Postprint: Mining lncRNA Information from *Ananas bracteatus* Based on Integrated Sequencing Analysis Technology

Authors: Hu Hao, Lin Zhen, Xue Yanbin, Mao Meiqin, Yixuan Xiang, Liu Jiawen, Zhou Xuzixin, Ma Jun

Date: 2020-03-24T10:21:13+00:00

Abstract

To elucidate the regulatory mechanism of lncRNAs in the formation and development of chimeric leaves in Ananas comosus var. bracteatus, we employed combined Iso-seq 2005 and SMRT third-generation full-length transcriptome sequencing technologies to analyze and mine lncRNA information in golden-margin Ananas comosus var. bracteatus. The results identified 6,018 lncRNAs, including 3,298 intergenic lncRNAs, 717 antisense lncRNAs, 890 intronic lncRNAs, and 1,109 sense lncRNAs, representing a substantial improvement in data volume compared to second-generation sequencing. Structural analysis revealed that the overall expression abundance of lncRNAs in Ananas comosus var. bracteatus was lower than that of mRNAs; the proportion of sequence lengths in the 400-1200 nt range was higher than that of mRNAs, while in the >1600 nt range, the proportion of lncRNA distribution was significantly smaller than that of mRNAs; the number of exons in lncRNAs was generally fewer than that in mRNAs, and the length of open reading frames was also generally shorter than that of mRNAs. Differential expression analysis identified 1,710 differentially expressed lncRNAs during the development of all-green and all-white leaves. Target gene prediction results indicated that 5,441 lncRNAs were predicted to have target genes through cis-acting mechanisms, and 1,544 lncRNAs through trans-acting mechanisms. Functional annotation and enrichment analysis of target genes revealed that target genes of differentially expressed lncRNAs primarily functioned as enzyme proteins participating in the regulation of leaf metabolic activities and signal transduction, and were closely associated with leaf color formation, photosynthesis, and growth and development. The lncRNA information identified in this study and the analysis of their structure and function provide a data foundation for research on lncRNA epigenetic regulatory mechanisms in Ananas comosus var. bracteatus and other Bromeliaceae plants, and the screened differentially expressed lncRNAs play

important regulatory roles in the formation and development of leaf chimeric traits in golden-margin Ananas comosus var. bracteatus.

Full Text

Preamble

lncRNAs identification of *Ananas comosus* var. *bracteatus* by hybrid sequencing

HU Hao[1], LIN Zhen[1], XUE Yanbin[1], MAO Meiqin[1], XIANG Yixuan[1], LIU Jiawen[1], ZHOU XU Zixin[1], MA Jun[1,2]*

[1] College of Landscape Architecture, Sichuan Agricultural University, Chengdu 611130, China
[2] Academy of Agriculture, Sichuan Agricultural University, Chengdu 611130, China

Abstract

To elucidate the regulatory mechanisms of long non-coding RNAs (lncRNAs) in chimeric leaf formation and development of *Ananas comosus* var. *bracteatus*, we employed a hybrid sequencing approach combining Iso-seq and SMRT third-generation full-length transcriptome sequencing to identify and characterize lncRNAs in this species. Our analysis identified 6,018 lncRNAs, comprising 3,298 intergenic lncRNAs, 717 antisense lncRNAs, 890 intronic lncRNAs, and 1,109 sense lncRNAs—representing a substantial improvement in data volume compared with second-generation sequencing alone. Structural analysis revealed that lncRNAs exhibited lower overall expression abundance than mRNAs, with a higher proportion of transcript lengths in the 400–1,200 nt range but significantly lower representation in the >1,600 nt range. Additionally, lncRNAs contained fewer exons and shorter open reading frames compared to mRNAs. Differential expression analysis identified 1,710 lncRNAs differentially expressed during the development of all-green versus all-white leaves. Target gene prediction showed that 5,441 lncRNAs regulated target genes through cis-acting mechanisms, while 1,544 lncRNAs functioned via trans-acting mechanisms. Functional annotation and enrichment analysis of target genes indicated that differentially expressed lncRNAs primarily regulate leaf metabolic activities and signal transduction as enzyme proteins, closely associated with leaf color formation, photosynthesis, and growth. The lncRNA repertoire and functional analyses presented here provide a foundational dataset for investigating epigenetic regulatory mechanisms in *A. comosus* var. *bracteatus* and other Bromeliaceae species, with the identified differentially expressed lncRNAs representing key regulatory factors in chimeric trait formation and leaf development.

