncRNAs had the highest TE content and the broadest range of TE superfamilies, with LTR/Gypsy and Helitron elements being the most frequent among TE-containing transcripts. In contrast, antisense lncRNAs had the lowest TE prevalence. Consistent with this, antisense lncRNAs also contained the smallest TE-derived segments, supporting their overall low TE

burden. Protein-coding genes showed intermediate TE content and the shortest associated TE regions. Together, these findings indicate a preference and distinct composition of TEs in intergenic lncRNAs compared with antisense lncRNAs and protein-coding loci.

Collectively, these results show that the insertion of repetitive and TE-derived sequences into lncRNA loci is not random but preferential. The abundance of diverse TE families in intergenic lncRNAs points to distinct evolutionary trajectories.

Distinct methylation and histone modifications underpin divergent chromatin states

We compared DNA methylation across lncRNAs and protein-coding genes in seedlings in the CG, CHG, and CHH contexts (Fig. 3A). Intergenic lncRNAs showed the highest CHG and CHH methylation, consistent with a more repressive chromatin state. Antisense lncRNAs and protein-coding genes had much lower CHG methylation, and antisense lncRNAs showed only a small increase in CHH methylation relative to protein-coding genes. In the CG context, all transcript classes had clear drops in methylation at TSS and TES. The depth of this reduction was greatest in intergenic lncRNAs and protein-coding genes, whereas antisense lncRNAs displayed a slightly weaker decrease. This shared CG dip reflected reduced nucleosome occupancy and increased accessibility at promoters and transcript ends. Overall, intergenic lncRNAs were linked to high non-CG methylation. In contrast, antisense lncRNAs and protein-coding genes had lower non-CG methylation and similar CG dynamics, consistent with more permissive chromatin.

Fig. 3. Integrative landscape of DNA methylation, histone modifications, RdDM association, and R-loop profiles at lncRNA and protein-coding loci. (**a**) Metagene profiles of DNA methylation levels in CG, CHG, and CHH contexts. (**b**) Heatmap of absolute Pearson correlation coefficients from ChIP-seq read counts showing co-occurrence strength among histone modification marks. (**c**) Metagene profiles of RdDM and chromatin-associated factors, including Pol V (ChIP, GRO, RIP), Pol II (ChIP, GRO), and ChIP-seq datasets for Pol IV, AGO4, SPT6L, RDM15, HDA6, LDL1, and DDM1. (d) R-loop metaprofiles and heatmaps (ssDRIP-seq) showing signal intensity. All profiles (RPGC-normalized) represent ±3 kb regions flanking TSS and TES of antisense (a) / intergenic (i) lncRNA, and protein-coding (pc) loci.

We analyzed 30 histone modifications to evaluate chromatin states across transcript types (Fig. 3B). Protein-coding genes showed a highly ordered chromatin architecture, with strong positive correlations among most marks and consistently higher levels, indicative of a stable, actively transcribed chromatin environment. Intergenic lncRNAs displayed moderate coordination among activating acetylation marks, reflecting partial association with active chromatin. Antisense lncRNAs showed the weakest and most variable correlations, including negative associations involving H2Bub, H3K4me3, and related marks, consistent with constrained chromatin arising from overlapping transcription. These distinctions were supported by ChIP-seq signal intensities: protein-coding genes had the highest signals of active histone marks, whereas intergenic and antisense lncRNAs showed lower and more variable levels (Supplementary Fig. S1).

Together, these analyses show clear epigenetic distinctions among lncRNA classes, defined by their specific DNA methylation and histone modification profiles. The modest positive correlations between activating and repressive marks likely reflect mixed chromatin states across cell types or regulatory regions, indicating that intergenic lncRNAs occupy dynamically regulated chromatin environments.

Pol V transcribes a restricted lncRNA repertoire for RdDM

Pol V and RdDM-associated factors displayed transcript type–specific profiles (Fig. 3C). Pol V ChIP showed its highest occupancy on intergenic lncRNAs, with lower signal at antisense lncRNAs and minimal enrichment on protein-coding genes (Fig. 3C, top left). Pol V RIP followed this pattern, whereas Pol V GRO-seq showed a strong signal over protein-coding genes (Fig. 3C, top middle–right), reflecting assay-specific background rather than Pol V activity. AGO4 ChIP closely matched the Pol V ChIP distribution, with incremental enrichment at intergenic lncRNAs (Fig. 3C, middle right). In contrast, Pol II ChIP and Pol II GRO-seq were restricted to protein-coding genes, with negligible occupancy at lncRNA loci (Fig. 3C, second row). The Pol II elongation factor SPT6L showed the same bias toward protein-coding genes (Fig. 3C, middle left). Additional RdDM components (Pol IV, RDM15, and DDM1) were also most co-enriched at intergenic lncRNAs (Fig. 3C, middle row). Repressive chromatin modifiers HDA6 and LDL1

accumulated at these loci as well (Fig. 3C, bottom left–center), consistent with their elevated DNA methylation and transcriptional repression. Together, these data indicate intergenic lncRNAs as the primary targets of Pol V-dependent RdDM activity and associated chromatin remodeling.

Arabidopsis lncRNAs exhibited a wide spectrum of polyadenylation states, with the dominant fraction (62.7%) displaying weak or undetectable polyadenylation (Supplementary Fig. S2). Poly(A)+ and non-poly(A) lncRNAs accumulated to closely comparable steady-state levels, as reflected by their strong concordance in expression ($r = 0.94$; Fig. S2A). Disruption of Pol V exerted minimal influence on global lncRNA abundance: only 6 of 1,098 non-poly(A) lncRNAs were significantly reduced in *nrpe1* (padj < 0.05; $\log_2FC < -1$) (Supplementary Table S3), indicating that Pol V directly contributes to the expression of only a small subset of loci. To delineate this subset within established Pol V transcriptional domains, we intersected our lncRNA catalog with genomic intervals defined as Pol V–transcribed by IPARE [63]. This analysis identified 215 lncRNAs residing within Pol V–occupied regions, the majority of which were intergenic (86.5%) and lacked detectable 3′ polyadenylation (70.7%), consistent with defining molecular features of Pol V-associated transcripts [64]. Collectively, these data indicate that while non-polyadenylated lncRNAs constitute a substantial component of the *Arabidopsis* transcriptome, only a discrete subset aligns with canonical Pol V transcriptional activity.

Altogether, these results define a Pol II–independent RdDM hierarchy in which Pol V and its cofactors selectively operate at lncRNA loci, although only a small subset of lncRNAs are bona fide Pol V transcripts.

R-loop landscapes reflect divergent lncRNA transcriptional dynamics

R-loop profiling resolved distinct signal distributions across lncRNA subclasses (Fig. 3D). Antisense lncRNAs showed moderate R-loop signals at TSS and lower levels throughout gene bodies, whereas intergenic lncRNAs showed elevated signals at TSS and a pronounced peak at TES. Protein-coding genes exhibited the characteristic promoter-centered R-loop peak accompanied by a secondary TES-associated signal. Only limited loci in any category reached high signal intensity, indicating that strong R-loops arise at restricted genomic positions. These profiles delineate subclass-specific transcriptional dynamics among lncRNAs and differentiate them from the canonical R-loop architecture of protein-coding genes.

lncRNAs engage broadly with sRNAs

Despite the recognized significance of the interplay between lncRNAs and miRNAs, a considerable portion of their interactions remains to be elucidated. To address this, we predicted lncRNA targets for 427 published miRBase miRNAs. We first identified miRNA targets within lncRNAs and subsequently examined shared miRNA targets with mRNAs, resulting in the identification of 2,560 lncRNA–miRNA interactions and 2,202 miRNA–mRNA interactions (Supplementary Table S6). To explore the connection between lncRNAs, miRNAs, and mRNAs, we functionally annotated the target gene sequences into 30 of 50 major functional categories, primarily encompassing genes related to enzymes, solute transport, polyamide metabolism, and RNA biosynthesis (Fig. 4A). We then performed enrichment analysis on genes targeted by lncRNA-associated miRNAs to identify significantly enriched pathways across multiple public databases (Supplementary Table S7). The analysis revealed *miR822*, which regulates megaspore degeneration and female gametogenesis in flowering [65], as one of the most significantly enriched target genes. Other represented miRNA–targeted gene pathways included *miR165*, *miR5661*, *miR161*, *miR846*, and *miR842*.

