Computational and Structural Biotechnology Journal 21 (2023) 3032–3044

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Long non-coding RNA and microRNA landscape of two major domesticated cotton species

Ajeet Singh [a,b,1], Vivek AT [a,1], Kanika Gupta [a,2], Shruti Sharma [a,2], Shailesh Kumar [a,*,3]

[a] Bioinformatics Lab, National Institute of Plant Genome Research, New Delhi 110067, India
[b] Postdoctoral Associate, Ophthalmology, Baylor College of Medicine, Houston, TX, USA

ARTICLE INFO

Article history:
Received 23 December 2022
Received in revised form 11 May 2023
Accepted 11 May 2023
Available online 12 May 2023

Keywords:
Non-coding RNAs
MiRNA
LncRNA
Cotton
Gossypium hirsutum
Gossypium barbadense

ABSTRACT

Allotetraploid cotton plants *Gossypium hirsutum* and *Gossypium barbadense* have been widely cultivated for their natural, renewable textile fibres. Even though ncRNAs in domesticated cotton species have been extensively studied, systematic identification and annotation of lncRNAs and miRNAs expressed in various tissues and developmental stages under various biological contexts are limited. This influences the comprehension of their functions and future research on these cotton species. Here, we report high confidence lncRNAs and miRNA collection from *G. hirsutum* accession and *G. barbadense* accession using large-scale RNA-seq and small RNA-seq datasets incorporated into a user-friendly database, CoNCRAtlas. This database provides a wide range and depth of lncRNA and miRNA annotation based on the systematic integration of extensive annotations such as expression patterns derived from transcriptome data analysis in thousands of samples, as well as multi-omics annotations. We assume this comprehensive resource will accelerate evolutionary and functional studies in ncRNAs and inform future breeding programs for cotton improvement. CoNCRAtlas is accessible at **http://www.nipgr.ac.in/CoNCRAtlas/**.

1. Introduction

Non-coding RNAs (ncRNAs) constitute a broad class of transcripts that are endogenously produced and play a crucial role in regulating cell function [1]. In addition to introducing new levels of regulation of gene expression, ncRNAs play a significant role in major cellular processes [2]. Although initially considered as "junk" RNA, they are now known to be involved in nearly every developmental process [3]. MicroRNAs have been identified as regulators of plant mRNAs, akin to their role in animals [4]. Furthermore, research on long noncoding RNAs (lncRNAs) has expanded our understanding of gene regulation. It is now apparent that the distinct processes of lncRNA transcription, processing, export, and turnover are directly linked to their diverse functions in a cell [5]. Together, knowledge of miRNAs and lncRNAs has significantly increased with high-throughput sequencing, facilitating genome-wide discovery in several plant processes in recent years [4,6–8].

Studies on cotton ncRNAs have grown over time, but more is needed about the quantity, nature, and expression patterns of lncRNAs compared to miRNAs in cultivated cotton species. Despite the discovery of several lncRNA sequences, extensive functional research on cotton lncRNAs is still in its infancy [9–13]. Over the years, several comprehensive plant ncRNA databases have been developed, but the composition of cotton-associated ncRNAs is limited [14–18]. Existing lncRNA and miRNA databases for cotton have several drawbacks, including a small sample size and lack of comprehensiveness. Many well-known lncRNA databases do not contain cotton lncRNAs, and some databases have few lncRNA annotations due to various factors, including the emphasis on including only experimentally confirmed lncRNAs. Although PLncDB has numerous lncRNAs for both *G. hirsutum* and *G. barbadense*, the number of samples applied to annotate lncRNAs is still limited. Even high quality databases such as miRBase [19], PmiREN [17], sRNAanno [20], and plant small RNA genes [21] have insufficient coverage of miRNA and lncRNA in domesticated cotton species. It is essential to note that *G. barbadense* has no miRNA records in miRBase, widely regarded as the gold standard for miRNA information. Thus, the scope

* Correspondence to: National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali Marg, New Delhi 110067, India.
E-mail address: shailesh@nipgr.ac.in (S. Kumar).
[1] Equal first authors.
[2] Equal second authors.
[3] ORCID ID: https://orcid.org/0000-0002-1872-9903

https://doi.org/10.1016/j.csbj.2023.05.011

and coverage of these cotton ncRNAs in these databases are limited. The non-uniform ncRNA data may hinder progress in cotton science and confuse researchers, given the complexities of ncRNA nomenclature and the distribution of ncRNA annotation across various resources, each with its own quality metrics and definitions of each ncRNA type [2,22]. Solving this problem requires large-scale curation and annotation, leading to a collection of reference ncRNA genes of higher quality [10,23]. Given the growing significance of ncRNAs as major regulators, a comprehensive reference atlas of lncRNA and miRNA would be an invaluable tool for both fundamental and cotton improvement research. Studies on noncoding RNAs in cotton have been largely limited to small RNAs until now, and RNA sequencing has helped to identify hundreds of small noncoding RNAs. To that end, we created CoNCRAtlas, a specialised database of two major ncRNAs, lncRNA, and miRNA that serves as a centralised access point to data from many publicly available samples of two major domesticated cotton species, *G. hirsutum,* and *G. barbadense*. Using transcriptomic datasets from multiple biotypes of both cotton genomes, we mapped miRNAs and lncRNAs that were specific to different tissues as well as those that were broadly transcribed, covering a wide range of expression levels. Our findings shed light on the tissue-dependent distribution of ncRNAs and other key annotations required to understand their regulation in domesticated cotton species.

2. Methods

2.1. Data collection and pre-processing

RNA-seq and small RNA-seq datasets of allopolyploids domesticated cotton species, *G. hirsutum,* and *G. barbadense* were exploited in this work as transcriptome data by mining literature and using NCBI, with the goal of constructing the lncRNA and miRNA landscape in two major cotton species (Fig. S1). The RNA-seq datasets that lacked tissue, development stage/age, or treatment information were removed. Whole genome references were downloaded from https://github.com/Genome-data-of-Gossypium-hirsutum (*G. hirsutum*) and https://ncbi.nlm.nih.gov/data-hub/genome/GCA_008761655.1 (*G. barbadense*). SRA toolkit (v3.0.0) (https://github.com/ncbi/sra-tools) was used to download RNA-seq and small RNA-seq data from NCBI SRA. Table S1 and Fig. S1 contain information on the RNA-seq samples that were analysed. TrimGalore (v0.6.5)(https://github.com/FelixKrueger/TrimGalore) and fastp (v0.23.2) [24] were used to trim adapter sequences before identifying and annotating miRNAs and lncRNAs using the procedures listed below (Fig. 1).

2.2. lncRNA identification and classification

A total of 266 and 1398 RNA-seq datasets for *G. barbadense* and *G. hirsutum* were used for lncRNA identification. To obtain high-confidence lncRNAs, a strict set of criteria was used, considering the redundancy, background noise, mapping error percentage, length, and coding potential. The analytical procedures were performed from raw RNA-seq data to produce high-confidence lncRNAs as described in Fig. 1. At first, low-quality reads were filtered, and adapter sequences were trimmed. Then, clean reads were mapped to the reference genome for each sample using HISAT2 (v2.2.1) [25]. Only libraries with more than 50% of reads mapping to the reference genome were used for further analysis. Further, the reference-based transcript assembly with StringTie (v2.2.0) [26] was performed based on the read-mapping results for each sample to the corresponding reference genome annotations. Transcriptome assemblies derived from the preceding stages were combined with StringTie-merge function to produce complete non-redundant transcripts for further analysis. To further narrow down the genes possibly

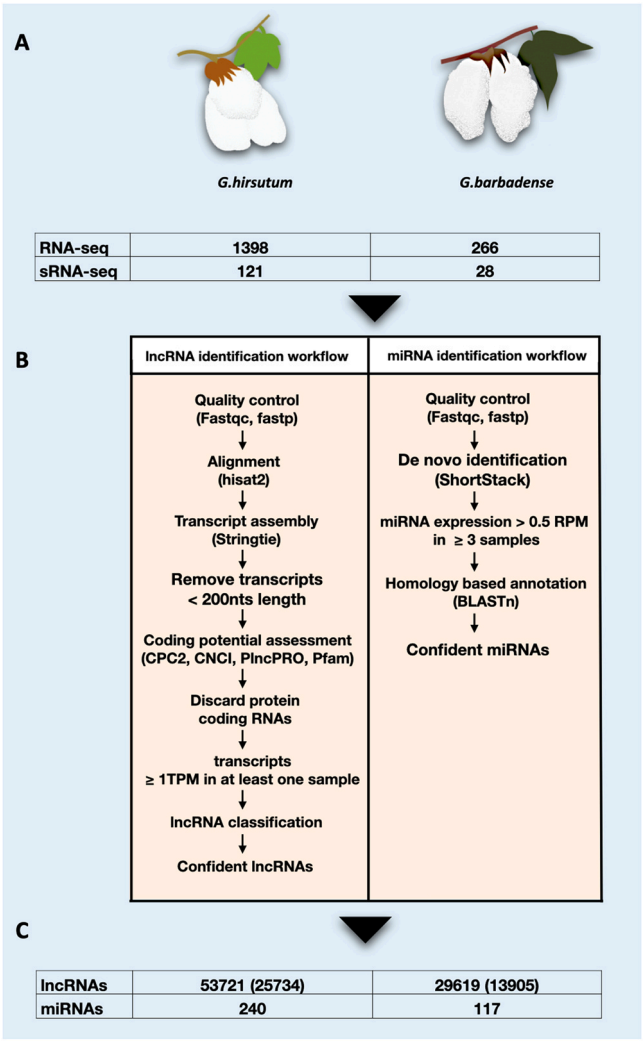

Fig. 1. Workflow diagram of lncRNA and miRNA identification using large-scale datasets. (A) Numbers of bulk RNA-seq datasets and small RNA-seq datasets used in this study. (B) Overview of lncRNA and miRNA annotation pipeline. The lncRNAs and miRNAs are identified using distinct pipelines that enabled expression profiles of cotton miRNAs and lncRNAs, respectively, across diverse biological contexts/conditions based on comprehensive integration of tools, extensive filtering steps, and specialized analysis. (C) Numbers of distinct lncRNA and miRNA in both cotton species. The numbers in brackets indicate the number of isoforms of lncRNAs genes in each cotton species.

encoding lncRNAs (Fig. 1), filtering criteria used for this purpose are as follows, (i) Candidate lncRNA transcripts with 1 TPM expression in at least one sample were included for further analysis [27]; (ii) candidate transcripts with less than 200nt; (iii) Three major tools, CPC2 (v1.0.1) [28], CNCI (v2.0) [29], and PlncPRO[30] were used to estimate coding potential, and only transcripts identified as non-coding by all three methods were retained and the Pfam 2021 [31] protein family database was put to use to remove domain containing sequences; (iv) FEELnc (v.0.2) [32] was used to evaluate the coding capacity of all unannotated transcripts using the options to "-m shuffle," and the shuffled protein-coding transcripts were used as a non-coding training set. If a gene was found to produce mRNA transcripts, it was excluded from the downstream analysis. LncRNAs were further classified relative to the protein-coding genes, and those that failed the FEELnc standards were excluded. Fig. 1 and Fig. S2 depict a comprehensive overview of the lncRNA identification pipeline used in this study, freely available at https://github.com/skbinfo/CoNCRAtlas. Finally, the entire set of annotated lncRNAs for each species is saved in GTF format.