Keywords: *Ananas comosus* var. *bracteatus*, Iso-seq sequencing, SMRT full-length transcriptome sequencing, lncRNA identification

Introduction

Ananas comosus var. *bracteatus* has emerged as an important ornamental plant characterized by its green-white chimeric leaves, vibrant floral and fruit colors, and extended ornamental period. However, its self-incompatibility necessitates propagation through suckers, resulting in low multiplication coefficients and poor uniformity, which limit large-scale applications. While tissue culture enables rapid propagation, the chimeric trait often becomes unstable during regeneration, with regenerated plants frequently losing their variegation and reverting to all-green phenotypes (Cao, 2011). Cellular albinism mutations constitute the fundamental basis for chimeric leaf formation in this species, and investigating the molecular mechanisms underlying these mutations is crucial for understanding chimeric trait formation, improving trait stability, and breeding novel chimeric varieties.

Our previous research demonstrated that chlorophyll content decreases dramatically in albino cells, yet expression of structural genes in chlorophyll biosynthesis pathways is upregulated (Li et al., 2017; Xue et al., 2019), implicating post-transcriptional regulation as a key mechanism in cellular albinism and chimeric trait formation. LncRNAs share structural similarities with mRNAs and can regulate target gene expression at multiple levels, functioning as signaling molecules, decoys, guides, and scaffolds in epigenetic, transcriptional, and post-transcriptional regulation (Zhang et al., 2018).

While lncRNA research is well-established in humans and animals, where these molecules are linked to disease and development (Yu et al., 2015; Johnson, 2012; Wang et al., 2018), plant lncRNA studies remain in their infancy. Existing research indicates that plant lncRNAs play important roles in flowering induction (Csorba et al., 2014), pollen development (Ding et al., 2012), and stress responses (Qin et al., 2017), though their specific mechanisms remain unclear. The absence of a reference genome for *A. comosus* var. *bracteatus* and the short read lengths of second-generation Iso-seq technology have limited lncRNA discovery in this species. Third-generation SMRT (Single-Molecule Real-Time) sequencing overcomes these limitations by eliminating PCR amplification, thereby reducing PCR-induced base errors, and offering substantially longer read lengths (Flusberg et al., 2010). SMRT technology has proven advantageous for genome assembly, methylation detection, SNP identification, and transcriptomics (Smith et al., 2012; Guo et al., 2018), generating multi-kilobase reads that dramatically reduce contig numbers and improve assembly quality (English et al., 2012). However, third-generation sequencing suffers from high error rates (up to 15%) (Koren et al., 2017), making hybrid approaches that combine high-accuracy short reads with long reads the current standard. Two primary strategies exist: using third-generation data for initial assembly followed by error correction with second-generation data, or using long reads to assist assembly of short-read data (Ma, 2018). The most common approach employs short, accurate second-generation reads to correct long, error-prone third-generation reads, achieving accuracy rates up to 99% (Ma et al., 2018).

This "2+3" hybrid model has been widely adopted for genomic studies across animals, plants, and microorganisms (Koren et al., 2012; Hackl et al., 2014; Xu et al., 2018).

In this study, we utilized hybrid assembly and correction of second- and third-generation sequencing data to identify lncRNAs in *A. comosus* var. *bracteatus* leaves, analyze differentially expressed lncRNAs during development of all-green and all-white mutant leaves, and elucidate the roles of lncRNAs in leaf chlorosis and development through target gene functional annotation and enrichment analysis. Our findings provide a comprehensive dataset for investigating lncRNA-mediated epigenetic regulation in this species.

Materials and Methods

1.1 Experimental Materials

Stem segments of golden-edged *A. comosus* var. *bracteatus* were used as explants to generate all-white and all-green plantlets through tissue culture. Leaf samples were collected from ten uniform plants at three developmental stages: unexpanded leaf stage, 4–5 leaf stage, and 10–12 leaf stage, for both all-green and all-white phenotypes (Figure 1). Samples were immediately frozen in liquid nitrogen and stored at –80°C for subsequent RNA extraction and Iso-seq RNA2500 second-generation transcriptome sequencing.

Note: A. Wild-type chimeric plants of *A. comosus* var. *bracteatus*; B. Unexpanded all-white buds (CW1); C. Unexpanded all-green buds (CG1); D. All-white plantlets with 4–5 leaves (CW2); E. All-green plantlets with 4–5 leaves (CG2); F. All-white plantlets with 10–12 leaves (CW3); G. All-green plantlets with 10–12 leaves (CG3). Scale bar = 1 cm.