Fig. 4. Identification and annotation of sRNAs associated with lncRNAs**.** (**a**) Enrichment analysis of miRNA target genes associated with lncRNAs, showing significant overrepresentation (Fisher's exact test, $p < 0.05$) across the top 20 MapMan bin functional categories. MapMan bins represent major biological processes. (**b**) Subnetwork of the miR859 pathway illustrating interactions between ath-miR859 and its predicted targets. Solid edges indicate interactions supported by degradome evidence. Protein-coding genes are shown as green ellipses, antisense lncRNAs as blue diamonds, and intergenic lncRNAs as grey diamonds. (**c**) Composition of sRNA loci overlapping lncRNAs. (**d**) Number of intergenic ("i") and antisense ("a") lncRNAs overlapping distinct sRNA locus types (siRNA, miRNA, and other sRNAs). (**e**) Intersection analysis of RdDM pathway components. Horizontal bars (left) indicate total set sizes for individual components: Pol IV/Pol V co-dependent regions and Pol IV-independent Pol V regions (from [43]), Pol V transcripts (from [38]), lncRNA-overlapped sRNAs, and AGO4-bound siRNAs. Vertical bars represent

intersection sizes, with connected dots indicating the specific components present in each set combination.

To further investigate the functional roles of lncRNAs, we used publicly available degradome datasets to explore their potential as competing endogenous RNAs (ceRNAs). As a case study, we examined miR859-targeted lncRNAs and built an interaction network to explore their roles in sRNA pathways via ceRNA mechanisms. Our analysis revealed interactions between miR859 and five lncRNAs (two antisense and three intergenic), as well as five mRNA targets encoding F-box proteins (Fig. 4B). Subsequent validation with degradome datasets confirmed cleaved targets, including AT3G16880.1 and MNTRG.39971.1 in leaf samples, and AT5G36200.1 and AT3G16820.1 in flower samples (Supplementary Table S8).

We also examined the overlap between sRNA precursors from the PSRG database and our identified lncRNAs (Supplementary Table S8). Within our lncRNA collection, we identified 23 miRNA precursors that majorly produce 21-nt miRNAs; all of these overlapped with intergenic lncRNAs (Fig. 4D). These included precursors from 13 miRNA families, with multiple representatives from *miR166*, *miR172*, *miR157*, and *miR319* families. In contrast, our lncRNA collection showed a predominant overlap with 24-nt siRNAs (n = 3,058), indicating a strong association with siRNA production. This overlap was primarily with the antisense class of lncRNAs (Fig. 4C and 4D)

To investigate the associations among RdDM pathway components, we performed an analysis incorporating Pol IV/Pol V-co-dependent regions and Pol IV-independent Pol V regions [42], Pol V transcripts [43], lncRNA-associated sRNAs, and AGO4-bound siRNAs (Fig. 4E) using data from previous studies. In our catalog, the largest association (9,627 regions) involved Pol IV/Pol V co-dependent regions together with AGO4-bound siRNAs, lncRNA-associated sRNAs, and Pol V transcripts, emphasizing their core role in RdDM. The second largest group (7,206 regions) consisted of Pol IV-independent Pol V regions associated with the same components, indicating a substantial number of loci where Pol V functions independently of Pol IV. Furthermore, the association between AGO4-bound siRNAs and lncRNA-associated sRNAs across 4,590 regions reinforces the role of lncRNAs in siRNA recruitment and transcriptional

silencing. This analysis suggests the extensive interplay between Pol IV, Pol V, and small RNAs, further reinforcing the fundamental role of lncRNAs in guiding RdDM.

In summary, the analyses demonstrate that lncRNAs maintain extensive associations with sRNAs, functioning both as miRNA targets and as sources of siRNAs that contribute centrally to RdDM pathway.

lncRNA-chromatin interactions point to cis regulation and stress responsiveness

Chromatin-associated regulation often involves lncRNAs that modulate gene transcription and genome organization by recruiting chromatin modifiers to specific loci. Therefore, to investigate whether lncRNAs contribute to triplex formation, we performed a computational analysis scanning the *Arabidopsis* genome. This analysis identified 286 potential triplex-forming sites where lncRNAs could serve as third-strand RNA binding partners in dsDNA:RNA triplex structures. These sites exhibited three major triplex motifs: purine (R), pyrimidine (Y), and mixed purine–pyrimidine (M). A total of 937 antisense and 743 intergenic lncRNAs were identified as having the potential to engage in such interactions, together accounting for the entire lncRNA population examined (Fig. 5A). We identified specific RNA–chromatin interactions mediated by lncRNAs, categorized as cis (≤1 kb), intra-chromosomal (>1 kb within the same chromosome), and inter-chromosomal (between different chromosomes). Among the 484 RNA–chromatin interactions mediated by 351 lncRNAs (including 355 antisense and 4 intergenic), the majority (65.7%) were cis, followed by 23.1% intra-chromosomal and 11.2% inter-chromosomal (Fig. 5B). These findings suggest that antisense lncRNAs predominantly mediate cis interactions. Moreover, lncRNAs interacted not only with their corresponding protein-coding genes but also with other lncRNA loci. However, none of the lncRNAs predicted to form PTS (potential triplex-forming sequences) were supported by interaction data from [47].

Fig. 5. Summary of lncRNA-chromatin interactions. (A) Number of potential PTS identified involving antisense and intergenic lncRNAs. Stacks represent triplex-forming motif types: Pyrimidine (Y), where CT-rich lncRNA sequences bind parallel to purine DNA; Purine (R), where

AG-rich sequences bind antiparallel to the DNA strand; and Mixed (M), which involves variable orientations of guanine- and thymine-rich sequences relative to purine-rich DNA. (C) Heatmap illustrates alterations in RNA-chromatin interactions mediated by lncRNAs in response to *Pst* infection and heat treatment. (D) Number of TF motif–associated peaks within ACRs overlapping lncRNA–DNA interaction loci. The annotated ACRs and motifs are from seedling ATAC-seq data.

We also investigated whether RNA–chromatin interactions respond to abiotic stresses, including heat stress (seedlings exposed to heat) and biotic stress (seedlings treated with *Pseudomonas syringae* pv. tomato (Pst)) GRID-seq samples. We observed alterations in lncRNA–chromatin interactions, with 31 interactions up-regulated and 33 interactions down-regulated in response to heat stress and 34 upregulated and 82 downregulated in response to Pst DC3000 (Fig. 5C), primarily mediated through antisense lncRNAs. Furthermore, we identified 13 protein-coding genes involved in both heat and Pst infection responses, encompassing various functional classes such as enzymes in sulfur metabolism (APS4, ATPAE5), regulatory proteins including a transcription factor (KTF1), chaperones (BIP3), and signaling proteins (AMK2, CIB2), indicating that lncRNAs mediate interactions across diverse gene regulatory networks involved in stress responses (Supplementary Table S9). In contrast, only a limited number of intergenic lncRNAs (n = 4) interacted with DNA and abiotic stress, implying a potential role like other DNA-interacting transcriptional regulators such as APOLO. To further investigate the potential transcriptional regulatory role of these lncRNA–chromatin interactions, we integrated our dataset with seedling ATAC-seq data to identify TFBMs overlapping lncRNA-associated ACRs (Fig. 5D). A distinct set of recurrent TFBMs was identified, primarily involving DOF, ERF, MYB, NAC, WRKY, bHLH, and bZIP families. Among these, ERF and DOF motifs were the most prevalent, suggesting their potential roles in mediating the observed chromatin interactions.