2.3. Conservation analysis of lncRNAs

G. hirsutum and *G. barbadense* lncRNAs were aligned to the *G. barbadense* genome and reciprocally to the *Gossypium hirsutum* genome to assess conservation using HISAT2 (v2.2.1) [25]. Genomic coordinates of mapped lncRNAs and overlapping regions were extracted using bedtools (v2.30.0) [33]. Conservation scores were calculated as the ratio of overlapping region length to lncRNA sequence length. In addition to that, the overlap of lncRNAs to the genome and transcriptome within and across related species, *G. arboreum* and *G. raimondii*, were examined.

2.4. miRNA identification

A total of 121 and 28 samples were considered for miRNA identification in *G. barbadense* and *G. hirsutum*, respectively (Supplementary Table 1). ShortStack (v3.8.5) [34] was used to identify small RNAs from clean sRNA reads. We filtered predicted sRNA clusters to include only those with > = 80% of readings within 20–24 nts in length and miRNAs with at least 2 RPM. The miRNA families were assigned based on homology-based annotation. The annotated miRNAs for each species, including mature miRNA, star miRNA, and precursor miRNA information, were stored in a GTF file. Raw counts and normalised counts were used for further representation and analysis.

2.5. Extraction of cotton ncRNAs from other sources

Information on experimentally confirmed miRNAs and lncRNAs was collected from published literature, and additional support was obtained by manually curating published research articles. The curated information included the name and biotype of the lncRNAs/miRNAs, their sequence and positional information, experimental techniques (such as microarray, Northern blot, and qRT-PCR), experimental samples (such as tissue), lncRNA expression patterns (whether upregulated or downregulated), and PubMed database hyperlinks (PubMed ID, year of publication, and title of the paper).

2.6. Assessment of tissue specificity of ncRNAs

To assess the tissue specificity of lncRNA and mRNA across *G. hirsutum* and *G. barbadense*, we profiled gene expression across all tissues and removed under expressed genes (< 0.1 RPKM). The normalised RPM and TPM expression values of miRNA and lncRNA, respectively, across tissues were used to calculate two tissue specificity metrics, tau [35] and tissue specificity index (TSI) [36], in addition to quantifying gene expression. Hierarchical clustering heatmap was generated by using the R package ComplexHeatmap (v3.1) [37] to examine the relationship between tissue-wise sample grouping with ncRNA expression profiles. Supplementary Table 1 have the name of all libraries categorised under tissues. Rtsne (v0.16) (https://github.com/jkrijthe/Rtsne) was used to perform a t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis. The log (TPM) transformed normalised TPM expression values were applied to all lncRNAs and tissues (Gh-28; Gb-14).

2.7. ncRNA-ncRNA interactions and transposon associations

With the default parameters, psRNATarget (2017 release) [38] was used to predict the mRNA and lncRNA target of microRNAs, as this method enabled the identify potential lncRNA/mRNA-miRNA interactions. An online tool Mercator [39], was used to functionally annotate the corresponding mRNAs of both cotton species to understand the functional aspect of the miRNA targets. Further, lncRNA overlapping miRNA precursors were identified by comparing lncRNA sequences to miRNA hairpin sequences using bedtools (v 2.30.0) [33]. As TE information was unavailable for both the cotton genomes, EDTA (v2.0.0) [40] was used with default parameters to re-identify TEs for both cotton species. Like identifying lncRNA overlapping miRNA precursors, the genomic coordinates of the identified TE were compared to cotton ncRNAs to find TE-lncRNA/miRNA associations.

2.8. Database implementation

CoNCRAtlas is currently hosted on a Linux, Apache, MySQL, and PHP stack. The web interface is built using HTML, CSS, and JavaScript, and is supported by a MySQL relational database that handles ncRNA and relevant annotations. The web interface provides search capabilities, data retrieval, and visualization, utilizing JavaScript and PHP. Interactive diagrams are added using data visualization functions from the Plotly libraries (https://plotly.com). CoNCRAtlas is integrated with standalone BLAST (v2.11.0) [41] for online similarity searches, ViennaRNA (v2.4.16) [42] for secondary structure visualization, and ORFfinder (v0.4.3) [43] for detecting sORFs within ncRNA-producing loci. CoNCRAtlas has been thoroughly tested with various browsers, including Firefox, Google Chrome, Edge, Safari, and Opera. Researchers can access CoNCRAtlas for free by visiting http://www.nipgr.ac.in/CoNCRAtlas/, which requires no registration or login.

2.9. Decoding fibre development associated ncRNAs

R programming language's WGCNA package (v1.71) [44] was utilised to investigate co-expression networks significantly related to fibre development. The hclust function performed hierarchical cluster analysis on the *G. hirsutum* and *G. barbadense* miRNA and lncRNA expression datasets. The normalized gene expression values in the CoNCRAtlas for fibre development stages were used in this study. Hub genes for the relevant modules (PCC cutoff 0.8) were identified using > =0.8 and > =0.8 MM and GS thresholds, respectively. Further, we selected miRNAs common to significant modules found in co-expression of *G. hirsutum* and *G. barbadense* lncRNAs. Annotations for these miRNAs were then searched in the CoNCRAtlas database to present a miRNA-lncRNA-mRNA network model in connection to fibre development. Cytoscape (v3.7.0) [45] was used to visualize the network.

3. Results

3.1. Composition of lncRNA in two major cotton species

The analysis conducted on the samples to obtain comprehensive annotation for lncRNAs in the genomes of *G. barbadense* (Gb) and *G. hirsutum* (Gh) is shown in Fig. 1. To systematically integrate data from multiple RNA-seq strategies and accurately identify lncRNAs, we created the lncRNA pipeline depicted in Fig. S2. A total of 39,639 cotton lncRNA genes were identified (Gh −25,734; Gb −13,905) based on a large-scale analysis of bulk RNA-seq libraries from various biotypes (Fig. S1C-E). Each species had about twice as many lncRNA transcripts as actual genes due to isoforms. In terms of length, *G. hirsutum* had a greater number of lncRNA transcripts than its counterpart, but this could be attributed to the limited RNA-seq libraries used for identification in the case of *G. barbadense* (Fig. 2A). However, it is worth noting that the proportion of lncRNAs found in each chromosome relative to each species is strikingly similar (Fig. 2B). Most of the genes had 1–5 isoforms, with only 55 and 35 lncRNA genes found to have more than 20 isoforms in *G. hirsutum* and *G. barbadense*, respectively (Fig. 2C). These genes may be useful candidates for understanding the complexities of alternative splicing in both cotton species. It is clear from the combined annotation that lncRNA transcripts often have two or more exons and/or fewer than

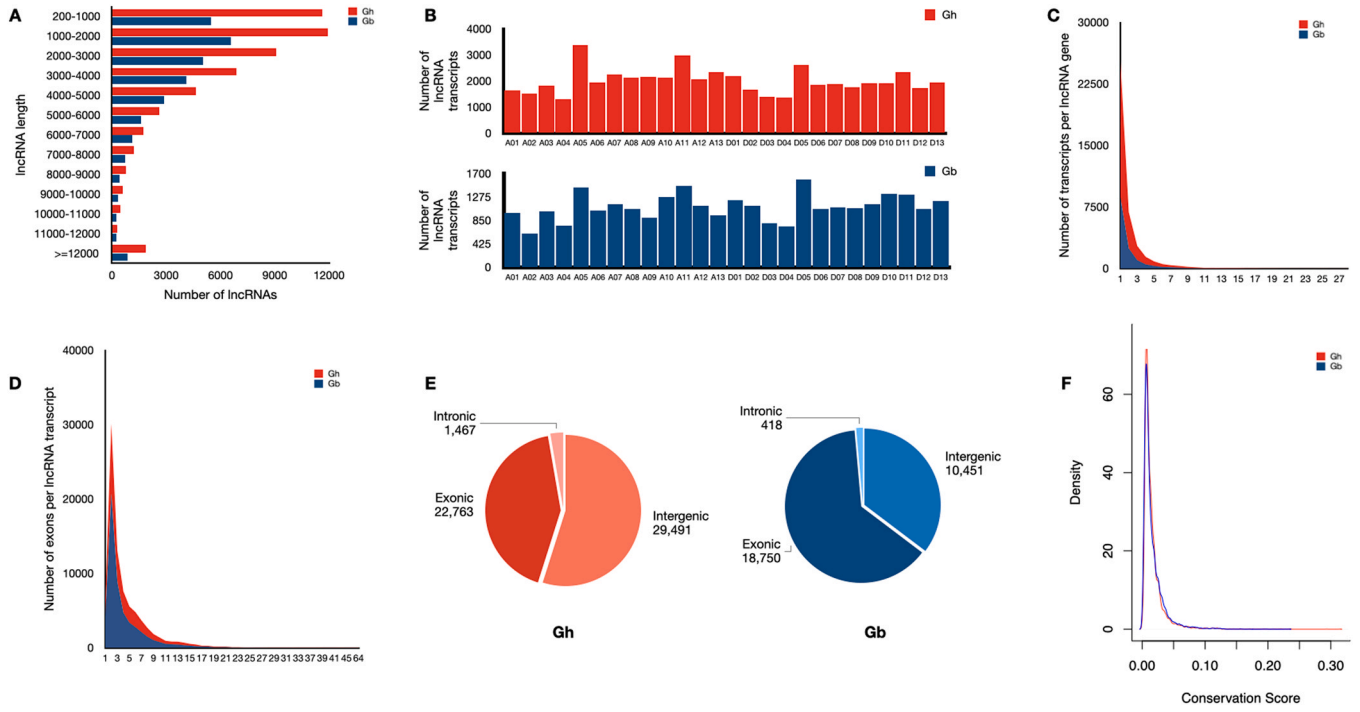

Fig. 2. Basic features of identified lncRNAs in domesticated cotton species. (A) Length-wise distribution of lncRNAs; (B) chromosome-wise distribution of lncRNAs; (C) distribution of the number of transcripts per lncRNA gene/locus; (D) overall distribution of the number of exons per transcript; and (E) proportion of identified lncRNAs based on their genomic locations relative to protein-coding genes. (F) Conservation scores of identified lncRNAs (Gh- *G. hirsutum*; Gb- *G. barbadense*).