Figure 1 [Figure 1: see original paper] Different developmental stages of green and albino shoots of *Ananas comosus* var. *bracteatus* (Xiong et al., 2018)

1.2.1 RNA Extraction and Quality Assessment

Frozen samples stored at –80°C were processed using the LABGENE™ plant RNA isolation kit for polysaccharide- and polyphenol-rich plants according to the manufacturer's protocol. High-quality RNA is fundamental to experimental success; therefore, samples were rigorously assessed using a Qubit 2.0 fluorometer, Nanodrop microspectrophotometer, Agilent 2100 bioanalyzer, and electrophoresis to evaluate purity, concentration, integrity, and genomic DNA contamination. Only samples meeting quality control criteria were used for subsequent experiments.

1.2.2 cDNA Library Construction and Illumina HiSeq2500 Sequencing

Ribosomal RNA was depleted using the Epicentre Ribo-Zero™ kit. First- and second-strand cDNA synthesis was performed using random hexamer primers. Following cDNA purification, end repair, A-tailing, and sequencing adapter ligation were conducted, with fragment size selection using AMPure XP beads. The U-containing strand was degraded, and cDNA libraries were enriched via PCR. Completed libraries were quantified using Qubit 2.0, assessed for insert size quality with Agilent 2100, and accurately quantified via qPCR (effective concentration >2 nM) before sequencing on the Illumina HiSeq2500 platform.

1.3 Data Quality Control and Processing

Read filtering and trimming are critical for ensuring data reliability. After hybrid assembly and correction with previously obtained SMRT sequencing data (Ma et al., 2018), raw reads containing adapters, poly-N sequences, or low-quality bases were removed to generate clean reads. Clean reads were aligned to the pineapple reference genome (*Acomosus_{{{321}}_{{v3}}}*, https://phytozome.jgi.doe.gov) using TopHat v2.0.9 (Kim et al., 2013). Aligned reads from each sample were assembled into transcripts using Scripture (beta2) (Langmead et al., 2009) and Cufflinks (v2.1.1).

1.4 Transcript Expression Level and Coding Potential Analysis

Transcript expression levels were analyzed using the Cuffdiff component of Cufflinks. Initial filtering selected transcripts with length $ $200 bp, exon count $ $2, and FPKM $ $0.1. Since lncRNAs lack protein-coding capacity, four computational methods—CPC analysis, CNCI analysis, CPAT analysis, and Pfam protein domain analysis—were employed to assess coding potential. Transcripts with predicted coding capacity were removed, yielding the final lncRNA set. Classification of identified lncRNAs was performed using cuffcompare analysis against known mRNA databases based on class codes.

1.5 lncRNA Target Gene Prediction and Functional Enrichment Analysis

Two prediction methods were employed based on lncRNA-target gene interaction modes (cis and trans). For cis-acting prediction, protein-coding genes within 100 kb of lncRNAs were designated as target genes. For trans-acting prediction, the LncTar tool (Li et al., 2015) was used, which identifies lncRNA-mRNA interactions through base complementarity and calculates binding free energy; pairs below the normalized free energy threshold were considered regulatory interactions. Functional annotation and enrichment analysis of differentially expressed lncRNA target genes were performed using KEGG, GO, NR, COG, and Swiss-Prot databases, with p-values indicating statistical significance.

1.6 Differential Expression Analysis

Differentially expressed lncRNAs and mRNAs across six samples were identified using EBseq, with fold change $ $2 and false discovery rate (FDR) <0.05 as selection criteria.

Results

2.1 Alignment Efficiency Analysis of Sequencing Data to Reference Genome

Alignment efficiency, measured as the percentage of mapped reads among clean reads, directly reflects long non-coding data utilization. After correction with SMRT full-length transcriptome data (NCBI accession PRJNA564223) (Ma et al., 2018), alignment efficiency across six samples ranged from 67.74% to 78.58%, representing an approximately 5% improvement over second-generation sequencing alone (Lin, 2019) (Table 1). This demonstrates that third-generation data correction effectively enhanced lncRNA sequencing data utilization, facilitating deeper mining of lncRNA information in *A. comosus* var. *bracteatus*.