Together, the results demonstrate that antisense lncRNAs primarily function in cis, exhibiting dynamic responses to environmental stress and modulating transcriptional networks through coordinated interactions with transcription factor binding.

GWAS–QTL colocalization identifies trait-associated lncRNAs

To investigate the contribution of lncRNA genes to phenotypic traits, we analyzed colocalization between GWAS signals, QTLs, and lncRNA loci across *A. thaliana*. A colocalization event was defined as a robust association between a trait feature and a GWAS/QTL locus, resulting in 4,157 lncRNA–trait colocalization events, including 28 distinct traits (Supplementary Table S10). Among all annotated lncRNAs, 22.3% (674 intergenic and 391 antisense) were linked to GWAS-QTLs associated with defined traits ($P < 0.05$, chi-squared test; Supplementary Fig. S3). Traits related to arsenic concentration, flowering time, leaf discoloration, metabolite content, relative root length, and root morphology showed significant associations with lncRNA loci. Flowering time was linked exclusively to antisense lncRNAs, whereas leaf trichome density was associated primarily with intergenic lncRNAs. Several other traits exhibited similar levels of lncRNA–QTL overlap, indicating that lncRNAs participate in diverse trait-related genomic interactions across the *Arabidopsis* genome.

Evidence for small peptides in lncRNA producing loci

Numerous studies reported the presence of stable, functional peptides encoded from lncRNA producing loci [63–65]. To systematically characterize sORFs and their encoded small peptides, we developed a pipeline to analyze both lncRNA and mRNA transcripts as outlined in Fig. 6A. Although guided by features typical of protein-coding genes, this method systematically identified a large collection of putative sORFs across various transcript types. Most candidate peptides derived from protein-coding genes, with far fewer from antisense or intergenic lncRNAs, indicating disparities in translational potential among RNA classes (Fig. 6A). Given the potential for this method to miss sORF-containing transcripts not resembling protein-coding genes, we validated the authenticity of putative small peptides by searching for sequences 5–100 amino acids in length with significant similarity to experimentally annotated peptides in *Arabidopsis* Peptide Atlas (BLASTp, E-value < 0.00001; percent identity ≥ 90%).

Fig. 6. Characterization of lncRNA-encoded peptides. (**a**) Outline of the sORF and small peptide prediction and validation workflow. (**b**) Distribution of peptide lengths encoded by lncRNAs (antisense, ‘a’; intergenic, ‘i’) compared with protein-coding transcripts (‘pc’). (**c**) Venn diagram showing the overlap of validated peptides derived from protein-coding transcripts and lncRNAs.

This analysis confirmed 4,203 validated peptides, with a median length of 18–24 amino acids (Fig. 6B, Supplementary Table S11), establishing their identity as bona fide small peptides. Most of these validated peptides (4,142) originated from protein-coding genes, with smaller subsets derived from antisense lncRNAs (414 peptides) and intergenic lncRNAs (36 peptides) (Fig. 6C). This distribution reveals predominance of peptide-encoding potential in antisense lncRNAs relative to intergenic lncRNAs. Furthermore, 61 peptides were identified as shared between protein-coding genes and antisense lncRNAs, while intergenic lncRNAs showed no such overlap with other transcript types. These findings suggest that the shared peptides may have dual functions in both coding and non-coding regulatory contexts. Overall, the results further support the translational capacity of lncRNAs, particularly antisense lncRNAs, emphasizing their contribution as a source of functional small peptides.

Intergenic lncRNAs show positional conservation and structural retention despite sequence divergence in Brassicaceae

To explore the evolutionary relationships of intergenic lncRNAs identified in *A. thaliana* in this study with those cataloged in the PLncDB database [22] for *Eutrema salsugineum*, *Capsella rubella*, *Arabidopsis lyrata*, and *Brassica rapa*, we conducted a comparative analysis, excluding antisense lncRNAs due to their partial overlap with protein-coding genes. Building on methodologies described by [53], we employed three strategies to understand potential interrelationships: sequence homology, secondary structure similarity, and synteny analysis (Fig. 7).

Fig. 7. Assessment of evolutionary relationships among intergenic lncRNAs across Brassicaceae species. (A) Venn diagrams depict the number of classified lncRNA families (left) and their members (right) across species based on BLAST, secondary structure, and synteny classifications. (B) A consensus circos plot illustrates pairwise syntenic relationships of lncRNAs among the species.

First, we identified one-to-one best reciprocal BLAST hits between each pair of species and applied a clustering approach to integrate pairwise comparisons, defining lncRNA families across species. As anticipated, this analysis revealed a limited number of conserved lncRNA families (n = 403), encompassing 2,233 lncRNAs (Fig. 7A, Supplementary Table S12), consistent with the well-documented pattern of low sequence conservation and substantial species divergence among lncRNAs. Second, we assessed structural similarities using the Beagle software, which performs pairwise alignments of secondary structures. This approach classified 3,135 lncRNAs into 756 structural families (Fig. 7A, Supplementary Table S12). These findings align with recent observations in Brassicaceae, suggesting the presence of conserved structural motifs with potential biological relevance [66]. However, the observed structural similarity may reflect either heightened evolutionary constraints or limitations in the specificity of the Beagle alignment method, necessitating caution in interpreting these results, as secondary structure alignments lack sufficient specificity for robust evolutionary inferences [59]. Third, we performed a synteny analysis using a validated methodology from [53] to identify evolutionarily related lncRNAs. This approach uncovered a substantially larger number of syntenic lncRNAs (n = 7,149) compared to the sequence homology and structural analyses, although fewer lncRNAs were shared across all five Brassicaceae species than in the structural analysis (Fig. 7A–B, Supplementary Table S12). Nevertheless, 66 families were common to all five species, which was the highest number among the three methods and represented 16.8% of annotated *A. thaliana* lncRNAs. This method, overall, reveals clear syntenic relationships among Brassicaceae species, particularly among those with closer evolutionary ties (Fig. 7B).

To further investigate potential functional or evolutionary relatedness, we examined syntenic families for conserved sequence motifs shared among transcripts within a family. Motif discovery using MEME identified 250 families with conserved motifs, while an independent graph-based approach implemented in lncLOOM detected 329 families with shared motifs. Comparative analysis of these results revealed 204 lncRNA families with conserved sequence motifs identified by both tools (Supplementary Table S13s). This observation is consistent with prior reports documenting short conserved sequences within lncRNAs [66].

Collectively, these findings indicate that most intergenic lncRNAs are species-specific, exhibiting considerable evolutionary divergence across Brassicaceae species, yet retaining a limited capacity for sequence conservation, as evidenced by the presence of short, conserved motifs within syntenic lncRNAs.