12 exons (Fig. 2D). Moreover, the existence of lncRNA transcripts with two exons is more prevalent in both species, following a similar pattern observed in many other species. Based on the combined lncRNA annotations, most lncRNAs were exonic in *G. barbadense*, whereas intergenic in *G. hirsutum* (Fig. 2E and Table S2). The number of exonic lncRNAs in both species was nearly identical, but the composition of intergenic lncRNAs differed significantly, with *G. hirsutum* containing three times the number of *G. barbadense*.

3.2. Comparative analysis of LncRNA sequence conservation in cotton species

LncRNA sequence conservation in *G. barbadense* and *G. hirsutum* revealed that these lncRNAs possess a greater degree of similarity with their respective genomes, but they still exhibit some level of conservation with related species *G. raimondii* and *G. arboreum*. In terms of transcriptomes, both species were observed to have a high level of similarity with all genomes, with the exception of the *G. hirsutum* genome, which may exhibit some divergence (Fig S4A).

Compared to coding genes, lncRNAs generally exhibit lower sequence conservation rates [46]. Our analysis revealed exceptionally low conservation among lncRNAs when comparing *G. hirsutum* and *G. barbadense*. The genomes and lncRNAs of both species were analysed to find the degree of overlap (Fig. 2F). Moreover, we computed conservation scores among cotton species transcripts using identified lncRNA transcripts and found that the majority of lncRNAs had low sequence conservation scores, with only a few demonstrating higher scores (Fig. S4B-C).

3.3. Expression-based annotation of cotton miRNAs

The annotation of miRNAs was performed by utilizing 124 sRNA-seq datasets from two domesticated cotton species (Gh-100 and Gb-24), with each RNA-seq library producing 1–10 million reads that mapped to *G. hirsutum* and *G. barbadense*, respectively. Our analysis revealed that the present miRBase annotation of cotton miRNAs is incomplete and can be expanded using new sRNA-seq data. With the aid of our built-in workflow, we successfully identified and annotated 357 cotton miRNAs (Gh-117; Gb-240) with full-length deep sequencing datasets, providing greater coverage (Fig. 1C). Of these, 280 miRNAs (Gh-198; Gb-28) were members of 78 families, while the remaining 77 miRNAs (Gh-42; Gb-35) were novel. Consistent with other plant species, we observed that the cultivated *Gossypium sp.* has the highest frequency of miRNAs with lengths of 21 nt, while miRNAs with lengths of 23 nt are the least abundant (Fig.S5B).

Due to the allotetraploid nature of the genomes, distinct miRNA gene families were identified, with many members in each family. We discovered 106 families in *G. hirsutum* and 53 families in *G. barbadense*, for a total of 116 distinct families in the two cotton species (Fig. 3A). The identified miRNAs belong to 78 different miRNA families, with miR166 being the most prominent. The most prevalent miRNA families across both species are MIR156, MIR166, MIR167, MIR171, MIR172, and MIR396. *G. hirsutum* has more unique miRNAs than *G. barbadense*, but this could change with more libraries for identification and annotation. Furthermore, we identified 38 common miRNA families in the two cotton species, of which 27 were unique to *G. hirsutum* and 10 to *G. barbadense*. For example, MiR396 is found in *G. hirsutum* but not in *G. barbadense*. We observed seven miRNA families with at least ten members in *G. hirsutum*, *G. barbadense*, or both (miR166, miR156, miR172, miR482, miR171, miR167, and miR396). According to our data, each species contains at least five members of the miR156, miR172, and miR482 families. The findings revealed that the members of miR156 and miR166 are widely distributed across both cotton species. The contrasting composition of members can be seen in miR396, which is low in *G. barbadense*. However, we anticipate that the discrepancies will become clearer as additional sRNA-seq samples become available. As a result of the workflow, all miRNA precursors were identified as high-confidence transcripts with broad coverage for mature and star sequences. Using the output of the ShortStack pipeline, we were also able to find accurate information on the star sequence for all miRNAs.

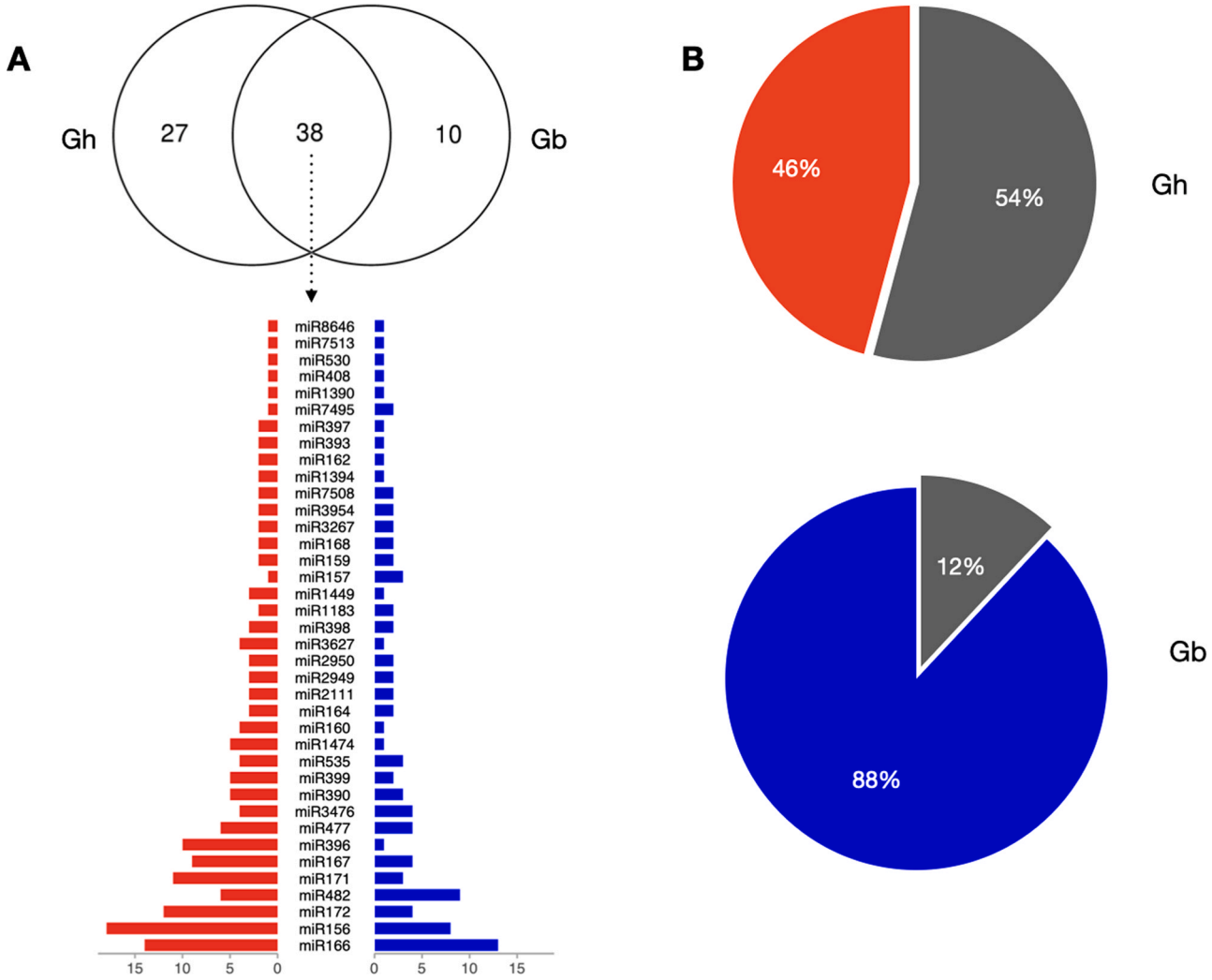

Fig. 3. miRNAs identified and annotated in domesticated cotton. (A) Venn diagram showing the number of miRNA families conserved between *G. hirsutum* and *G. barbadense*. Conservation analysis indicates that miRNA families are conserved between the two cotton species, and their distribution of miRNA members across both cotton species is shown in a bar chart; (B) A pie chart depicts the proportion of miRNAs found in the literature, with grey wedges indicating the number of validated miRNAs in each species.

3.4. Known miRNAs and few lncRNAs by literature mining

To conduct a literature search in the PubMed and Google Scholar search engine, we used the following keyword strategy: (lncrna or long noncoding or long non-coding RNA or noncoding or miRNA or microRNA or *Gossypium hirsutum* or *Gossypium barbadense* or cotton). The results were sorted by cotton species and limited to publications up to December 2020 and ranked based on meeting our criteria. We carefully assessed each publication's title, abstract, keywords, and entire text to identify studies that presented ncRNA annotations, databases, and functions. Only high-quality associations with multiple lines of strong experimental evidence, confirmed by RNAi, in vitro knockdown, Western blot, qRT-PCR, or luciferase reporter assays, were considered. We re-checked all selected studies for the miRNA/lncRNA names and replaced some with official or recommended names. We also collected other names, including aliases and synonyms, for both miRNAs/lncRNAs in this step.

Compared to protein-coding genes, lncRNAs remain poorly understood as most lack functional annotation. Nonetheless, the number of publications, including the keyword "long non-coding RNA" has increased in recent years, although the literature is skewed towards a few well-studied lncRNAs. Despite this, at least 25 lncRNAs have been functionally investigated (Fig. S6). Unfortunately, sequencing and annotating lncRNAs is challenging due to the lack of official names, reliable identities, or independent identifiers. Through manual curation of lncRNA papers in various species, several annotation groups have accumulated valuable functional resources (Bai et al., 2019; Gallart et al., 2016b; Zhou et al., 2021). In contrast to model plants like *Arabidopsis thaliana* and *Oryza sativa*, cotton lncRNAs have not been successfully annotated for sequence research or included in lncRNA-specific databases. Literature mining has yielded more validated cotton lncRNAs than database searches. However, for miRNAs, literature mining found 89 (Gh-72 and Gb-17) research associated with cotton miRNAs, 24 of which reported experimentally validated miRNAs (Fig. 3B). In contrast, only one publication exists for 124 lncRNAs (Gh-110; Gb-14), indicating a dearth of such studies for lncRNAs so far (Fig. S6). Literature mining methods annotated over 50% of *G. hirsutum* with previously published literature, while only 12% of *G. barbadense* miRNAs were associated with known literature (Fig. 3B). Finally, data mining-based ncRNA information is collated and added to the CoNCRAtlas database.