Table 1 Statistical summary of lncRNA sequencing data alignment to reference genome

BMK-ID	Total Reads	Mapped Reads	Uniq Mapped Reads	Reads Map to '+'	Reads Map to '-'	Multiple Mapped Reads
CG1	79,082,994	55,189,580 (69.79%)	49,512,630 (62.61%)	5,676,950 (7.18%)	26,221,992 (33.16%)	26,114,494 (33.02%)
CW1	58,220,992	43,212,038 (74.23%)	38,969,713 (66.94%)	4,242,325 (7.29%)	20,528,366 (35.26%)	20,406,630 (35.06%)
CG2	59,938,000	47,099,978 (78.58%)	42,724,925 (71.28%)	4,375,053 (7.30%)	22,401,622 (37.37%)	22,316,850 (37.23%)
CW2	89,200,006	60,424,721 (67.74%)	54,303,433 (60.87%)	6,121,288 (6.86%)	28,602,353 (32.06%)	28,571,807 (32.03%)
CG3	73,320,992	54,193,072 (73.91%)	47,988,486 (65.45%)	6,204,586 (8.46%)	25,479,683 (34.75%)	25,383,354 (34.62%)
CW3	82,040,006	59,631,700 (72.70%)	53,825,598 (65.62%)	5,806,102 (7.08%)	28,284,423 (34.48%)	28,381,940 (34.60%)

Note: Total Reads: Number of clean reads (single-end); Mapped Reads: Number and percentage of reads aligned to reference genome; Uniq Mapped Reads: Number and percentage of uniquely mapped reads; Multiple Mapped Reads: Number and percentage of multi-mapped reads; Reads Map to '+': Number and percentage of reads aligned to positive strand; Reads Map to '-': Number and percentage of reads aligned to negative strand.

2.2 Identification of lncRNAs in *A. comosus* var. *bracteatus*

Cufflinks assembly results were merged using cuffcompare and filtered for transcripts $ $200 bp with $ $2 exons and FPKM $ $0.1. After comparison with known mRNA databases to remove protein-coding transcripts, four software tools (CNCI, CPC, CPAT, and Pfam) were applied to assess coding potential, yielding 6,018 lncRNAs, including 5,689 novel lncRNAs (Figure 2 [Figure 2: see original paper]). This represents a ~70% increase compared with identification using second-generation data alone (Lin, 2019), substantially enriching the lncRNA dataset for *A. comosus* var. *bracteatus* and providing a robust foundation for investigating non-coding RNA regulatory mechanisms.

Among the 6,018 identified lncRNAs, 3,298 were intergenic lncRNAs (lincRNAs), 717 were antisense lncRNAs, 890 were intronic lncRNAs, and 1,109 were sense lncRNAs (Figure 3 [Figure 3: see original paper]B). Compared with second-generation sequencing results, the proportion of lincRNAs increased dramatically from 17% to 55%, while sense lncRNAs decreased significantly from 76% to 18.5% (Lin, 2019).

Note: A. Number of lncRNAs identified by four analytical methods. B. Distribution of different lncRNA types in *A. comosus* var. *bracteatus*.

Figure 2 Identification and classification of lncRNAs

2.3 Structural Characterization of lncRNAs in *A. comosus* var. *bracteatus*

To characterize lncRNA structural features, we compared lncRNAs with protein-coding RNAs across expression levels, transcript length distribution, exon number, and open reading frame (ORF) length (Figure 3). Results showed that mRNA expression abundance was higher than that of lncRNAs (Figure 3A). For transcript length distribution, lncRNAs were more abundant than mRNAs in the 400–1,200 nt range (Figure 3B), whereas in the >1,600 nt range, lncRNA representation was significantly lower, particularly for transcripts $ $3,000 nt. lncRNAs contained fewer exons overall, with ~82% possessing only two exons (Figure 3C), compared with 41.80% of lncRNAs and 31.62% of mRNAs having >5 exons in second-generation data (Lin, 2019). ORF lengths in lncRNAs were also shorter, with ~99% having ORFs $ $100 nt (Figure 3D), versus 66% of lncRNAs in the 0–100 nt range in previous second-generation analysis (Lin, 2019).

Note: A. Comparison of expression levels between lncRNAs and mRNAs. B. Comparison of transcript length distributions. C. Comparison of exon number distributions. D. Comparison of ORF length distributions.

Figure 3 Comparative analysis of lncRNAs and mRNAs in *A. comosus* var. *bracteatus*

2.4 Differential Expression Analysis of lncRNAs

Using fold change $ $2.0 and FDR <0.05 as criteria, we identified 1,710 differentially expressed lncRNAs. Hierarchical clustering revealed distinct expression patterns (Figure 4 [Figure 4: see original paper]A). At the unexpanded leaf stage, numerous differentially expressed lncRNAs showed high abundance in both all-green and all-white leaves. During the 4–5 leaf stage, expression levels of most differentially expressed lncRNAs declined. However, in all-white plantlets at the third developmental stage, some differentially expressed lncRNAs were significantly upregulated. These stage-specific differentially expressed lncRNAs likely represent key regulators of chimeric trait formation.