Discussion

Over the past decade, lncRNAs have emerged as critical regulators of gene expression, genome organization, and chromatin stability in plants and other eukaryotes. Our integrated analysis of over 700 stranded RNA-seq datasets renders a comprehensive *A. thaliana* lncRNA catalog (including antisense and intergenic lncRNAs) with a broad range of annotations, shedding new light on their global properties and testing or generalizing prior hypotheses. Nevertheless, our collection of lncRNAs is by no means exhaustive, as typical RNA-seq inevitably misses many actively transcribed but rapidly degraded transcripts. Future studies employing nascent RNA sequencing approaches (e.g., GRO-seq, plaNET-seq) or nuclear RNA degradation mutants (e.g., *Arabidopsis* exosome mutants) will be essential to achieve a more complete representation of the lncRNA transcriptome. In this study, we broadened the lncRNA repertoire in the *A. thaliana* catalog, not only complements previous genome-wide annotation efforts [10, 24, 25, 67] but also uncovers numerous genomic regions previously unidentified as lncRNA-producing loci. Another important observation is that lncRNAs identified in our study, as well as in others, show limited overlap with existing databases and previous catalogs. This points to the need for strategically consolidating these diverse data into a unified repository, like the approach taken for human lncRNA collections over the years. Such integration and additional experimental evidence would unify fragmented annotations, ensuring a comprehensive lncRNA repository.

Consistent with earlier findings, we observe features (such as reduced length, decreased GC content, and increased sequence variability) that are indicative of relaxed selective pressure and rapid evolutionary divergence [69, 71–74]. These characteristics correlate with tissue-specific expression, emphasizing the role of lncRNAs in precise regulatory mechanisms. For instance, the depletion of intergenic lncRNAs in fruit likely reflects maturation-linked transcriptional reprogramming, where chromatin compaction and hormone-driven pathways transiently suppress their activation. It would be interesting to study these lncRNAs as these might be involved in regulatory shifts during fruit maturation, potentially modulating mechanisms tied to cell expansion or hormone signaling (e.g., auxin, ethylene) [75]. This specificity, also seen across plant species [76], reinforces the context-dependent role of lncRNAs, though low expression levels may inflate perceived specificity. Still, differentiating genuine regulatory specificity from effects caused by low transcript detectability remains a significant challenge.

Unlike protein-coding genes, lncRNAs contain repetitive sequences, particularly in intergenic regions for regulatory functions [77, 78]. A substantial fraction of lncRNA loci contain TE-derived fragments, with intergenic loci exhibiting a mild yet broader diversity of TEs compared to antisense or coding loci. The selective retention of TEs in noncoding regions acts as an evolutionary strategy, enabling these regions to acquire regulatory functions while preserving transcriptional stability, supporting evidence that TEs influence lncRNA biogenesis, stability, and function [79]. These TE-derived patches, often short motifs that do not fully match annotated TEs, are especially prevalent in intergenic lncRNAs, suggesting they are not trivial remnants but active players in genome dynamics. Such sequences appear to guide epigenetic modifications, repressing gene expression or compacting chromatin, a role amplified by their TE-like signatures that may recruit silencing machinery, as supported by [25] and consistent with frequent TE origins in intergenic regions [80, 81]. The relationship between lncRNAs and TEs complicates efforts to distinguish them, yet it suggests a compelling possibility: embedded TE elements may serve as regulatory switches that link lncRNA activity to the epigenetic landscape, influencing expression variability and genome organization.

Our findings reveal distinct epigenetic landscapes of lncRNAs, indicating their specialized regulatory roles compared to protein-coding genes. In intergenic lncRNAs, elevated CHG and CHH methylation likely enforces repressive chromatin, whereas antisense lncRNAs maintain CG and non-CG methylation reminiscent of protein-coding genes, suggesting these loci retain a more accessible chromatin state conducive to transcription despite overlapping units. The relationship between DNA methylation and expression is generally negative, particularly in gene bodies [24], and our data also suggest that high methylation levels correlate with low lncRNA expression. Although our data provides hints to methylation-mediated regulation, additional research that combines promoter architecture analysis, as shown by [82], methylation and expression studies will be essential to establish more definitive conclusions regarding lncRNA regulation. Besides this, additional evidence from our study shows that lncRNAs carry both activating and repressive histone marks, and class-specific correlations position them along a continuum from permissive to constrained chromatin states, indicating a direct role in establishing or maintaining heterochromatin architecture.

Our study builds on previous findings by demonstrating how lncRNAs actively participate in shaping the RdDM pathway, beyond merely being passive transcriptional byproducts. We find that RdDM depends on Pol V, which our data reveal targets lncRNA loci, occupying increasingly at intergenic regions. The recruitment of Pol V, AGO4, and Pol V-associated sRNAs at intergenic

lncRNAs suggests a sequential mechanism, where Pol V transcripts facilitate AGO4 binding and siRNA-guided DNA methylation [83, 84]. The co-enrichment of chromatin remodelers such as HDA6 and LDL1 further implicates chromatin modifications in RdDM regulation [43], although we found no convincing evidence for a connection with SPT6L. Previous research [64] reported that Pol V transcripts are generally non-polyadenylated; however, we observed that only around 4% of non-polyadenylated transcripts are directly transcribed by Pol V. Nonetheless, this finding aligns with Pol V's role in generating lncRNAs that guide the methylation machinery [64, 84].

Our data also paint a clear picture of R-loops tied to lncRNAs, revealing transcriptional patterns that differ from protein-coding genes and vary between lncRNA types. Intergenic lncRNAs form prominent R-loops at TES that may facilitate Pol V engagement and chromatin remodeling to support RdDM, with antisense lncRNAs displaying weaker, diffuse R-loops likely due to transcriptional interference from overlapping coding sequences. These differences cast R-loops as active players shaped by genomic context and align with evidence that R-loop-associated lncRNAs influence epigenetic regulation in plants [85]. Given that R-loop structures have been proposed to recruit or stabilize RdDM components, this invites further investigation into how R-loops may serve as intermediates linking lncRNA transcription to locus-specific DNA methylation.

Certain lncRNAs act as ceRNAs, interacting with miRNAs like miR859 to regulate F-box mRNA levels, as validated by degradome sequencing. Specific lncRNA genes seem to harbor sRNAs that are biologically processed into shorter functional RNAs such as microRNAs (such as miR166 and miR172 via DCL1 cleavage) for post-transcriptional silencing in *Arabidopsis* [86–88]. Beyond miRNAs, other sRNAs emerge from distinct biogenesis routes, including tasiRNAs and phasiRNAs from miRNA-triggered lncRNAs, both shaping regulatory outcomes [89, 90]. In addition to this, the observation of widespread co-localization of lncRNA-associated sRNAs, AGO4-bound siRNAs, and Pol V transcripts, reinforcing the central role of lncRNAs in RdDM [91]. The identification of Pol IV-independent Pol V regions further suggests that some lncRNAs facilitate RdDM without Pol IV-derived siRNAs [92]. Altogether, these collectively support a refined view of RdDM, where lncRNAs act as central regulatory nodes integrating transcriptional, siRNA-mediated, and chromatin-based silencing mechanism.

To manage environmental stress, plants have developed specialized gene regulatory mechanisms, including the activity of antisense lncRNAs. Our findings indicate that antisense lncRNAs dominantly drive cis-chromatin interactions, possibly aligning with local regulatory roles [93]. These interactions shift under stress conditions, inducing stress-responsive functions. The

presence of motifs (such as ERF and DOF) within lncRNA loci suggests that stress-induced transcription factors may be recruited through antisense transcriptional activity [94]. Such evidence also supports the emerging view that antisense transcription acts as an activating regulatory layer in stress responses [95, 96]. Nevertheless, predicted triplexes hint at additional chromatin interactions, but their functional relevance awaits validation.

