the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2024) **120,** 799–817 doi: 10.1111/tpj.17013

RESOURCE

A multispecies study reveals the diversity and potential regulatory role of long noncoding RNAs in cucurbits

Pascual Villalba-Bermell, Joan Marquez-Molins[†] and Gustavo Gomez[*]

Institute for Integrative Systems Biology (I2SysBio), Consejo Superior de Investigaciones Científicas (CSIC) - Universitat de València (UV), Parc Científic, Cat. Agustín Escardino 9, 46980 Paterna, Spain

Received 2 April 2024; revised 31 July 2024; accepted 23 August 2024; published online 10 September 2024.
*For correspondence (e-mail gustavo.gomez@csic.es)
[†]Present address: Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences and Linnean Center for Plant Biology, Uppsala, Sweden

SUMMARY

Plant long noncoding RNAs (lncRNAs) exhibit features such as tissue-specific expression, spatiotemporal regulation, and stress responsiveness. Although diverse studies support the regulatory role of lncRNAs in model plants, our knowledge about lncRNAs in crops is limited. We employ a custom pipeline on a dataset of over 1000 RNA-seq samples across nine representative species of the family Cucurbitaceae to predict 91 209 nonredundant lncRNAs. The lncRNAs were characterized according to three confidence levels and classified by their genomic context into intergenic, natural antisense, intronic, and sense-overlapping. Compared with protein-coding genes, lncRNAs were, on average, expressed at low levels and displayed significantly higher specificity when considering tissue, developmental stages, and stress responsiveness. The evolutionary analysis indicates higher positional conservation than sequence conservation, probably linked to the conserved modular motifs within syntenic lncRNAs. Moreover, a positive correlation between the expression of intergenic/natural antisense lncRNAs and their closest/parental gene was observed. For those intergenic, the correlation decreases with the distance to the neighboring gene, supporting that their potential cis-regulatory effect is within a short-range. Furthermore, the analysis of developmental studies showed that a conserved NAT-lncRNA family is differentially expressed in a coordinated way with their cognate sense protein-coding genes. These genes code for proteins associated with phloem development, thus providing insights about the potential involvement of some of the identified lncRNAs in a developmental process. We expect that this extensive inventory will constitute a valuable resource for further research lines focused on elucidating the regulatory mechanisms mediated by lncRNAs in cucurbits.

Keywords: lncRNA-mediated regulation, *Cucurbitaceae* family, computational lncRNAs prediction, transcriptional dataset, gene expression regulation, bioinformatics and agriculture.

INTRODUCTION

Plants, as sessile organisms, have evolved intricate mechanisms to adapt to changing environmental conditions and regulate their development (Zhu, 2016). Deciphering the complex regulatory networks that underlie these processes has been a long-standing challenge in plant biology. Traditionally, protein-coding (PC) genes have been the primary focus of research, but recent advancements in high-throughput sequencing and functional genomics have unveiled a vast landscape of noncoding RNAs (ncRNAs), including a group identified as long noncoding RNAs (lncRNAs), which were initially considered "transcriptional noise" (Mattick, 2009).

Long noncoding RNAs represent a diverse class of transcripts characterized by their length (>200 nucleotides) and lack of protein-coding potential. These molecules are generally expressed at lower levels compared to PC genes. However, they exhibit distinct features, such as tissue-specific expression patterns, spatiotemporal regulation, expression related to environmental changes, and subcellular localization, suggesting their involvement in specific

biological processes (Engreitz, Ollikainen, & Guttman, 2016; Mattick et al., 2023).

Regarding their biogenesis, lncRNAs are usually transcribed by RNA polymerase (Pol) II and subjected to 5′ capping and 3′ polyadenilation (Marquardt et al., 2014; Mattick et al., 2023). Moreover, the plant-specific Pol V has also been linked to lncRNA production (Wierzbicki, Blevins, & Swiezewski, 2021). Occasionally, plant lncRNAs can be transcribed by RNA Pol III and/or Pol IV (Wierzbicki, Haag, & Pikaard, 2008; Wu et al., 2012); although these noncanonical lncRNAs are poorly characterized mainly due to their low expression and high instability (Huang, Zhou, Zhang, & Li, 2023).

Based on their genomic origin, orientation, and proximity to protein-coding genes, plant lncRNAs can be commonly classified into four main classes: (i) intergenic lncRNAs (lincRNAs) that do not overlap with other genes; (ii) sense-overlapping lncRNAs (SOT-lncRNAs) that overlap (total or partially) with the same strand of its associated PC gene; (iii) natural antisense lncRNAs (NAT-lncRNAs) that overlap (total or partially) with the opposite strand of its associated PC gene; and (iv) intronic lncRNAs (int-lncRNAs), located within an intron of the associated PC gene. Both NAT-lncRNAs and lincRNAs constitute the predominant classes of lncRNAs described and characterized in plants (Budak, Kaya, & Cagirici, 2020; Palos, Yu, Railey, Nelson Dittrich, & Nelson, 2023).