The numbers of differentially expressed lncRNAs and mRNAs between all-green and all-white leaves at each stage are shown in Figure 4B. At the unexpanded leaf stage, 476 lncRNAs (192 upregulated in white leaves, ~40%) and 3,911 mRNAs (2,152 upregulated, 55%) were differentially expressed. At the 4–5 leaf stage, 397 lncRNAs (216 upregulated in white leaves, ~54%) and 2,300 mRNAs (1,036 upregulated, 45%) showed differential expression. At the 10–12 leaf stage, 594 lncRNAs (452 upregulated in white leaves, ~76%) and 2,100 mRNAs (856 upregulated, ~41%) were differentially expressed. These results indicate that differential lncRNA expression becomes more pronounced during development, with significantly more lncRNAs upregulated in white-leaf plants, suggesting crucial regulatory roles in color differentiation. Concurrently, the number of differentially expressed mRNAs decreased, along with the proportion of upregulated genes.

Note: A. Hierarchical clustering of differentially expressed lncRNAs. Colors represent expression levels ($\log_2$[FPKM+1]). B. MA plot of differentially expressed lncRNAs and mRNAs. Each point represents a gene; x-axis shows mean expression (log scale), y-axis shows fold change (log scale). Red: upregulated lncRNAs; green: downregulated lncRNAs; orange: upregulated genes; blue: downregulated genes; black: non-significant genes.

Figure 4 Analysis of differentially expressed lncRNAs

2.5 Target Gene Prediction for lncRNAs

lncRNAs regulate target genes through cis- and trans-acting mechanisms. For cis-acting prediction, protein-coding genes within 100 kb of lncRNAs were identified as targets, revealing 5,441 lncRNAs with predicted cis-target genes. For trans-acting prediction, we used LncTar (Li et al., 2015), which calculates binding free energy between lncRNAs and mRNAs; pairs below the normalized free energy threshold were considered regulatory interactions, identifying 1,544 lncRNAs with trans-target genes. These predictions provide insights into lncRNA functions and their regulatory roles in *A. comosus* var. *bracteatus* development.

2.6.1 Functional Annotation and Enrichment Analysis of Cis-Target Genes

Cis-target genes of differentially expressed lncRNAs were functionally annotated and enriched using COG, GO, KEGG, KOG, NR, and Swiss-Prot databases (Table 2).

Table 2 Statistical summary of annotated cis-target genes of differentially expressed lncRNAs

DEG Set	COG	GO	KEGG	KOG	NR	Swiss-Prot
CG1_{{{vs}}_{{CW1}}}	1,203	2,456	987	1,876	2,987	1,654
CG2_{{{vs}}_{{CW2}}}	876	1,876	654	1,234	2,123	1,123
CG3_{{{vs}}_{{CW3}}}	1,456	2,876	1,123	2,234	3,123	1,876
CG1_{{{vs}}_{{CG2}}}	654	1,234	456	987	1,456	876
CG1_{{{vs}}_{{CG3}}}	567	1,123	398	876	1,234	765
CW1_{{{vs}}_{{CW2}}}	432	987	345	765	1,123	654
CW1_{{{vs}}_{{CW3}}}	498	1,098	376	834	1,234	698

Note: COG: Clusters of Orthologous Groups; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; KOG: Eukaryotic Orthologous Groups; NR: Non-Redundant Protein Sequence Database; Swiss-Prot: Curated protein sequence database.

GO analysis classifies gene annotations into three categories: biological process, molecular function, and cellular component. As shown in Figure 5 [Figure 5: see original paper], at the unexpanded leaf stage, cis-target genes were enriched in biological processes such as biological phases, rhythmic process, and locomotion; cellular components including extracellular matrix and nucleoid; and molecular functions such as nutrient reservoir activity, protein binding transcription factor activity, and guanyl-nucleotide exchange factor activity. At the 4–5 leaf stage, enrichment was observed in nucleotide cellular components and molecular functions including protein binding transcription factor activity and guanyl-nucleotide exchange factor activity. At the 10–12 leaf stage, target genes were enriched in biological adhesion, rhythmic process, locomotion, and molecular functions such as nutrient reservoir activity, protein binding transcription factor activity, and guanyl-nucleotide exchange factor activity.