Determining the functional roles of lncRNA genes in phenotypes remains a major challenge, with few cases directly linked to observable effects. The observed lncRNA–trait associations based on colocalization of GWAS and QTL signals indicate that most traits are majorly influenced by intergenic or antisense lncRNAs, with some traits affected by both. This imply that lncRNAs may influence phenotypic variation either independently or through coordinated regulation with their cognate genes, as seen in examples such as *asCOOLAIR–FLC* (affects flowering time) [97], and *asDOG1–DOG1* (controls seed dormancy) [98].

Emerging evidence indicates that lncRNA loci can produce small peptides, and identification of lncRNA-derived peptides from this catalog supports their role as a functional reservoir, consistent with translational activity reported in other plant species [99, 100]. LncRNAs regulate epigenetic and stress responses, and the presence of peptide-coding potential, especially in antisense lncRNAs, points to dual functionality in regulation and translation, broadening their evolutionary scope.

This study further compiles an extensive catalog of conserved intergenic lncRNA families across Brassicaceae, establishing a resource aimed to advance investigations into the lncRNA evolution. The rapid sequence turnover of lncRNAs has long confounded comparative analyses; however, our findings reiterate that the synteny-based method provides the most reliable framework for identifying conserved lncRNAs, outperforming sequence or structural similarity. Despite the divergence among Brassicaceae genomes, syntenic lncRNAs retain short, conserved motifs, affirming genomic position as the key determinant of evolutionary retention [101]. This supports the “RNA modular code” hypothesis, in which conserved sequence modules form specific structural or interaction domains, while surrounding regions evolve rapidly [102, 103]. Such a modular organization may enable functional persistence even as primary sequences diverge.

The evolutionary lability of lncRNAs raises a central question: how is functional conservation maintained amid rapid sequence turnover? This turnover appears largely driven by TE insertions, gene duplications, and pseudogenization events that generate new transcripts while eroding detectable sequence homology [104, 105]. Yet, synteny comparisons indicate that

transcriptional activity at conserved loci is often preserved, and the retained motifs within these lncRNAs may safeguard essential molecular interfaces. Known examples such as *IPS1* [106]**,** *ENOD40* [108] and *COOLAIR* [109], *HID1* [107] demonstrate that functional conservation can persist through shared molecular mechanisms rather than sequence similarity. Thus, lncRNA conservation reflects the maintenance of functionally critical modules within evolving genomic contexts, reframing the central question from whether these conserved lncRNAs are functional to how their specific roles are preserved across evolution.

Structural analyses further suggest evolutionary constraints on RNA folding, although the limitations of secondary structure alignments temper the strength of these evolutionary inferences. Nevertheless, the presence of conserved structural features in certain lncRNAs [108–110] suggests that RNA folding may underpin stable, functionally critical interactions. The model of intergenic lncRNA evolution we propose, alongside our catalog of conserved lncRNAs, lays the groundwork for future efforts to validate and elucidate their functional contributions.

The functional significance and the precise number of genuine lncRNAs in plants remain debated. We propose expanding lncRNA annotation efforts in *A. thaliana* and other plant species by leveraging comprehensive RNA-seq datasets, multi-omic profiles, native-state sequencing, RNA polymerase IV/V-dependent transcription studies, and long-read sequencing technologies to better understand lncRNA functions. As sequencing efforts increasingly target lncRNAs, our work provides a foundation for refining their annotations, expression patterns, and functional roles, thereby elucidating their contributions to plant genome regulation.

Conclusion

This investigation presents a comprehensive atlas of lncRNAs in *A. thaliana*, assembled by integrating strand-specific RNA-seq data with diverse multi-omics resources. Extending beyond mere enumeration, our approach aimed at detailed functional annotations that illuminate the regulatory roles of these transcripts. We report lncRNA associations with TEs, DNA methylation, and histone modifications, alongside ties to features such as Pol II/IV/V occupancy, R-loops, and RdDM pathway involvement. We uncovered lncRNAs encoding small peptides and others linked to phenotypic traits, reinstating the regulatory potency of these transcripts. A part of conserved lncRNAs, marked by shared sequence motifs, spanned Brassicaceae species, implying selective pressures on their roles. Overall, we envision this resource as a reference for dissecting the intricate contributions of lncRNAs to plant development and adaptation, guiding broader lncRNA research.

Abbreviations

ACR Accessible Chromatin Region
AGO4 Argonaute 4
ATAC seq – Assay for Transposase-Accessible Chromatin using Sequencing
BHLH Basic Helix-Loop-Helix
BLAST Basic Local Alignment Search Tool
BWA Burrows–Wheeler Aligner
BZIP Basic Leucine Zipper
ceRNA Competing Endogenous RNA
ChIP-seq Chromatin Immunoprecipitation Sequencing
eQTL Expression Quantitative Trait Locus
ERF Ethylene Response Factor
FPKM Fragments per Kilobase of transcript per Million mapped reads
GRID-seq Global RNA Interactions with DNA by Sequencing
GRO-seq Global Run-On Sequencing
GWAS Genome-Wide Association Study
H2Bub Histone H2B Monoubiquitination
H3K4me3 Histone H3 Lysine 4 Trimethylation
H3K27ac Histone H3 Lysine 27 Acetylation
H3K36me3 Histone H3 Lysine 36 Trimethylation
HDA6 Histone Deacetylase 6
LINE Long Interspersed Nuclear Element
lncRNA Long Non-Coding RNA
LTR Long Terminal Repeat
miRNA MicroRNA
mRNA Messenger RNA

NRPE1 Nuclear RNA Polymerase E1 (Pol V Subunit)
ONT Oxford Nanopore Technologies
ORF Open Reading Frame
PCC Pearson Correlation Coefficient
phasiRNA Phased Secondary Small Interfering RNA
Pol II Polymerase II
Pol IV Polymerase IV
Pol V Polymerase V
PTS Potential Triplex-Forming Sequence
QTL Quantitative Trait Locus
RdDM RNA-Directed DNA Methylation
RIP-seq RNA Immunoprecipitation Sequencing
R-loop RNA:DNA Hybrid Structure
ssDRIP-seq Single-Strand DNA Ligation from DNA:RNA Hybrid Immunoprecipitation Sequencing
SINE Short Interspersed Nuclear Element
SNP Single Nucleotide Polymorphism
sORF Small Open Reading Frame
SRA Sequence Read Archive
tasiRNA Trans-acting Small Interfering RNA
TE Transposable Element
TFBM Transcription Factor Binding Motif
TPM Transcripts Per Million
TSS Transcription Start Site
TES Transcription End Site
WGBS Whole-Genome Bisulfite Sequencing

Acknowledgements

AT Vivek acknowledges the Department of Biotechnology, Government of India, and the National Institute of Plant Genome Research (NIPGR) for providing the research fellowship. The authors are grateful to the DBT e-Library Consortium (DeLCON) for providing access to e-material. The

authors express their gratitude to Dr. Hrant Hovhannisyan for assisting in plotting variation data during the analysis and for useful discussions. The authors also acknowledge Dr. Prabhakaran Soundararajan, Scientist, BRIC-NIPGR, for helpful discussions regarding synteny analysis.

Authorship contribution

Vivek AT: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Writing - Original Draft, Visualization; **Kiran H**: Formal analysis; **Namrata Sahu:** Formal analysis; **Garima Kalakoti:** Visualization; **Supriya Swain**: Data Curation; **Shailesh Kumar:** reviewed & edited the manuscript, conceived, and coordinated the project and provided overall supervision.

Funding

This research is supported by the BT/PR40146/BTIS/137/4/2020 project grant from the Department of Biotechnology (DBT), Government of India, and by the core grant of the National Institute of Plant Genome Research (NIPGR) in the laboratory of SK.