3.5. Tissue-specific expression of miRNAs and lncRNAs

In different sample groups of tissues from *G. hirsutum* and *G. barbadense*, we examined the expression of ncRNAs. Due to the lack of availability, samples from some tissue groups in one of the species

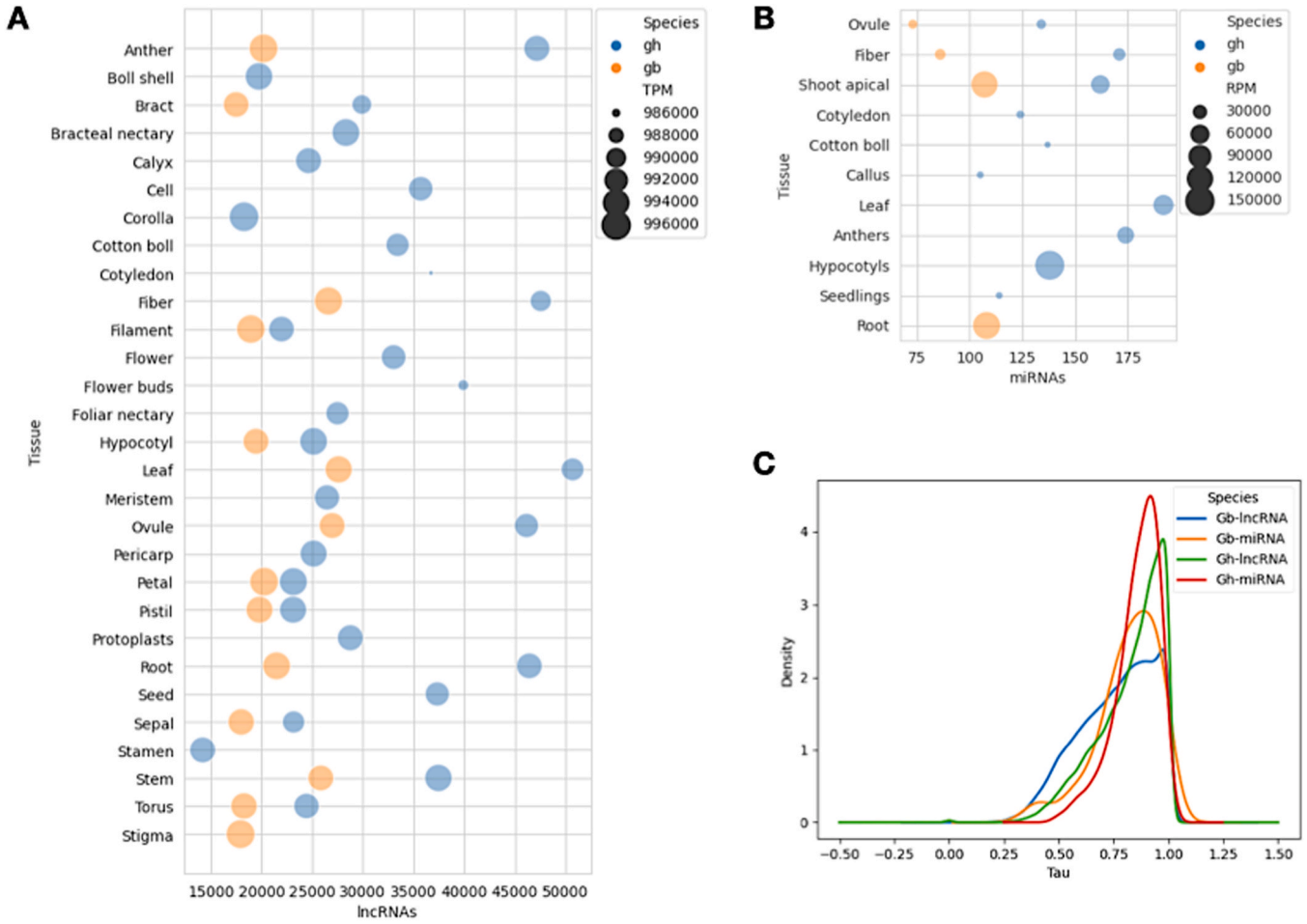

Fig. 4. Tissue-specific expression of lncRNAs and miRNAs in *G. hirsutum* and *G. barbadense*; (A) Coverage of lncRNAs and miRNAs in the profiled tissues. If a lncRNA was found in a tissue, it was considered transcribed if it was found at > 1 TPM, and for miRNAs, > 1 RPM; (B) Average Tau score of identified lncRNAs and miRNAs in *G. hirsutum* and *G. barbadense* across tissue sample groups. The density function in R was used to generate the kernel density estimates shown in this graph.

might not be represented. In most tissues in both species, approximately 18,000–30,000 lncRNAs are normally observed to be expressed with a cumulative expression > 10,000,00 TPM (Fig. 4A). Although cotyledon and flower buds have low expression levels, the number of lncRNAs expressed is higher than in other tissue types. The anther, fibre, and root tissues of *G. hirsutum* and *G. barbadense* differ in the lncRNAs responsible for overall expression. This difference can also be seen in the proportion of miRNA numbers in expression. Notably, the number and expression of miRNAs are higher in roots and shoots in *G. barbadense*, whereas hypocotyl miRNAs show their maximum representation in *G. hirsutum* (Fig. 4B). Additionally, we observed that miRNAs and lncRNAs are the main contributors to tissue specificity, with Tau > 0.8 in both species (Fig. 4C). According to our findings, less than 20% of all ncRNAs were found in only one tissue (Tau = 1). The remaining ncRNAs were either widely expressed or only found in certain tissues. Thus, ncRNAs appear to be expressed in a more context-specific manner than protein-coding RNAs.

We performed a hierarchical clustering of the gene expression data for lncRNA and miRNA independently to validate the annotation and expression profile analyses. The log transformed TPM and RPM expression data from all tissues of the two species were used for this investigation. Visualization of expression patterns shows that each sample group has a distinct ncRNA profile that distinguishes it from the others and has uncovered the presence of distinct transcriptional signatures to tissue types (Figs. 4A and 5). As it is evident from several other studies [9,11,12,47] and from Fig. 5, the majority of the lncRNAs are expressed at low in comparison to protein-coding genes. This approach provides an atlas of the ncRNA expression in domesticated cotton.

3.6. ncRNA transcriptomes are broadly rewired by transposable elements

A number of lncRNAs have been found to be derived from transposons in plants, and it has been discovered that TE-related lncRNAs exhibit tissue-specific transcription and play important roles in plant abiotic stress responses [48–50]. In the case of cotton, at least one study shows that LINEs derived from TEs play an important role in the origin of lncRNAs [51]. In this study, we set out to discover the overlapping regions of ncRNAs and transposable elements. The distribution of lncRNAs was found to be similar to that of TE, which was found all over the genome (Fig. 6). In *G. hirsutum* and *G. barbadense*, we found 213171 and 103969 TEs overlapping lncRNAs, respectively, indicating a twofold increase in the number of lncRNAs overlapping different classes of TEs (Table S3). We observed that the presence of TEs near the ends is unusually abundant, and we observed an increase in the expression of lncRNAs overlapping those regions (Fig. 6). Some miRNAs are thought to have originated naturally from TEs in cotton [52]. However, the number of precursor miRNAs overlapping TEs in *G. hirsutum* (18) was significantly lower than in *G. barbadense* (3). Following that, we discovered many lncRNAs (588 in *G. hirsutum* and 230 in *G. barbadense*) overlapping 128 and 56 precursor miRNAs in *G. hirsutum* and *G. barbadense*,

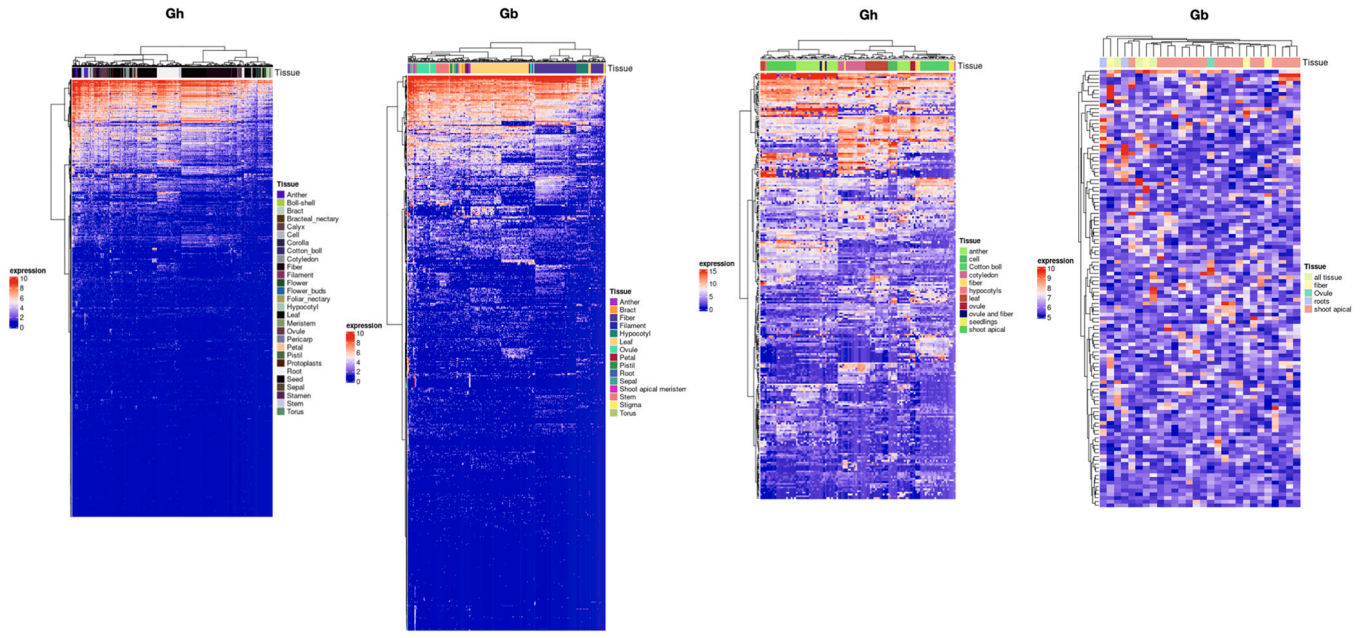

Fig. 5. Cluster heatmaps of identified lncRNAs and miRNAs in *G. hirsutum* and *G. barbadense* across diverse tissue sample groups. The first two heatmaps from the left show heatmaps generated based on the expression of lncRNAs, followed by two heatmaps of miRNAs.

respectively. Interestingly, we discovered that two of the precursor miRNAs, two transposons, and nine lncRNA transcripts of *G. hirsutum* overlapped, whereas such a combination of overlap was not observed in *G. barbadense*. Further research into these overlapping molecules may reveal novel molecular mechanisms that regulate ncRNAs and transposable elements.