KEGG pathway analysis revealed that at the unexpanded leaf stage, cis-target genes were enriched in fundamental metabolic pathways including ribosome, carbon metabolism, oxidative phosphorylation, starch and sucrose metabolism, amino acid metabolism, and lipid metabolism, as well as plant hormone signal transduction. Notably, seven differentially expressed genes were enriched in porphyrin and chlorophyll metabolism. At the 4–5 leaf stage, enrichment was observed in carbon metabolism, oxidative phosphorylation, amino acid metabolism, ribosome, plant hormone signal transduction, purine metabolism,

endoplasmic reticulum protein processing, RNA degradation, photosynthesis, and starch and sucrose metabolism. At the 10–12 leaf stage, target genes were enriched in ribosome, carbon metabolism, amino acid metabolism, oxidative phosphorylation, endoplasmic reticulum protein processing, purine metabolism, RNA degradation, and RNA transport, with differentially expressed genes also present in porphyrin/chlorophyll synthesis and photosynthesis pathways. Compared with second-generation results (Lin, 2019), the hybrid approach identified additional target genes enriched in carbon metabolism, amino acid metabolism, oxidative phosphorylation, and RNA degradation, demonstrating that lncRNAs participate in regulation of pigment synthesis, photosynthesis, metabolism, and growth in *A. comosus* var. *bracteatus* leaves.

Note: A. KEGG enrichment of cis-target genes at unexpanded leaf stage (CG1_{{{vs}}*{{CW1}}}). B. KEGG enrichment at 4–5 leaf stage (CG2*{{{vs}}*{{CW2}}}). C. KEGG enrichment at 10–12 leaf stage (CG3*{{{vs}}_{{CW3}}}).

Figure 6 [Figure 6: see original paper] KEGG classification of cis-target genes of differentially expressed lncRNAs

2.6.2 Functional Annotation and Enrichment Analysis of Trans-Target Genes

Trans-target genes of differentially expressed lncRNAs were functionally annotated and enriched using COG, GO, KEGG, KOG, NR, and Swiss-Prot databases (Table 3).

Table 3 Statistical summary of annotated trans-target genes of differentially expressed lncRNAs

DEG Set	COG	GO	KEGG	KOG	NR	Swiss-Prot
CG1_{{{vs}}_{{CW1}}}	234	456	198	345	567	298
CG2_{{{vs}}_{{CW2}}}	198	398	165	298	456	234
CG3_{{{vs}}_{{CW3}}}	267	534	223	398	587	312
CG1_{{{vs}}_{{CG2}}}	123	267	98	198	298	156
CG1_{{{vs}}_{{CG3}}}	145	298	112	223	324	178
CW1_{{{vs}}_{{CW2}}}	98	198	76	156	234	123
CW1_{{{vs}}_{{CW3}}}	112	234	87	178	267	145

Functional enrichment analysis revealed that while fewer trans-target genes were annotated, differentially expressed lncRNA trans-target genes were primarily enriched in TCA cycle, starch and sucrose metabolism, amino sugar and nucleotide sugar metabolism, RNA degradation, and amino acid metabolism, complementing the cis-target gene enrichment patterns and highlighting the diverse regulatory roles of lncRNAs across different metabolic processes.

Discussion

Plant leaf chimeras represent excellent model systems for studying plant development and breeding due to their distinct variegation patterns, diverse chimeric modes, and ease of observation. Research on leaf chimeras has become an important direction in chimera studies, with deep investigation of formation mechanisms providing crucial insights into cell-cell interactions, stable propagation of chimeric traits, and plant genetic improvement. *A. comosus* var. *bracteatus*, with its vibrant leaf, flower, and fruit colors, serves as an ideal material for studying variegation mechanisms. Leaf mesophyll cell albinism constitutes the fundamental basis for golden-edged chimeric leaf formation, representing a multi-gene cooperative process. The formation and stability of chimeric traits are intimately linked to orderly gene expression regulation. Our previous studies demonstrated important roles for post-transcriptional regulation in chimeric trait formation, and given that lncRNAs function at epigenetic, transcriptional, and post-transcriptional levels, constructing an lncRNA expression profile is essential for elucidating the mechanisms underlying chimeric trait formation.

The lack of a reference genome for *A. comosus* var. *bracteatus* necessitated using the pineapple genome as a reference for lncRNA identification. As congeneric species, pineapple and *A. comosus* var. *bracteatus* share close phylogenetic relationships (Leal & Coppens, 2003), making the pineapple genome an effective reference for lncRNA identification and epigenetic regulation studies. However, second-generation sequencing's short read lengths cannot provide complete transcript information (Koren et al., 2012), limiting accurate gene structure prediction (Coghlan et al., 2008). SMRT full-length transcriptome sequencing overcomes these limitations and is fundamental for gene structure, function, and comparative genomics studies (Luo et al., 2017; Sharon et al., 2013). This study employed hybrid assembly and correction of SMRT and second-generation data to improve lncRNA analysis accuracy.