Availability of data and materials

All source codes and processed data used in this study are publicly available at https://github.com/skbinfo/lncFETCHER.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent of Publication

Not Applicable

Competing interests

The authors have declared no competing interests.

References

1. Chekanova JA. Long non-coding RNAs and their functions in plants. Curr Opin Plant Biol. 2015;27:207–16. https://doi.org/10.1016/j.pbi.2015.08.003.

2. Yu Y, Zhang Y, Chen X, Chen Y. Plant Noncoding RNAs: Hidden Players in Development and Stress Responses. Annual Review of Cell and Developmental Biology. 2019;35 Volume 35, 2019:407–31. https://doi.org/10.1146/annurev-cellbio-100818-125218.

3. Lucero L, Ferrero L, Fonouni-Farde C, Ariel F. Functional classification of plant long noncoding RNAs: a transcript is known by the company it keeps. New Phytol. 2021;229:1251–60. https://doi.org/10.1111/nph.16903.

4. Zhang Y-C, He R-Q, Cheng Y, Wang D, Ariel F, Chen Y-Q. Long noncoding RNAs as molecular architects: Shaping plant functions and physiological plasticity. Mol Plant. 2025;18:1643–71. https://doi.org/10.1016/j.molp.2025.09.008.

5. Zhou B, Zhao H, Yu J, Guo C, Dou X, Song F, et al. EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. Nucleic Acids Res. 2018;46:D100–5. https://doi.org/10.1093/nar/gkx677.

6. Klapproth C, Sen R, Stadler PF, Findeiß S, Fallmann J. Common Features in lncRNA Annotation and Classification: A Survey. Non-Coding RNA. 2021;7:77. https://doi.org/10.3390/ncrna7040077.

7. Cao H, Wahlestedt C, Kapranov P. Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. Trends Genet. 2018;34:704–21. https://doi.org/10.1016/j.tig.2018.06.002.

8. Vivek AT, Kumar S. Computational methods for annotation of plant regulatory non-coding RNAs using RNA-seq. Brief Bioinform. 2021;22:bbaa322. https://doi.org/10.1093/bib/bbaa322.

9. Zheng H, Brennan K, Hernaez M, Gevaert O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. GigaScience. 2019;8:giz145. https://doi.org/10.1093/gigascience/giz145.

10. Zhu Y, Chen L, Hong X, Shi H, Li X. Revealing the novel complexity of plant long non-coding RNA by strand-specific and whole transcriptome sequencing for evolutionarily representative plant species. BMC Genomics. 2022;23:381. https://doi.org/10.1186/s12864-022-08602-9.

11. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. https://doi.org/10.1186/1471-2105-12-323.

12. 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. https://doi.org/10.1038/nmeth.4197.

13. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 2008;36 Database issue:D1009-1014. https://doi.org/10.1093/nar/gkm965.

14. 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. https://doi.org/10.1038/s41587-019-0201-4.

15. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5. https://doi.org/10.1038/nbt.3122.

16. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. https://doi.org/10.1038/nbt.1883.

17. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research. 2020;9:ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.2.

18. Sims D, Ilott NE, Sansom SN, Sudbery IM, Johnson JS, Fawcett KA, et al. CGAT: computational genomics analysis toolkit. Bioinformatics. 2014;30:1290–1. https://doi.org/10.1093/bioinformatics/btt756.

19. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017;45:e57–e57. https://doi.org/10.1093/nar/gkw1306.

20. Oróstica KY, Verdugo RA. chromPlot: visualization of genomic data in chromosomal context. Bioinformatics. 2016;32:2366–8. https://doi.org/10.1093/bioinformatics/btw137.

21. Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I. CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. Methods Mol Biol Clifton NJ. 2019;1933:415–29. https://doi.org/10.1007/978-1-4939-9045-0_26.

22. Jin J, Lu P, Xu Y, Li Z, Yu S, Liu J, et al. PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. Nucleic Acids Res. 2021;49:D1489–95. https://doi.org/10.1093/nar/gkaa910.

23. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J Cell Mol Biol. 2017;89:789–804. https://doi.org/10.1111/tpj.13415.

24. Palos K, Nelson Dittrich AC, Yu L, Brock JR, Railey CE, Wu H-YL, et al. Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. Plant Cell. 2022;34:3233–60. https://doi.org/10.1093/plcell/koac166.

25. Kornienko AE, Nizhynska V, Molla Morales A, Pisupati R, Nordborg M. Population-level annotation of lncRNAs in Arabidopsis reveals extensive expression variation associated with transposable element–like silencing. Plant Cell. 2024;36:85–111. https://doi.org/10.1093/plcell/koad233.

26. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.

27. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma Oxf Engl. 2014;30:923–30. https://doi.org/10.1093/bioinformatics/btt656.

28. Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schütz F, et al. Mechanisms and Evolutionary Patterns of Mammalian and Avian Dosage Compensation. PLOS Biol. 2012;10:e1001328. https://doi.org/10.1371/journal.pbio.1001328.

29. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14. https://doi.org/10.1093/bib/bbw008.

30. Schikora-Tamarit MÀ, Gabaldón T. PerSVade: personalized structural variant detection in any species of interest. Genome Biol. 2022;23:175. https://doi.org/10.1186/s13059-022-02737-4.

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.

32. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinforma Oxf Engl. 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560.

33. M. Vasimuddin, S. Misra, H. Li, S. Aluru. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019. p. 314–24. https://doi.org/10.1109/IPDPS.2019.00041.

34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.

35. Zhou Q, Lim J-Q, Sung W-K, Li G. An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. BMC Bioinformatics. 2019;20:47. https://doi.org/10.1186/s12859-018-2593-4.

36. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42:W187–91. https://doi.org/10.1093/nar/gku365.

37. Bonenfant Q, Noé L, Touzet H. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. Bioinforma Adv. 2023;3:vbac085. https://doi.org/10.1093/bioadv/vbac085.

38. Yuan Y, Liu Y, Han L, Li Y, Qi Y. An RdDM-independent function of Pol V transcripts in gene regulation and plant defence. Nat Plants. 2024;10:1562–75. https://doi.org/10.1038/s41477-024-01774-0.

39. Lunardon A, Johnson NR, Hagerott E, Phifer T, Polydore S, Coruh C, et al. Integrated annotations and analyses of small RNA-producing loci from 47 diverse plants. Genome Res. 2020;30:497–513. https://doi.org/10.1101/gr.256750.119.

40. Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. 2011;39 Web Server issue:W155-159. https://doi.org/10.1093/nar/gkr319.

41. LOHSE M, NAGEL A, HERTER T, MAY P, SCHRODA M, ZRENNER R, et al. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ. 2014;37:1250–8. https://doi.org/10.1111/pce.12231.

42. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36:2628–9. https://doi.org/10.1093/bioinformatics/btz931.

43. Liu Y, Shu J, Zhang Z, Ding N, Liu J, Liu J, et al. A conserved Pol II elongator SPT6L mediates Pol V transcription to regulate RNA-directed DNA methylation in Arabidopsis. Nat Commun. 2024;15:4460. https://doi.org/10.1038/s41467-024-48940-8.

44. Johnson NR, Yeoh JM, Coruh C, Axtell MJ. Improved Placement of Multi-mapping Small RNAs. G3 GenesGenomesGenetics. 2016;6:2103–11. https://doi.org/10.1534/g3.116.030452.

45. Amatria-Barral I, González-Domínguez J, Touriño J. PATO: genome-wide prediction of lncRNA–DNA triple helices. Bioinformatics. 2023;39:btad134. https://doi.org/10.1093/bioinformatics/btad134.