Even though understanding the precise molecular functions of lncRNAs is still limited, only a little about their evolution is known. However, much remains to learn about how lncRNAs emerge from TE sequences in plants. A catalogue of the overlapping non-coding RNAs to the TEs is necessary to uncover the origins of lncRNAs associated with TEs. It can assume that transposable elements are responsible for a sizable fraction of the cotton ncRNA sequences. The considerable differences in transposable-element load among cotton genomes, at least 60% of the *G. hirsutum* genome to 70% of the *G. barbadense* genome, made it possible to determine the association of these elements to the emergence of ncRNAs (Table S2). This comprehensive study revealed overlaps that suggest these TE regions were a part of miRNA and lncRNA genes during cotton evolution. Furthermore, ncRNAs linked to TE sequences perform a variety of

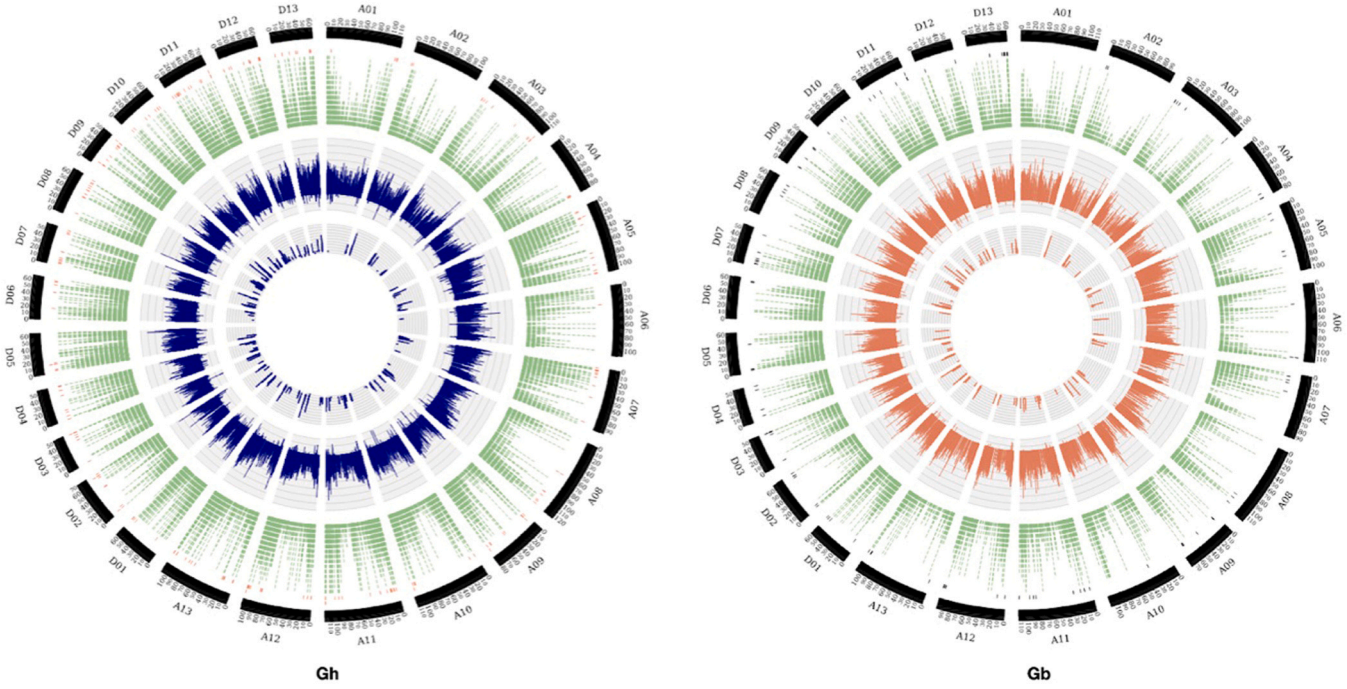

Fig. 6. Genomic map of ncRNA expression and overlapping transposons in cotton. The outermost track represents cotton genome ideograms for all chromosomes. The presence of miRNAs and lncRNAs can be seen in the innermost and next-to-innermost tracks, respectively. The bars show the log-transformed normalised expression count of ncRNAs (*G. hirsutum* = Blue, *G. barbadense* = Orange). The green tiles in the outer track represent transposable elements widespread across the chromosomes.

regulatory functions. For example, these TEs have the potential to act as regulatory signals for lncRNA genes and could act as "sponges" for miRNAs. Thus, cotton-related ncRNAs and TEs may present more interesting biology in the future.

Despite limited knowledge regarding the precise molecular functions of lncRNAs, their evolution remains poorly understood. To investigate the origins of lncRNAs associated with TEs in plants, a comprehensive catalog of overlapping non-coding RNAs to the TEs is required. It is believed that transposable elements constitute a significant proportion of cotton ncRNA sequences, with at least 60% of the *G. hirsutum* genome and 70% of the *G. barbadense* genome comprising these elements (Table S2). The identified overlaps suggest these TE regions were part of miRNA and lncRNA genes during cotton evolution. Furthermore, TEs can act as regulatory signals for lncRNA genes and "sponges" for miRNAs, providing insight into the regulatory functions of ncRNAs linked to TE sequences in cotton [50,53].

3.7. Dynamics of miRNA targeting mRNAs and lncRNAs

Plant miRNAs direct post-transcriptional regulation by binding to their targets with perfect or nearly perfect complementarity, leading to mRNA cleavage or translation suppression. Besides controlling mRNAs, studies have shown that miRNA interactions with lncRNAs further regulate the transcriptome. Additionally, lncRNAs can also carry out their roles by being targeted by the miRNAs [54–56]. Thus, miRNA targeting lncRNAs can be predicted in the same way as miRNA targeting mRNAs, using the miRNA-target complementarity [57]. Our analysis using the scoring schema v2 (2017 release) suggested by Axtell (2013) identified 156,211 mRNA targets (Gh-122485; Gb-33726) and 264,966 lncRNA targets (Gh-201369; Gb-63597). A miRNA typically targeted 10 mRNAs, and a lncRNA was targeted by at least 2 miRNAs, demonstrating extensive regulation networks. In total, 5558 and 5606 mRNAs were predicted as miRNA targets in *G. hirsutum* and *G. barbadense*, respectively, accounting for 7.6% and 7.5% of the PCGs (Table S4). The functional annotation of mRNA targets revealed that most belonged to functional categories (Mapman bins) associated with large enzyme families and solute transport proteins. Therefore, these results will contribute to understanding key interactions induced by miRNAs, competition between miRNAs and lncRNAs for the same mRNA target, miRNA biogenesis from lncRNAs, and lncRNAs functioning as miRNA decoys in cotton.

3.8. Database content and features

CoNCRAtlas offers a user-friendly web interface for reliable database searches, even for users with minimal bioinformatics knowledge. Users can easily explore the database through the Browse, Search, and BLAST modules (Fig. 7).

3.8.1. Search

The CoNCRAtlas database offers users distinct search engines for miRNAs and lncRNAs in each of the cotton species (Fig. 7A). The miRNA search module includes four filtering fields: organism (*G. hirsutum*/*G. barbadense*), CoNCRAtlas ID (e.g., CoMIRGB001), and locus (e.g., A01:1000–9000). The locus filtering field allows users to search for all lncRNAs within specified coordinates by specifying chromosome, start, and end positions. Once the filtering options are selected, the "Search module" generates a table that lists the set of lncRNAs/miRNAs with essential information.

Extensive filtering options, such as an identifier (e.g., CoLNCGB00001), location (upstream, downstream, intronic, or exonic), lncRNA type (intergenic or genic), orientation (sense, antisense, or unknown), exon numbers, locus coordinates, and lncRNA length, are available in the lncRNA search module. Users can filter the results based on the chosen species. Additionally, the webpage header has a search engine for keywords or lncRNA identifiers to explore the complete database fields. The outcome will be a comprehensive list of all relevant ncRNAs.

CoNCRAtlas also offers a sequence-based search engine that uses an online BLAST interface. Relevant search engines are provided to look up ncRNA-related information, such as genomic coordinates and expression values. Dedicated buttons simplify downloading the results for each module. Additionally, users can find publications, biological sources, miRNA interactions with lncRNA/mRNA, associated transposons, predictions of miRNA targets, and related functional data. These details make it easy to interpret the biological significance of the results.

3.8.2. Browse

The Browse module facilitates exploration of miRNAs and lncRNAs in various biotypes, presenting a set of expressed ncRNAs categorized by samples. Users can browse all ncRNAs by selecting the organism, ncRNA type, and sample type, directing them to individual lncRNA and miRNA species lists. The Browse page's hierarchical structure enables convenient search of ncRNA sets from species to sample biotype (Fig. 7B). The output table displays the precise number of ncRNA types expressed in the sample biotype, and users can customize the number of records shown per page (Fig. 7C). LncRNAs and miRNAs are categorized based on tissue (e.g., leaf, fiber), developmental stage (e.g., 0 DPA, 3 DPA), and treatment (e.g., *Fusarium oxysporum* infected) for each cotton species. The Table of Entries presents brief information on each lncRNA/miRNA, including their genomic coordinates. The CoNCRAtlas ID can be used to access individual lncRNA detailed annotations (as shown in Fig. 7C&D).

3.8.3. Annotation details page of lncRNA

CoNCRAtlas assigns a unique accession number to each transcript, which corresponds to a specific webpage containing basic information such as symbol, genomic context, length, exon number, GC content (as a percentage), classification, sequence, coding potential, and multi-omics data including expression and lncRNA/mRNA-miRNA interaction (as shown in Fig. 7D-L). The transcriptome datasets expression levels are displayed categorically based on various samples such as tissues, developmental stages, and treatments. CoNCRAtlas profiles expression levels of lncRNAs across collected tissues and visualizes them in a bar chart, enabling users to explore their functional significance.

Furthermore, the additional information section allows users to investigate lncRNAs overlapping with TE, miRNA targets, and sORFs within lncRNA loci (as shown in Fig. 7E & H). Like the lncRNAs, miRNA targets are also linked to extensive annotations, enabling users to decode the functional mechanisms displayed in the figures. Although only a small fraction of identified miRNAs and lncRNAs in cotton have experimental evidence with supported publications, CoNCRAtlas provides comprehensive function annotations based on manual curation for all the featured ncRNAs. CoNCRAtlas describes each miRNA-target association using Mapman vocabularies [39] that outline the functions and biological processes in which they might be involved. Also, it is important to note that, at present, not all known validated lncRNAs can be linked back to their chromosomal locations and are therefore excluded from the analysis in CoNCRAtlas.