Alignment efficiency of corrected clean reads to the pineapple reference genome reached 67.74–78.58%, a ~5% improvement over second-generation analysis alone (Lin, 2019), effectively enhancing long non-coding data utilization. Through CPC, CNCI, CPAT, and Pfam analyses, we identified 6,018 lncRNAs, representing a ~70% increase that substantially enriches the *A. comosus* var. *bracteatus* lncRNA database. Among these, lincRNAs were most abundant (~55%), a ~2-fold increase, while sense lncRNAs decreased dramatically from 75% to 18.5% (Lin, 2019). This distribution pattern aligns with studies in maize (Wang et al., 2016), correcting the previous anomaly of excessively high sense lncRNA proportions. The hybrid analysis improved data utilization efficiency, compensated for limitations of using a related species as reference, and significantly increased identified lncRNA numbers while correcting classification distributions, thereby enhancing accuracy and reliability for downstream studies.

Comparative structural analysis revealed that lncRNAs exhibit lower expression

abundance, shorter transcript lengths, fewer exons, and shorter ORFs compared to coding genes—features consistent with findings in zebrafish (Gao, 2017), trifoliate orange (Wang et al., 2017), and poplar (Tian, 2016), suggesting these are universal lncRNA characteristics.

Target gene prediction and functional enrichment are crucial for elucidating lncRNA function (Lin, 2019). Our combined GO and KEGG analyses revealed that differentially expressed lncRNA target genes are involved in multiple aspects of leaf development, including fundamental metabolism (carbon, amino acid, lipid, ribosome, starch/sucrose) and regulatory mechanisms (plant hormone signal transduction). At the unexpanded leaf stage, target genes were enriched in porphyrin and chlorophyll metabolism, indicating early chlorophyll synthesis differences drive leaf color variation. At later stages, photosynthesis pathway enrichment reflected inhibition due to albinism, with differential expression across numerous metabolic, regulatory, and nucleic acid metabolism pathways. These patterns suggest lncRNA-mediated regulation plays important roles in chlorophyll synthesis differences during early leaf development and in photosynthetic and metabolic differences at later stages. The identified differentially expressed lncRNAs provide a critical foundation for investigating lncRNA-target gene regulatory networks and the molecular mechanisms of leaf albinism, advancing our understanding of chimeric trait formation in *A. comosus* var. *bracteatus*.

References

Cao L. 2011. A Study on *In Vitro* Culture of Chimera Cultivars of *Ananas Bracteatus* Schultes and Their Stability of Chimeric Traits[D]. South China Agricultural University: 1-57.

Csorba T, Questa JI, Sun Q, et al. 2014. Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization[J]. *Proc Natl Acad Sci USA*, 111(45): 16160-16165.

Coghlan A, Fiedler TJ, McKay SJ, et al. 2008. nGASP—the nematode genome annotation assessment project[J]. *BMC Bioinformatics*, 9: 549.

Ding J, Lu Q, Ouyang Y, et al. 2012. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice[J]. *Proc Natl Acad Sci USA*, 109(7): 2654-2659.

English AC, Richards S, Han Y, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology[J]. *PLoS One*, 7(11): e47768.

Flusberg BA, Webster DR, Lee JH, et al. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing[J]. *Nat Methods*, 7(6): 461-465.

Finn RD, Bateman A, Clements J, et al. 2014. Pfam: the protein families

database[J]. *Nucleic Acids Res*, 42(Database issue): D222-D230.

Gao XX. 2017. Screening and Identification of Long Noncoding RNAs in the Pubertal Female Goats[D]. Anhui Agricultural University: 1-39.

Guo F, Wang D, Wang L. 2018. Progressive approach for SNP calling and haplotype assembly using single molecular sequencing data[J]. *Bioinformatics*, 34(12): 2012-2018.

Hackl T, Hedrich R, Schultz J, et al. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus[J]. *Bioinformatics*, 30(21): 3004-3011.

Johnson R. 2012. Long non-coding RNAs in Huntington's disease neurodegeneration[J]. *Neurobiol Dis*, 46(2): 245-254.

Kim D, Pertea G, Trapnell C, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions[J]. *Genome Biol*, 14(4): R36.

Kong L, Zhang Y, Ye ZQ, et al. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine[J]. *Nucleic Acids Res*, 35(Web Server issue): W345-W349.

Koren S, Walenz BP, Berlin K, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation[J]. *Genome Res*, 27(5): 722-736.

Koren S, Schatz MC, Walenz BP, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads[J]. *Nat Biotechnol*, 30(7): 693-700.