46. Li L, Luo H, Lim D-H, Han L, Li Y, Fu X-D, et al. Global profiling of RNA–chromatin interactions reveals co-regulatory gene expression networks in Arabidopsis. Nat Plants. 2021;7:1364–78. https://doi.org/10.1038/s41477-021-01004-x.

47. Ding K, Sun S, Luo Y, Long C, Zhai J, Zhai Y, et al. PlantCADB: A Comprehensive Plant Chromatin Accessibility Database. Genomics Proteomics Bioinformatics. 2023;21:311–23. https://doi.org/10.1016/j.gpb.2022.10.005.

48. Singh U, Wurtele ES. orfipy: a fast and flexible tool for extracting ORFs. Bioinformatics. 2021;37:3019–20. https://doi.org/10.1093/bioinformatics/btab090.

49. Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H, et al. PsORF: a database of small ORFs in plants. Plant Biotechnol J. 2020;18:2158–60. https://doi.org/10.1111/pbi.13389.

50. van Wijk KJ, Leppert T, Sun Q, Boguraev SS, Sun Z, Mendoza L, et al. The Arabidopsis PeptideAtlas: Harnessing worldwide proteomics data to create a comprehensive community proteomics resource. Plant Cell. 2021;33:3421–53. https://doi.org/10.1093/plcell/koab211.

51. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol AMB. 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.

52. Mattei E, Pietrosanto M, Ferrè F, Helmer-Citterich M. Web-Beagle: a web server for the alignment of RNA secondary structures. Nucleic Acids Res. 2015;43 Web Server issue:W493–7. https://doi.org/10.1093/nar/gkv489.

53. Pegueroles C, Iraola-Guzmán S, Chorostecki U, Ksiezopolska E, Saus E, Gabaldón T. Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis. RNA Biol. 2019;16:320–9. https://doi.org/10.1080/15476286.2019.1572438.

54. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238. https://doi.org/10.1186/s13059-019-1832-y.

55. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015;43:W39-49. https://doi.org/10.1093/nar/gkv416.

56. Ross CJ, Rom A, Spinrad A, Gelbard-Solodkin D, Degani N, Ulitsky I. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. Genome Biol. 2021;22:29. https://doi.org/10.1186/s13059-020-02247-1.

57. Kang C, Liu Z. An Easy-to-Follow Pipeline for Long Noncoding RNA Identification: A Case Study in Diploid Strawberry Fragaria vesca. Methods Mol Biol Clifton NJ. 2019;1933:223–43. https://doi.org/10.1007/978-1-4939-9045-0_13.

58. Jiang S, Cheng S-J, Ren L-C, Wang Q, Kang Y-J, Ding Y, et al. An expanded landscape of human long noncoding RNA. Nucleic Acids Res. 2019;47:7842–56. https://doi.org/10.1093/nar/gkz621.

59. Hovhannisyan H, Gabaldón T. The long non-coding RNA landscape of Candida yeast pathogens. Nat Commun. 2021;12:7317. https://doi.org/10.1038/s41467-021-27635-4.

60. Singh A, AT V, Gupta K, Sharma S, Kumar S. Long non-coding RNA and microRNA landscape of two major domesticated cotton species. Comput Struct Biotechnol J. 2023;21:3032–44. https://doi.org/10.1016/j.csbj.2023.05.011.

61. Pronozin AY, Afonnikov DA. ICAnnoLncRNA: A Snakemake Pipeline for a Long Non-Coding-RNA Search and Annotation in Transcriptomic Sequences. Genes. 2023;14:1331. https://doi.org/10.3390/genes14071331.

62. Lu Z, Xia X, Jiang B, Ma K, Zhu L, Wang L, et al. Identification and characterization of novel lncRNAs in Arabidopsis thaliana. Biochem Biophys Res Commun. 2017;488:348–54. https://doi.org/10.1016/j.bbrc.2017.05.051.

63. Tsuzuki M, Sethuraman S, Coke AN, Rothi MH, Boyle AP, Wierzbicki AT. Broad noncoding transcription suggests genome surveillance by RNA polymerase V. Proc Natl Acad Sci. 2020;117:30799–804. https://doi.org/10.1073/pnas.2014419117.

64. Wierzbicki AT, Haag JR, Pikaard CS. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell. 2008;135:635–48. https://doi.org/10.1016/j.cell.2008.09.035.

65. Tovar-Aguilar A, Grimanelli D, Acosta-García G, Vielle-Calzada JP, Badillo-Corona JA, Durán-Figueroa N. The miRNA822 loaded by ARGONAUTE9 modulates the monosporic female gametogenesis in Arabidopsis thaliana. Plant Reprod. 2023;1:1–16. https://doi.org/10.1007/S00497-023-00487-2/FIGURES/7.

66. Liu W, Duttke SH, Hetzel J, Groth M, Feng S, Gallego-Bartolome J, et al. RNA-directed DNA methylation involves co-transcriptional small-RNA-guided slicing of polymerase V transcripts in Arabidopsis. Nat Plants. 2018;4:181–8. https://doi.org/10.1038/s41477-017-0100-y.

67. Huang J-Z, Chen M, Chen D, Gao X-C, Zhu S, Huang H, et al. A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. Mol Cell. 2017;68:171-184.e6. https://doi.org/10.1016/j.molcel.2017.09.015.

68. Fesenko I, Shabalina SA, Mamaeva A, Knyazev A, Glushkevich A, Lyapina I, et al. A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants. Nucleic Acids Res. 2021;49:10328–46. https://doi.org/10.1093/nar/gkab816.

69. Chen L, Zhu Q-H. The evolutionary landscape and expression pattern of plant lincRNAs. RNA Biol. 2022;19:1190–207. https://doi.org/10.1080/15476286.2022.2144609.

70. Corona-Gomez JA, Coss-Navarrete EL, Garcia-Lopez IJ, Klapproth C, Pérez-Patiño JA, Fernandez-Valverde SL. Transcriptome-guided annotation and functional classification of long non-coding RNAs in Arabidopsis thaliana. Sci Rep. 2022;12:14063. https://doi.org/10.1038/s41598-022-18254-0.

71. Zheng X, Chen Y, Zhou Y, Shi K, Hu X, Li D, et al. Full-length annotation with multistrategy RNA-seq uncovers transcriptional regulation of lncRNAs in cotton. Plant Physiol. 2021;185:179–95. https://doi.org/10.1093/plphys/kiaa003.

72. Singh A, Vivek AT, Kumar S. AlnC: An extensive database of long non-coding RNAs in angiosperms. PLOS ONE. 2021;16:e0247215. https://doi.org/10.1371/journal.pone.0247215.

73. Villalba-Bermell P, Marquez-Molins J, Gomez G. A multispecies study reveals the diversity and potential regulatory role of long noncoding RNAs in cucurbits. Plant J. 2024;120:799–817. https://doi.org/10.1111/tpj.17013.

74. Yadav VK, Jalmi SK, Tiwari S, Kerkar S. Deciphering shared attributes of plant long non-coding RNAs through a comparative computational approach. Sci Rep 2023 131. 2023;13:1–13. https://doi.org/10.1038/s41598-023-42420-7.

75. Wang Y, Deng XW, Zhu D. From molecular basics to agronomic benefits: Insights into noncoding RNA-mediated gene regulation in plants. J Integr Plant Biol. 2022;64:2290–308. https://doi.org/10.1111/jipb.13420.

76. Zhao X, Li J, Lian B, Gu H, Li Y, Qi Y. Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. Nat Commun. 2018;9:5056. https://doi.org/10.1038/s41467-018-07500-7.