3.8.4. Annotation details page of miRNA

Each identified miRNA is assigned a unique accession number and has a dedicated webpage displaying detailed molecular features, including miRNA ID, miRNA family, location, miRNA sequence, length, stem-loop sequence, miRNA star sequence, and miRNA precursors. The webpage also provides information on targets, associated transposon elements, and expression profiles across

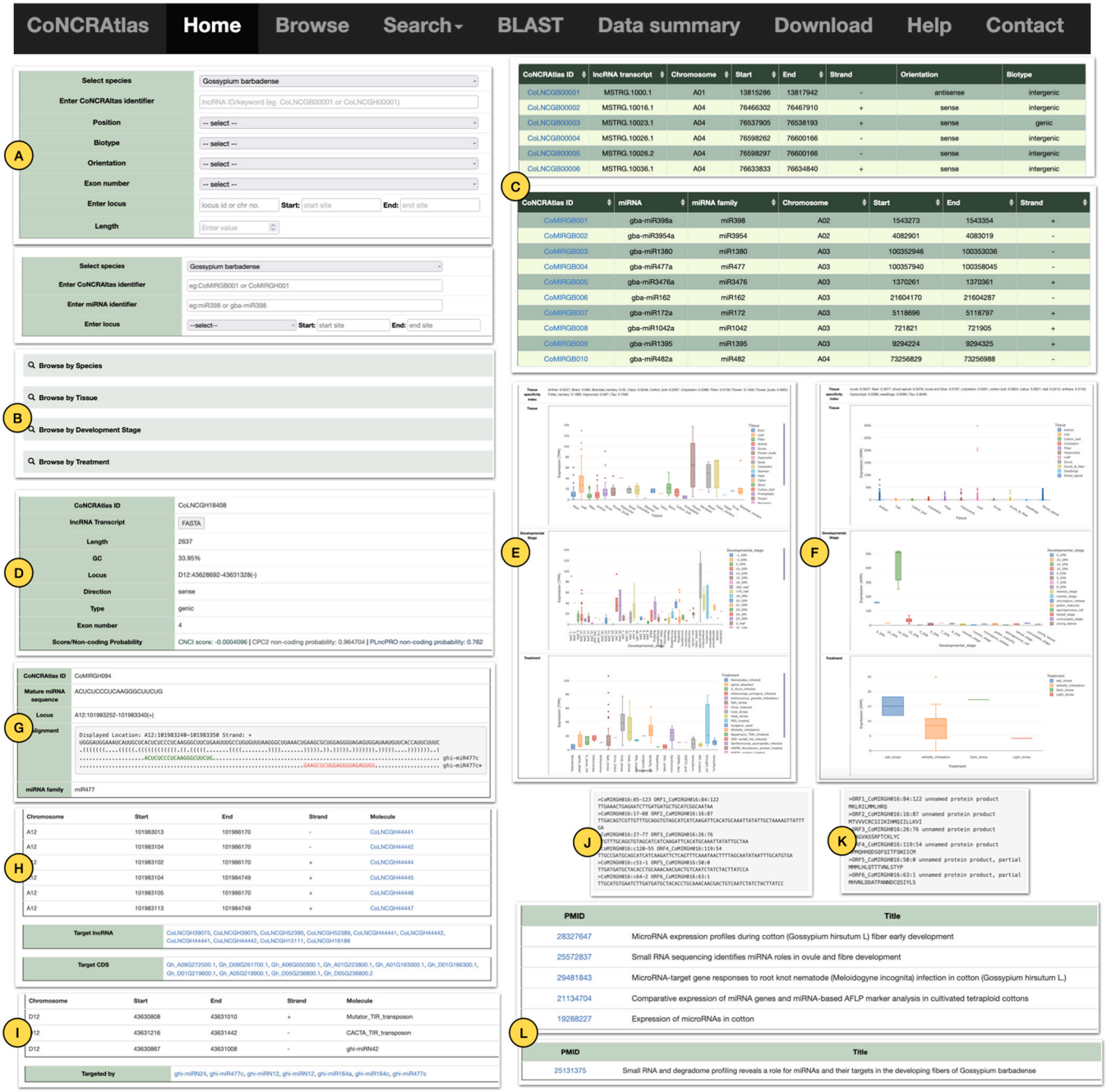

Fig. 7. Snapshots depicting the CoNCRAtlas interface. Users can explore various database functionalities and interactive visualisations through a dedicated menu bar; (A) lncRNAs/miRNAs can be refined using various filtering options for each species by querying with miRNA and/or lncRNA names, genomic location, and/or other multiple filtering combinations; (B) Hierarchical structure of browse page option section for each cotton species under diverse tissues, developmental stages, and treatment samples; (C) The user's browse and search results in a list of miRNAs and lncRNAs that can be arranged in ascending or descending order. These lncRNA/miRNA are supplemented with active links to the annotation details page (D-L). lncRNA and miRNA transcripts are annotated with basic characteristics (D & G) and with an abundance of multi-omics data, including expression profile and tissue specificity (E & F), overlap with lncRNAs, miRNAs (H & I), and transposons, along with miRNA targets, sORFs, and small peptides (J & K), and published literature on Ghi-miR177a and Gba-miR177a (L).

categorized RNA-seq samples, as shown in Fig. 7G-K. Dynamic bar charts are available to view or download tissue specificity scores and their average expression across different tissue types.

To provide details on annotated interaction partners of mRNAs/lncRNAs and miRNAs, extensive literature curation has been conducted. The webpage provides the miRNA target, its regulation mode, and an associated PubMed reference (Fig. 7 L). Users can quickly reference this information when searching through published literature resources for experimentally verified miRNA-mRNA targets and other essential information related to cotton.

3.8.5. Help page

To encourage users to become familiar with the database, CoNCRAtlas offers help throughout the website, providing explanations on how to access the database, tabular results, data statistics, and complete documentation on the usage of each database module.

3.8.6. Download

CoNCRAtlas allows users to download all basic information on ncRNAs in each species. This can be done in bulk or for each cotton species using the Download button in the toolbar. On the Download

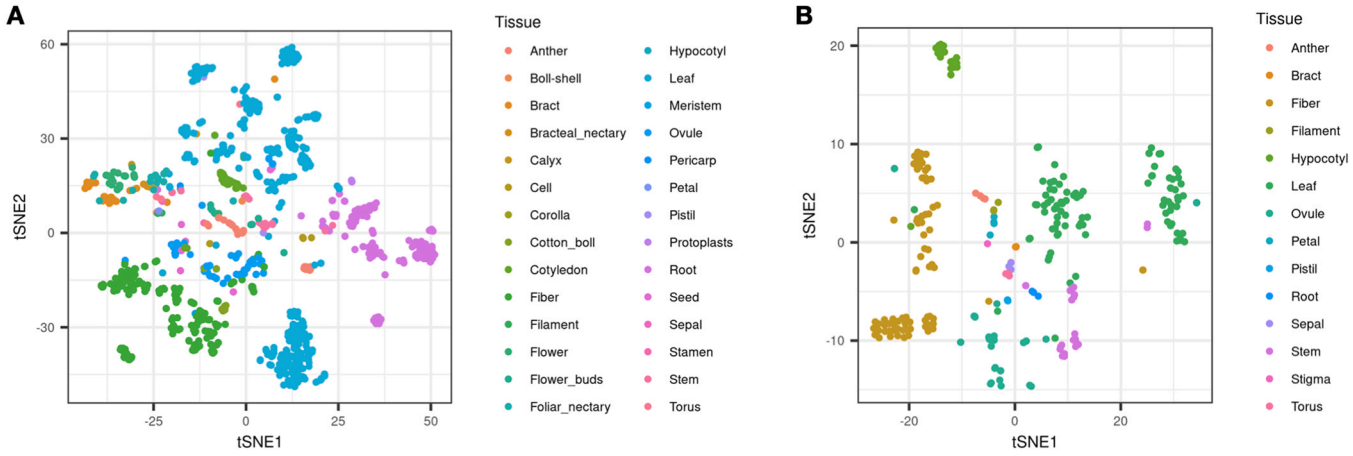

Fig. 8. t-SNE analysis of CoNCRAtlas samples using lncRNA expression data of (a) *G. hirsutum;* and (b) *G. barbadense*.

page, users are presented with three different ports to download data. Firstly, users can download all available information in bulk as text files. Secondly, using the user-defined download port, users can select the information they want by selecting species and data categories. Finally, users can access the genomic coordinates of ncRNAs, which can be easily downloaded. In addition, a backup copy of all data in CoNCRAtlas with the ID 7057078 has been uploaded to Zenodo (https://zenodo.org/).

4. Technical validation

To validate the annotation and expression profiling analysis, we utilized t-SNE on the gene expression data for lncRNA in *G. hirsutum* and *G. barbadense* separately. This allowed us to investigate the gene expression similarity between tissues and across CoNCRAtlas samples, as well as to summarize the lncRNA tissue-specific expression. In this study, we used TPM-normalized lncRNA expression data from all tissues of both species. Using t-SNE to reduce the dimensionality of lncRNA resulted in a robust clustering of samples based on tissue types (Fig. 8). The cluster heatmaps and t-SNE of lncRNA expressions reveal that each tissue has its transcriptional signature, which can be used to differentiate between tissues (Figs. 5 and 8).

The aim of this study was to investigate the roles of long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) in the development of cotton fibres. We utilized the weighted gene co-expression network analysis (WGCNA) approach to identify co-expression networks and hub lncRNAs significantly linked with fibre development. The DPA-related samples included in the analysis are outlined in Table S7. So, we performed co-expression analysis for *G. hirsutum* lncRNAs, *G. barbadense* lncRNAs, and *G. hirsutum* miRNAs, resulting in the partitioning of the WGCNA networks into 66, 58, and 5 co-expression modules, respectively. We identified associations among several lncRNA-lncRNA pairs in these co-expression networks, suggesting their involvement in the same fibre developmental stages and specific biological processes.

To investigate the potential role of lncRNAs and miRNAs in fibre development, we used co-expressed lncRNAs from *G. barbadense* 5DPA (MEmagenta module) and *G. hirsutum* 6DPA (MECyan module) based on significant module trait assessment. We also attempted to determine the function of a specific set of miRNAs in both species using CoNCRAtlas data (Fig. 9). Nine common members of the miR1183, miR171, miR3476, miR477, miR482, and miR530 families were identified as potential targets for co-expressed lncRNAs from the WGCNA analysis using CoNCRAtlas miRNA data. We further examined the associated data of the top lncRNAs and mRNAs connected with it, as well as the transposons, to represent a biologically relevant network model on fibre development for both species.