Since the discovery in 1994 of ENOD40 (EARLY NODULIN 40) the first lncRNA described in plants (Crespi et al., 1994), and with the advancement of sequencing technologies, an increasing number of lncRNAs have been identified and functionally validated in diverse crops and model plants (Palos et al., 2023). These studies have demonstrated that plant lncRNAs contribute to various aspects of plant biology, including development, stress response, genome stability, photomorphogenesis, reproduction, and flowering (Yang, Ariel, & Wang, 2023). Under a functional viewpoint, lncRNAs exert their regulatory roles through diverse mechanisms, such as transcriptional regulation, chromatin remodeling, epigenetic modifications, RNA splicing, and post-transcriptional regulation (Lucero, Ferrero, Fonouni-Farde, & Ariel, 2021). Although functional studies of plant lncRNAs are at an early stage, it is currently accepted that the availability of a comprehensive atlas of lncRNAs in crop plants will enable us to use lncRNAs as potential biomarkers and/or traits in breeding for tailoring stress-tolerant plants (Jha et al., 2020; Yang et al., 2023).

The *Cucurbitaceae* (cucurbits) constitute an important family of plants, including most of 900 species in over 90 genera that are mainly distributed in tropical and subtropical areas (Guo et al., 2020). Some cucurbit members that were domesticated and subsequently cultivated for thousands of years represent agronomically important crops, including cucumber (*Cucumis sativus*), melon (*Cucumis melo*), watermelon (*Citrullus lanatus*), pumpkin and squash (*Cucurbita pepo*, *C. moschata*, *C. maxima*, and *C. argyrosperma*), and bitter gourd (*Momordica charantia*). Others are used as medicinal plants (*Citrullus colocynthis* and *Siraitia grosvenorii*) or have practical uses, for example, as a bottle (*Lagenaria siceraria*) (Martínez & Jamilena, 2021). In 2020, the area cultivated with cucurbits worldwide was 10.42 million hectares, with a yield of near 360 million tons (https://faostat.fao.org). Furthermore, members of the *Cucurbitaceae* family have been extensively used as model organisms to study fundamental biological processes such as vascular development (Lough & Lucas, 2006), RNA trafficking (Gómez, Torres, & Pallás, 2005), sex determination (Bhowmick & Jha, 2015), fruit ripening (Pech, Bouzayen, & Latché, 2008), and host epigenetic alterations associated with infection (Martinez, Castellano, Tortosa, Pallas, & Gomez, 2014). Consequently, deciphering the molecular mechanisms underlying cucurbit biology is of significant importance for improving crop yield, quality, and resilience in response to changing environments (Ranjan et al., 2022).

Recent studies have revealed the potential importance of lncRNAs as modulators of the development (Dey et al., 2022; Tian et al., 2019), or the response to biotic (Gao et al., 2020; Tian et al., 2022; Zhou et al., 2020) or abiotic (He et al., 2020; Kęska, Szcześniak, Adamus, & Czernicka, 2021) stress conditions in diverse *Cucurbitaceae* family members. However, detailed information about lncRNAs in cucurbits is limited, and only a list of unclassified and barely characterized lncRNAs in cucumber, melon, and watermelon is currently available in GreeNC (Di Marsico, Paytuvi Gallart, Sanseverino, & Aiese Cigliano, 2022), PLncDB V2.0 (Jin et al., 2021) or CANTATAdb 2.0 (Szcześniak, Bryzghalov, Ciomborowska-Basheer, & Makałowska, 2019), the most relevant databases containing plant lncRNAs. Notably, no standardized information about lncRNAs can be found in the Cucurbit Genomics Database (Zheng et al., 2019), which is the reference portal for the genomic annotation of members of this family. Thus revealing the lack of studies able to offer detailed information about the global landscape of these regulatory RNAs in cucurbits.

Here, we use a vast dataset of more than 1000 RNA-seq samples to gain insights into the identity, characteristics, and expression of lncRNAs in nine representative species of the family *Cucurbitaceae*. Using a custom pipeline, we have identified intergenic, natural antisense, intronic, and sense-overlapping lncRNAs, finding that distinct molecular features are associated with each type. Next, we have analyzed the evolutionary relationships of the predicted lncRNAs and their expression in different tissues, developmental stages, and stress conditions. Finally, we also studied their potential role in the regulation of gene expression in cucurbits. Overall, this exhaustive study

reveals the diversity and molecular features of these emerging regulatory players in cucurbits, providing the first comprehensive map of lncRNAs in this agronomically relevant family.