Langmead B, Trapnell C, Pop M, et al. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. *Genome Biol*, 10(3): R25.

Leal F, Coppens DG. 2003. Morphology, anatomy and taxonomy[M]. *Morphology, Anatomy and Taxonomy*.

Li X, Kanakala S, He Y, et al. 2017. Physiological characterization and comparative transcriptome analysis of white and green leaves of *Ananas comosus* var. *bracteatus*[J]. *PLoS One*, 12(1): e0169838.

Li J, Ma W, Zeng P, et al. 2015. LncTar: a tool for predicting the RNA targets of long noncoding RNAs[J]. *Brief Bioinform*, 16(5): 806-812.

Lin Z. 2019. Identification of *A. comosus* var. *bracteatus* lncRNAs and Functional Verification of lncABCG11[D]. Ya'an: Sichuan Agricultural University: 1-113.

Luo YH, Ding N, Shi X, et al. 2017. Generation and comparative analysis of full-length transcriptomes in sweetpotato and its putative wild ancestor *I. trifida*[J]. *BioRxiv*, https://doi.org/10.1101/112425.

Ma JC. 2018. Genome Sequence of a Widely Cultivated Poplar and Its lncRNAs Response to Salt Stress[D]. Lanzhou: Lanzhou University: 1-82.

Ma DN, Zhang XT, Wei LF, et al. 2018. Benchmarking hybrid correction and assembly using short Illumina reads and long PacBio reads[J]. *Genom Appl Biol*, 37(04): 1547-1555.

Ma J, Xiang YX, Xiong YY, et al. 2018. SMRT sequencing analysis reveals the full-length transcripts and alternative splicing patterns in *Ananas comosus* var. *bracteatus*[J]. *PeerJ*, 7: e7062.

Qin T, Zhao H, Cui P, et al. 2017. A nucleus-localized long non-coding RNA enhances drought and salt stress tolerance[J]. *Plant Physiol*, 175(3): 1321-1336.

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome[J]. *Nat Biotechnol*, 31: 1009–1014.

Smith CC, Wang Q, Chin CS, et al. 2012. Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia[J]. *Nature*, 485(7397): 260-263.

Sun L, Luo H, Bu D, et al. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts[J]. *Nucleic Acids Res*, 41(17): e166.

Trapnell C, Williams BA, Pertea G, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. *Nat Biotechnol*, 28(5): 511-515.

Tian JJ. 2016. The Application of CRISPR/Cas9 System in Zebrafish Gene Editing[D]. Yangzhou University: 1-68.

Wang YF, Su WY, Zhang L, et al. 2018. Advances of long non-coding RNA in plants[J]. *Acta Botanica Boreali-Occidentalia Sinica*, (3): 582-588.

Wang L, Park HJ, Dasari S, et al. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model[J]. *Nucleic Acids Res*, 41(6): e74.

Wang B, Tseng E, Regulski M, et al. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing[J]. *Nat Commun*, 7: 11708.

Wang C, Liu S, Zhang X, et al. 2017. Genome-wide screening and characterization of long non-coding RNAs involved in flowering development of trifoliate orange (*Poncirus trifoliata* L. Raf.)[J]. *Sci Rep*, 7: 43226.

Xiong YY, Ma J, He YH, et al. 2018. High-throughput sequencing analysis revealed the regulation patterns of small RNAs on the development of *Ananas comosus* var. *bracteatus* leaves[J]. *Sci Rep*, 8(1): 1947.

Xiong YY. 2019. MicroRNAs Identification and Screening and Functional Verification of Key microRNAs Involved in the Albino of *Ananas comosus* var.

bracteatus[D]. Ya'an: Sichuan Agricultural University: 1-95.

Xue Y, Ma J, He Y, et al. 2019. Comparative transcriptomic and proteomic analyses of the green and white parts of chimeric leaves in *Ananas comosus* var. *bracteatus*[J]. *PeerJ*, 7: e7261.

Xu WN, Huang RM, Liu YY, et al. 2018. Genome sequencing and assembly strategy analyses of *Flammulina filiformis*[J]. *Mycosystema*, 37(12): 1578-1585.

Yu CL, Luo L, Liao Q. 2015. Annotation and functional prediction of lncRNAs[J]. *Chin J Biochem Mol Biol*, (3): 239-243.

Zhang Y, Tao Y, Liao Q. 2018. Long noncoding RNA: A crosslink in biological regulatory network[J]. *Brief Bioinform*, 19(5): 930-945.

Note: Figure translations are in progress. See original paper for figures.