Our analysis identified three lncRNAs, CoLNCGH20453, CoLNCGH252390, and CoLNCGH18853, which may partially explain the development of fibres in *G. hirsutum*. Our analysis suggests that cotton lncRNAs can act as miRNA sponges, reducing the regulatory effect of miRNAs on mRNAs and adding another layer of complexity to the miRNA-target interaction network. Using the CoNCRAtlas database, we identified miRNAs preferentially expressed in fibres, where the miRNA precursor is targeted by the same miRNA produced by the locus and overlaps with the corresponding antisense/sense lncRNA. This approach identified a possible regulatory module involving miR477b-COLNCGH39075 in fibre development in *G. hirsutum*. The expression data suggest that miR477b plays a crucial role in regulating the GRAS type transcription factor, poly-lycopene isomerase enzyme, and cellulose-synthase like protein G3, indicating an important function during fibre development. We also discovered a similar miR477b-CoLNCGB10457 module of miRNA-mediated mechanism coupled with sense lncRNA in *G. barbadense*, which could explain the differences in fibre development between the two species. It is worth noting that transcript expression alone may not be sufficient, as transcription and splicing are not expected to be the same in every tissue type, leading to differences in abundance and splicing.

4.1. Summary and outlook

In this study, a comprehensive analysis of miRNAs and lncRNAs in two important domesticated cotton species, *G. hirsutum,* and *G. barbadense* is presented. We annotated numerous high-confidence lncRNA genes, miRNAs and identified tissue-specific expression patterns of ncRNAs, providing insights into lncRNA annotation and isoform diversity. Additionally, we found that the current miRBase annotations of cotton miRNAs are incomplete, and further sRNA-seq samples are needed to expand the annotation.

To understand the potential functionality of lncRNAs, researchers conducted transcript-level assessments of their conservation across other cotton plant species. The findings are consistent with current trends regarding the conservation of lncRNAs, indicating that conservation level may be associated with lncRNA function, which may vary among different cotton species [11,58–60]. However, we recognize that conservation scores alone may not be conclusive in identifying functional lncRNAs, and additional experimental validation is essential to confirm their roles in gene regulation. Further, the analysis of non-coding RNA (ncRNA) expression in different tissue samples of two cotton plant species, *G. hirsutum,* and *G. barbadense*, revealed that ncRNAs are expressed in more context-specific manner than protein-coding RNAs, and that lncRNAs are expressed at lower levels compared to protein-coding genes. The tissue-wise data

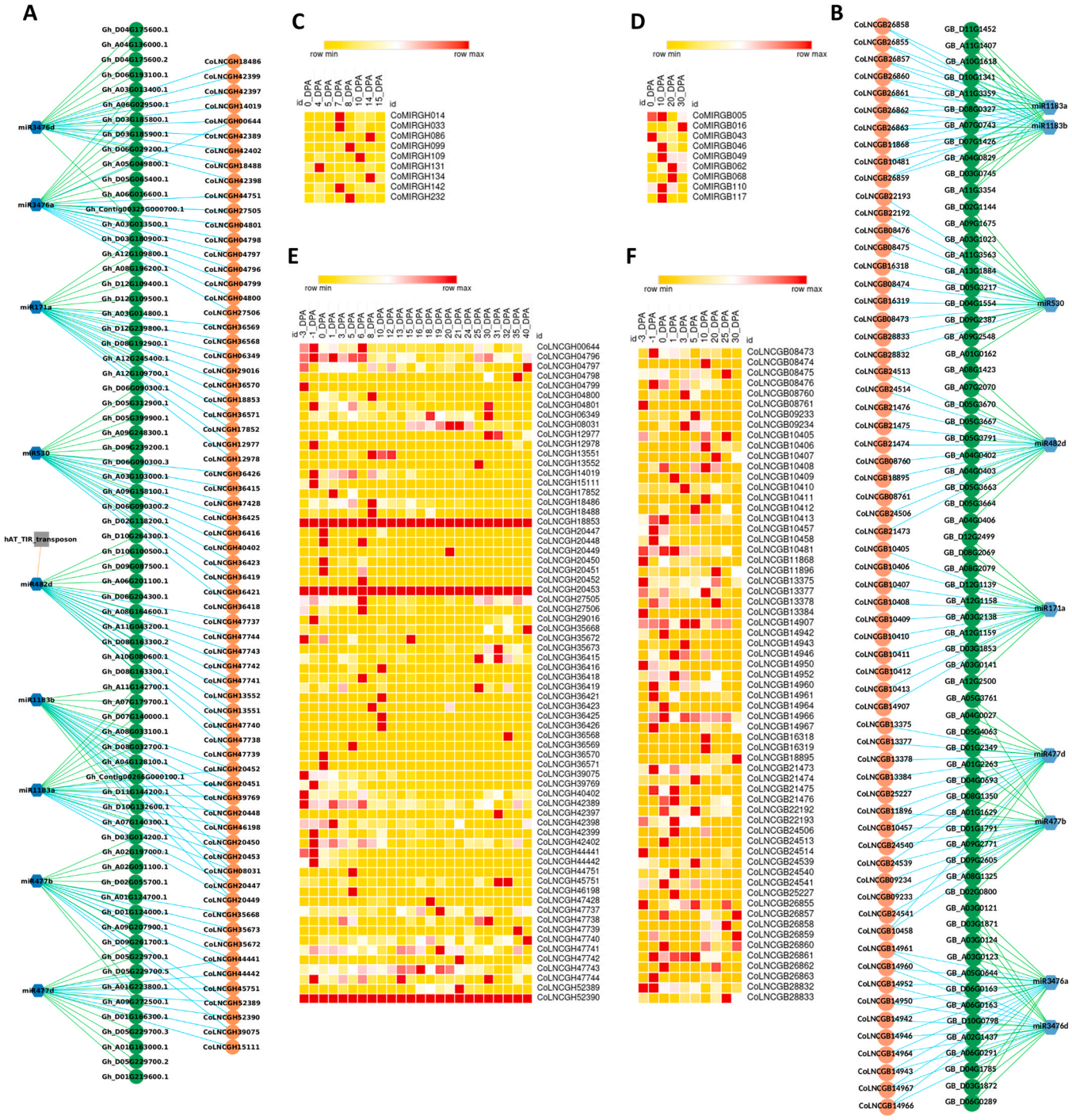

Fig. 9. Network model illustrating lncRNA-miRNA-mRNA interactions during fibre development in (A) *G. hirsutum* (left) and (B) *G. barbadense* (right), along with a heat map summarizing expression data for miRNAs (panels C and D) and lncRNAs (panels E and F) across samples available in CoNCRAtlas. Blue hexagons represent miRNAs, green circles represent mRNAs, orange circles represent lncRNAs, and gray squares represent transposons. Expression values for miRNAs and lncRNAs are in FPKM and RPM, respectively. The stages of fibre development, including initiation, elongation, secondary wall synthesis, and maturation, are shown over developmental time.

provides an overview of ncRNA expression in domesticated cotton, highlighting the unique characteristics of ncRNAs, such as tissue-specific expression and abundance, and their important regulatory role in shaping distinct transcriptional profiles in different cell types. Although lncRNAs are sparsely represented in bulk-tissue RNA sequencing datasets, their regulatory nature and low copy numbers enable easy detection in specific cells [10]. The complexity of lncRNA function is further increased by environmental factors [5,61]. Thus, their restricted expression patterns help shape distinct transcriptional profiles in different cell types, underlining their important regulatory role in gene expression programs.

The integration of tissue-level ncRNA measurements with other omics data holds the potential for inferring the activity of ncRNAs. Previously unknown tissue-specific miRNAs and lncRNAs were identified, and a significant proportion of cotton ncRNA sequences were found to be transposable elements (TEs) with regulatory functions in ncRNAs associated with TEs. Additionally, extensive miRNA-lncRNA-mRNA regulatory networks in cotton were uncovered, suggesting competition between miRNAs and lncRNAs for

the same mRNA targets, miRNA synthesis from lncRNAs, and lncRNAs functioning as miRNA decoys. These findings provide a valuable resource for future research in cotton genomics and can be applied to other plant species.

The primary objective of our study was to identify miRNAs and lncRNAs in cotton and develop a comprehensive resource for them, integrated into the CoNCRAtlas web interface. Compared to existing databases, our resource includes a broader range of ncRNA features, such as expression profiles, lncRNA-miRNA interactions, lncRNA as potential miRNA precursors, and transposable element (TE)-related ncRNAs (Table S6). The findings also highlight the significant involvement of TEs in the diversification, regulation, and potential function of lncRNAs and miRNAs, suggesting their role in the emergence of lncRNAs associated with TEs in plants. Instructions on how to use the CoNCRAtlas web-interface are provided in the results section, and a comparison with other databases is presented in Supplementary Table S6. The web interface is regularly updated with new discoveries and feature expression analyses, functional features, lncRNA-protein coding gene associations, and validated ncRNA-phenotype associations, making it a valuable tool for cotton research communities worldwide. Researchers are encouraged to provide suggestions for creating an up-to-date and comprehensive cotton ncRNA database. Furthermore, including miRNA-lncRNA interactions can help explain the function and intricate interplay of these components in the cotton genome. Overall, the study provides a valuable resource for future research in cotton genomics and can be extended to other plant species, enhancing our understanding of the molecular mechanisms underlying plant growth and development and offering opportunities for crop improvement.

Funding

This study was financially supported by the core grant of the National Institute of Plant Genome Research (NIPGR) in the laboratory of SK.

CRediT authorship contribution statement

Ajeet Singh and Vivek AT conceived and designed the experiment, carried out all aspects of the analysis, generated figures, and wrote the manuscript. Ajeet Singh and Vivek AT contributed to database curation, miRNA and lncRNA identification, expression, WGCNA and plot analyses, figure generation, and manuscript writing. Shruti Sharma and Kanika Gupta contributed to database curation and analysis. Shailesh Kumar supervised the project and reviewed and edited the manuscript. The final manuscript was read and approved by all authors.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

A.S. and A.T.V. gratefully acknowledge the Council of Scientific and Industrial Research and the Department of Biotechnology, respectively, for providing research fellowships. We would also like to thank Dr. Citu for her invaluable assistance with the WGCNA analysis. The authors extend their gratitude to the DBT e-Library Consortium (DeLCON) for providing access to e-material and to the Computational Biology & Bioinformatics Facility (CBBF) of the National Institute of Plant Genome Research (NIPGR) for their support. SK acknowledges the BT/PR40146/BTIS/137/4/2020 project grant from the Department of Biotechnology (DBT), Government of India.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.05.011.