77. Wang D, Qu Z, Yang L, Zhang Q, Liu Z-H, Do T, et al. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. Plant J. 2017;90:133–46. https://doi.org/10.1111/tpj.13481.

78. Lv Y, Hu F, Zhou Y, Wu F, Gaut BS. Maize transposable elements contribute to long non-coding RNAs that are regulatory hubs for abiotic stress response. BMC Genomics. 2019;20:864. https://doi.org/10.1186/s12864-019-6245-5.

79. Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet TIG. 2014;30:439–52. https://doi.org/10.1016/j.tig.2014.08.004.

80. Kannan S, Chernikova D, Rogozin IB, Poliakov E, Managadze D, Koonin EV, et al. Transposable Element Insertions in Long Intergenic Non-Coding RNA Genes. Front Bioeng Biotechnol. 2015;3:71. https://doi.org/10.3389/fbioe.2015.00071.

81. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLOS Genet. 2013;9:e1003470. https://doi.org/10.1371/journal.pgen.1003470.

82. Tokizawa M, Kusunoki K, Koyama H, Kurotani A, Sakurai T, Suzuki Y, et al. Identification of Arabidopsis genic and non-genic promoters by paired-end sequencing of TSS tags. Plant J Cell Mol Biol. 2017;90:587–605. https://doi.org/10.1111/tpj.13511.

83. Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet. 2014;15:394–408. https://doi.org/10.1038/nrg3683.

84. Zilberman D, Cao X, Jacobsen SE. ARGONAUTE4 Control of Locus-Specific siRNA Accumulation and DNA and Histone Methylation. Science. 2003;299:716–9. https://doi.org/10.1126/science.1079695.

85. Fonouni-Farde C, Christ A, Blein T, Legascue MF, Ferrero L, Moison M, et al. The Arabidopsis APOLO and human UPAT sequence-unrelated long noncoding RNAs can modulate DNA and histone methylation machineries in plants. Genome Biol. 2022;23:181. https://doi.org/10.1186/s13059-022-02750-7.

86. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–97. https://doi.org/10.1016/s0092-8674(04)00045-5.

87. Borges F, Martienssen RA. The expanding world of small RNAs in plants. Nat Rev Mol Cell Biol. 2015;16:727–41. https://doi.org/10.1038/nrm4085.

88. Kurihara Y, Watanabe Y. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proc Natl Acad Sci. 2004;101:12753–8. https://doi.org/10.1073/pnas.0403115101.

89. Fei Q, Xia R, Meyers BC. Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. Plant Cell. 2013;25:2400–15. https://doi.org/10.1105/tpc.113.114652.

90. Axtell MJ, Jan C, Rajagopalan R, Bartel DP. A two-hit trigger for siRNA biogenesis in plants. Cell. 2006;127:565–77. https://doi.org/10.1016/j.cell.2006.09.032.

91. Lahmy S, Pontier D, Bies-Etheve N, Laudié M, Feng S, Jobet E, et al. Evidence for ARGONAUTE4–DNA interactions in RNA-directed DNA methylation in plants. Genes Dev. 2016;30:2565–70. https://doi.org/10.1101/gad.289553.116.

92. You W, Lorkovic ZJ, Matzke AJM, Matzke M. Interplay among RNA polymerases II, IV and V in RNA-directed DNA methylation at a low copy transgene locus in Arabidopsis thaliana. Plant Mol Biol. 2013;82:85–96. https://doi.org/10.1007/s11103-013-0041-4.

93. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 2013 151. 2013;15:7–21. https://doi.org/10.1038/nrg3606.

94. Gao R, Liu P, Irwanto N, Loh DR, Wong S-M. Upregulation of LINC-AP2 is negatively correlated with AP2 gene expression with Turnip crinkle virus infection in Arabidopsis thaliana. Plant Cell Rep. 2016;35:2257–67. https://doi.org/10.1007/s00299-016-2032-9.

95. Zacharaki V, Quevedo M, Nardeli SM, Meena SK, Monte E, Kindgren P. Convergent antisense transcription primes hosting genes for stress responsiveness in plants. Mol Plant. 2025;18:1920–31. https://doi.org/10.1016/j.molp.2025.10.001.

96. Meena SK, Quevedo M, Nardeli SM, Verez C, Bhat SS, Zacharaki V, et al. Antisense transcription from stress-responsive transcription factors fine-tunes the cold response in Arabidopsis. Plant Cell. 2024;36:3467–82. https://doi.org/10.1093/plcell/koae160.

97. Swiezewski S, Liu F, Magusin A, Dean C. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nat 2009 4627274. 2009;462:799–802. https://doi.org/10.1038/nature08618.

98. Fedak H, Palusinska M, Krzyczmonik K, Brzezniak L, Yatusevich R, Pietras Z, et al. Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript. Proc Natl Acad Sci. 2016;113:E7846–55. https://doi.org/10.1073/pnas.1608827113.

99. Liu W, Sun J, Li J, Liu C, Si F, Yan B, et al. Reproductive tissue-specific translatome of a rice thermo-sensitive genic male sterile line. J Genet Genomics. 2022;49:624–35. https://doi.org/10.1016/j.jgg.2022.01.002.

100. Lin X, Lin W, Ku Y-S, Wong F-L, Li M-W, Lam H-M, et al. Analysis of Soybean Long Non-Coding RNAs Reveals a Subset of Small Peptide-Coding Transcripts1 [OPEN]. Plant Physiol. 2020;182:1359–74. https://doi.org/10.1104/pp.19.01324.

101. Mohammadin S, Edger PP, Pires JC, Schranz ME. Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. BMC Plant Biol. 2015;15:217. https://doi.org/10.1186/s12870-015-0603-5.

102. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012;482:339–46. https://doi.org/10.1038/nature10887.

103. Ross CJ, Ulitsky I. Discovering functional motifs in long noncoding RNAs. WIREs RNA. 2022;13:e1708. https://doi.org/10.1002/wrna.1708.

104. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat Rev Genet. 2016;17:601–14. https://doi.org/10.1038/nrg.2016.85.

105. Simopoulos CMA, Weretilnyk EA, Golding GB. Molecular Traits of Long Non-protein Coding RNAs from Diverse Plant Species Show Little Evidence of Phylogenetic Relationships. G3 GenesGenomesGenetics. 2019;9:2511–20. https://doi.org/10.1534/g3.119.400201.

106. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, et al. Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet 2007 398. 2007;39:1033–7. https://doi.org/10.1038/ng2079.

107. Wang Y, Fan X, Lin F, He G, Terzaghi W, Zhu D, et al. Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. Proc Natl Acad Sci. 2014;111:10359–64. https://doi.org/10.1073/pnas.1409457111.

108. Hawkes EJ, Hennelly SP, Novikova IV, Irwin JA, Dean C, Sanbonmatsu KY. COOLAIR antisense RNAs form evolutionarily conserved elaborate secondary structures. Cell Rep. 2016;16:3087–96. https://doi.org/10.1016/j.celrep.2016.08.045.

109. Vandivier LE, Anderson SJ, Foley SW, Gregory BD. The Conservation and Function of RNA Secondary Structure in Plants. Annual Review of Plant Biology. 2016;67 Volume 67, 2016:463–88. https://doi.org/10.1146/annurev-arplant-043015-111754.

110. Dong Q, Yang B, Sun W, Liang J, Xing Q, Ren L, et al. In vivo RNA structure influences the translation and stability of plant long noncoding RNAs. Plant Commun. https://doi.org/10.1016/j.xplc.2025.101575.