References

[1] Yu Y., Zhang Y., Chen X., Chen Y. Plant Noncoding RNAs: Hidden Players in Development and Stress Responses. Https://DoiOrg/101146/Annurev-Cellbio-100818–125218 2019;35:407–431. https://doi.org/10.1146/ANNUREV-CELLBIO-100818–125218.
[2] Vivek AT, Kumar S. Computational methods for annotation of plant regulatory non-coding RNAs using RNA-seq. Brief Bioinform 2021;22:1–24. https://doi.org/10.1093/BIB/BBAA322
[3] Palazzo AF, Lee ES. Non-coding RNA: What is functional and what is junk? Front Genet 2015;5:2. https://doi.org/10.3389/FGENE.2015.00002/BIBTEX
[4] Axtell MJ, Meyers BC. Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. Plant Cell 2018;30:272–84. https://doi.org/10.1105/TPC.17.00851
[5] Wierzbicki A.T., Blevins T., Swiezewski S. Long Noncoding RNAs in Plants. Https://DoiOrg/101146/Annurev-Arplant-093020–035446 2021;72:245–271. https://doi.org/10.1146/ANNUREV-ARPLANT-093020–035446.
[6] Derks KWJ, Misovic B, van den Hout MCGN, Kockx CEM, Gomez CP, Brouwer RWW, et al. Deciphering the RNA landscape by RNAome sequencing. RNA Biol 2015;12:30–42. https://doi.org/10.1080/15476286.2015.1017202
[7] Budak H, Kaya SB, Cagirici HB. Long non-coding RNA in plants in the era of reference sequences. Front Plant Sci 2020;11:276. https://doi.org/10.3389/FPLS.2020.00276/BIBTEX
[8] 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
[9] Lu X, Chen X, Mu M, Wang J, Wang X, Wang D, et al. Genome-Wide Analysis of Long Noncoding RNAs and Their Responses to Drought Stress in Cotton (Gossypium hirsutum L.). PLoS One 2016;11:e0156723. https://doi.org/10.1371/JOURNAL.PONE.0156723
[10] 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
[11] Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, et al. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). N Phytol 2015;207:1181–97. https://doi.org/10.1111/NPH.13429
[12] Salih H, Gong W, He S, Xia W, Odongo MR, Du X. Long non-coding RNAs and their potential functions in Ligon-lintless-1 mutant cotton during fiber development. BMC Genom 2019;20:1–16. https://doi.org/10.1186/S12864-019-5978-5/FIGURES/6
[13] Wang L, Han J, Lu K, Li M, Gao M, Cao Z, et al. Functional examination of lncRNAs in allotetraploid Gossypium hirsutum. BMC Genom 2021;22:1–13. https://doi.org/10.1186/S12864-021-07771-3/FIGURES/5
[14] Gallart AP, Pulido AH, De Lagrán IAM, Sanseverino W, Cigliano RA. GREENC: a Wiki-based database of plant lncRNAs. Nucleic Acids Res 2016;44:D1161–6. https://doi.org/10.1093/NAR/GKV1215
[15] Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I. CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. Methods Mol Biol 2019;1933:415–29. https://doi.org/10.1007/978-1-4939-9045-0_26/COVER
[16] 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. D1489–95 Nucleic Acids Res 2021;49. https://doi.org/10.1093/NAR/GKAA910
[17] Guo Z, Kuang Z, Wang Y, Zhao Y, Tao Y, Cheng C, et al. PmiREN: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Res 2020;48:D1114–21. https://doi.org/10.1093/NAR/GKZ894
[18] Fei Y, Wang R, Li H, Liu S, Zhang H, Huang J. DPMIND: degradome-based plant miRNA–target interaction and network database. Bioinformatics 2018;34:1618–20. https://doi.org/10.1093/BIOINFORMATICS/BTX824
[19] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47:D155–62. https://doi.org/10.1093/NAR/GKY1141
[20] Chen C, Li J, Feng J, Liu B, Feng L, Yu X, et al. sRNAanno—a database repository of uniformly annotated small RNAs in plants. Hortic Res 2021 81 2021;8:1–8. https://doi.org/10.1038/s41438-021-00480-8
[21] Xu Y., Zhang T., Li Y., Miao Z. Integrated Analysis of Large-Scale Omics Data Revealed Relationship Between Tissue Speci fi city and Evolutionary Dynamics of Small RNAs in Maize (Zea mays), 2020;11:1–16. https://doi.org/10.3389/fgene.2020.00051.
[22] 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
[23] Volders PJ, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res 2019;47:D135–9. https://doi.org/10.1093/NAR/GKY1031
[24] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90. https://doi.org/10.1093/BIOINFORMATICS/BTY560

[25] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019 378 2019;37:907–15. https://doi.org/10.1038/s41587-019-0201-4
[26] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;333(33):290–5. https://doi.org/10.1038/nbt.3122
[27] Yan X, Ma L, Yang MF. Identification and characterization of long non-coding RNA (lncRNA) in the developing seeds of Jatropha curcas. Sci Rep 2020;101(10):1–10. https://doi.org/10.1038/s41598-020-67410-x
[28] Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res 2017;45:W12–6. https://doi.org/10.1093/NAR/GKX428
[29] Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. e166–e166 Nucleic Acids Res 2013;41. https://doi.org/10.1093/NAR/GKT646
[30] Singh U, Khemka N, Rajkumar MS, Garg R, Jain M. PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. e183–e183 Nucleic Acids Res 2017;45. https://doi.org/10.1093/NAR/GKX866
[31] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2021;49:D412–9. https://doi.org/10.1093/NAR/GKAA913
[32] 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. e57–e57 Nucleic Acids Res 2017;45. https://doi.org/10.1093/NAR/GKW1306
[33] 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
[34] Axtell MJ. ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA 2013;19:740–51. https://doi.org/10.1261/RNA.035279.112
[35] Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 2005;21:650–9. https://doi.org/10.1093/BIOINFORMATICS/BTI042
[36] 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
[37] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;32:2847–9. https://doi.org/10.1093/BIOINFORMATICS/BTW313
[38] Dai X, Zhuang Z, Zhao PX. psRNATarget: a plant small RNA target analysis server (2017 release. Nucleic Acids Res 2018;46:W49–54. https://doi.org/10.1093/NAR/GKY316
[39] 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/SUPPINFO
[40] Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 2019;20:1–18. https://doi.org/10.1186/S13059-019-1905-Y/FIGURES/6
[41] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinforma 2009;10:1–9. https://doi.org/10.1186/1471-2105-10-421/FIGURES/4
[42] Lorenz R., Bernhart S.H., Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol, 2011;6:1–14. https://doi.org/10.1186/1748–7188-6–26/TABLES/2.
[43] Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF-FINDER: a vector for high-throughput gene identification. Gene 2002;282:33–41. https://doi.org/10.1016/S0378-1119(01)00819-8
[44] Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinforma 2008;9:1–13. https://doi.org/10.1186/1471-2105-9-559/FIGURES/4
[45] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 2003;13:2498–504. https://doi.org/10.1101/GR.1239303
[46] Johnsson P, Lipovich L, Grandér D, Morris KV. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta - Gen Subj 2014;1840:1063–71. https://doi.org/10.1016/J.BBAGEN.2013.10.035
[47] Zhang X, Dong J, Deng F, Wang W, Cheng Y, Song L, et al. The long non-coding RNA lncRNA973 is involved in cotton response to salt stress. BMC Plant Biol 2019;19:459. https://doi.org/10.1186/S12870-019-2088-0/FIGURES/7
[48] Pedro DLF, Lorenzetti APR, Domingues DS, Paschoal AR. PlaNC-TE: a comprehensive knowledgebase of non-coding RNAs and transposable elements in plants. Database 2018:2018. https://doi.org/10.1093/DATABASE/BAY078
[49] Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, 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
[50] Cho J. Transposon-derived non-coding RNAs and their function in plants. Front Plant Sci 2018;9:600. https://doi.org/10.3389/FPLS.2018.00600/BIBTEX
[51] Zhao T, Tao X, Feng S, Wang L, Hong H, Ma W, et al. LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. Genome Biol 2018;19:1–17. https://doi.org/10.1186/S13059-018-1574-2/FIGURES/8
[52] Yin Z, Li Y, Yu J, Liu Y, Li C, Han X, et al. Difference in miRNA expression profiles between two cotton cultivars with distinct salt sensitivity. Mol Biol Rep 2012;39:4961–70. https://doi.org/10.1007/S11033-011-1292-2
[53] Zhang Z, Xu Y, Yang F, Xiao B, Li G. RiceLncPedia: a comprehensive database of rice long non-coding RNAs. Plant Biotechnol J 2021;19:1492–4. https://doi.org/10.1111/PBI.13639
[54] Huang D, Feurtado JA, Smith MA, Flatman LK, Koh C, Cutler AJ. Long noncoding miRNA gene represses wheat β-diketone waxes. Proc Natl Acad Sci USA 2017;114:E3149–58. https://doi.org/10.1073/PNAS.1617483114/SUPPL_FILE/PNAS.1617483114.SD05.XLSX
[55] Xu XW, Zhou XH, Wang RR, Peng WL, An Y, Chen LL. Functional analysis of long intergenic non-coding RNAs in phosphate-starved rice using competing endogenous RNA network. Sci Rep 2016;61(6):1–12. https://doi.org/10.1038/srep20715.
[56] Yuan J, Zhang Y, Dong J, Sun Y, Lim BL, Liu D, et al. Systematic characterization of novel lncRNAs responding to phosphate starvation in Arabidopsis thaliana. BMC Genom 2016;17:1–16. https://doi.org/10.1186/S12864-016-2929-2/FIGURES/6
[57] Meng X, Li A, Yu B, Li S. Interplay between miRNAs and lncRNAs: Mode of action and biological roles in plant development and stress adaptation. Comput Struct Biotechnol J 2021;19:2567–74. https://doi.org/10.1016/J.CSBJ.2021.04.062
[58] Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Rep 2015;11:1110–22. https://doi.org/10.1016/J.CELREP.2015.04.023
[59] Chen L, Shen E, Zhao Y, Wang H, Wilson I, Zhu QH. The Conservation of Long Intergenic Non-Coding RNAs and Their Response to Verticillium dahliae Infection in Cotton. Int J Mol Sci 2022:23. https://doi.org/10.3390/IJMS23158594/S1
[60] Deng P, Liu S, Nie X, Weining S, Wu L. Conservation analysis of long non-coding RNAs in plants. Sci China Life Sci 2018;61:190–8. https://doi.org/10.1007/S11427-017-9174-9/METRICS
[61] Jha UC, Nayyar H, Jha R, Khurshid M, Zhou M, Mantri N, et al. Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation. BMC Plant Biol 2020;201(20):1–20. https://doi.org/10.1186/S12870-020-02595-X