RESULTS

Strategy for lncRNA prediction

A graphical resume of the pipeline developed in this work for the identification, classification, and characterization of lncRNAs in cucurbits is depicted in Figure 1. All the available RNA-seq data corresponding to 12 cucurbit species, comprising 3494 RNA-seq libraries and 271 RNA-seq projects, were downloaded from the SRA database (53). SRA accession codes, quality control, and strandedness information are detailed in Table S1. After quality control and strandedness inference, three species having only one strand-specific RNA-seq sample (*B. hispida*, *L. aegyptiaca*, *and S. grosvenorii*) were removed from our dataset. Consequently, 9 species (*C. sativus*, *C. melo*, *C. lanatus*, *L. siceraria*, *C. moschata*, *C. argyrosperma*, *C. pepo*, *C. maxima*, and *M. charantia*) comprising 1116 cleaned and strand-specific RNA-seq libraries and 78 RNA-seq projects were included in the subsequent analysis (Table 1). The nine analyzed species constitute representative members of three different tribes and five genera in the *Cucurbitaceae* family (Guo, Xu, et al., 2020).

Cleaned reads were mapped to the corresponding reference genome and assembled to create a merged transcriptome for each analyzed species (Figure 1a; see Experimental Procedures for details). Comparable levels of genome coverage were observed across the selected species (Figure S1). Assembled transcripts were categorized according to their position in the genome relative to protein-coding (PC) genes, and those classified as intergenic ("u"), antisense ("x"), intronic ("i"), and sense ("o" or "e") were selected. Only nonredundant transcripts longer than 200 nucleotides and expression levels above 0.3 Fragments per Kilobase Million (FPKM) in at least one of the experiments were considered.

As detailed in Experimental Procedures, we used three tools: CPC2 (Kang et al., 2017), FEELnc (Wucher et al., 2017), and CPAT (Wang et al., 2013) to evaluate the coding potential of the analyzed transcripts. Moreover, we looked for homology with known proteins and domains in the SwissProt (Anon, 2023) and Pfam (Mistry et al., 2021) databases, respectively. It is important to note that our strategy for lncRNA prediction was designed to keep a balance between sensitivity and robustness. Consequently, inferred lncRNAs were classified into three confidence levels according to the fulfillment of the following criteria: (i) *High* (HC): predicted as lncRNA by the three software (CPC2, CPAT, and FEELnc) and lacking homology with ORFs in the protein databases; (ii) *Medium* (MC): predicted as lncRNA by the three software but have similarity with one of the protein databases or lack similarities with the two protein databases but are only predicted by two of the three software; and (iii) *Low* (LC): predicted as lncRNA by two of the software or only one software and lacks similarity with one of the protein databases (Figure 1b).

Once potential lncRNAs homologous to miRNA precursors, ribosomal, transfer, small nuclear and small nucleolar RNAs were excluded (see Experimental Procedures for details), transcripts were aligned and compared against those deposited in three plant lncRNA databases (PLncDB V2.0, CANTATAdb 2.0, and GreeNC) (Figure 1b).

The predicted lncRNAs were classified according to their genomic location into intergenic lncRNAs (lincRNAs), natural antisense lncRNAs (NAT-lncRNAs), intronic lncRNAs (int-lncRNAs), and sense-overlapping lncRNAs (SOT-lncRNAs). Finally, further analyses were performed to compare the characteristics, conservation, and expression profiles of the identified lncRNAs (Figure 1c).

Wide ranges of lncRNAs are distributed through cucurbit genomes

Following the strategy previously described, we were able to predict 91 209 nonredundant lncRNAs in the analyzed cucurbits (Figure 2a and Table S2). The highest number of potential lncRNAs (29140) was identified in *C. melo*. Comparable amounts of lncRNAs were described in *C. sativus* (13563), *C. lanatus* (11189), and *C. argyrosperma* (10077). While *C. pepo* (8172), *C. moschata* (5472), *C. maxima* (4672), *M. charantia* (4515), and *L. siceraria* (4409) were the species with the fewest predicted lncRNAs (Table S2). The lncRNAs, categorized as High-Confidence (HC), were the predominant class identified in the nine analyzed cucurbits (with percentages ranging from 68% to 84%). In contrast, those classified as Low-Confidence (LC) were less abundant (4.5%–12.4%). Intermediate values (9.3%–19.3%) were obtained for lncRNAs categorized as Medium-Confidence (MC) (Table S2).

According to their genomic context, intergenic lncRNAs (lincRNAs) were the most abundant type in the nine analyzed species, with 58 907 potential lncRNAs representing 64.5% of the total. They were followed by those classified as natural antisense lncRNAs (NAT-lncRNAs), which, with a total number of 18 050 lncRNAs (except in *C. argyrosperma*), were the second most abundant class (Table S2). The sense-overlapping lncRNAs (SOT-lncRNAs) and intronic lncRNAs (int-lncRNAs), with 6742 and 7510 identified transcripts, respectively, were the least represented, and their prevalence varied depending on the species (Table S2).

To determine whether the predominant identification of lincRNAs may be associated with the genome size of the cucurbit species, we searched for correlation between: (*i*) the percentage of the genome covered by intergenic

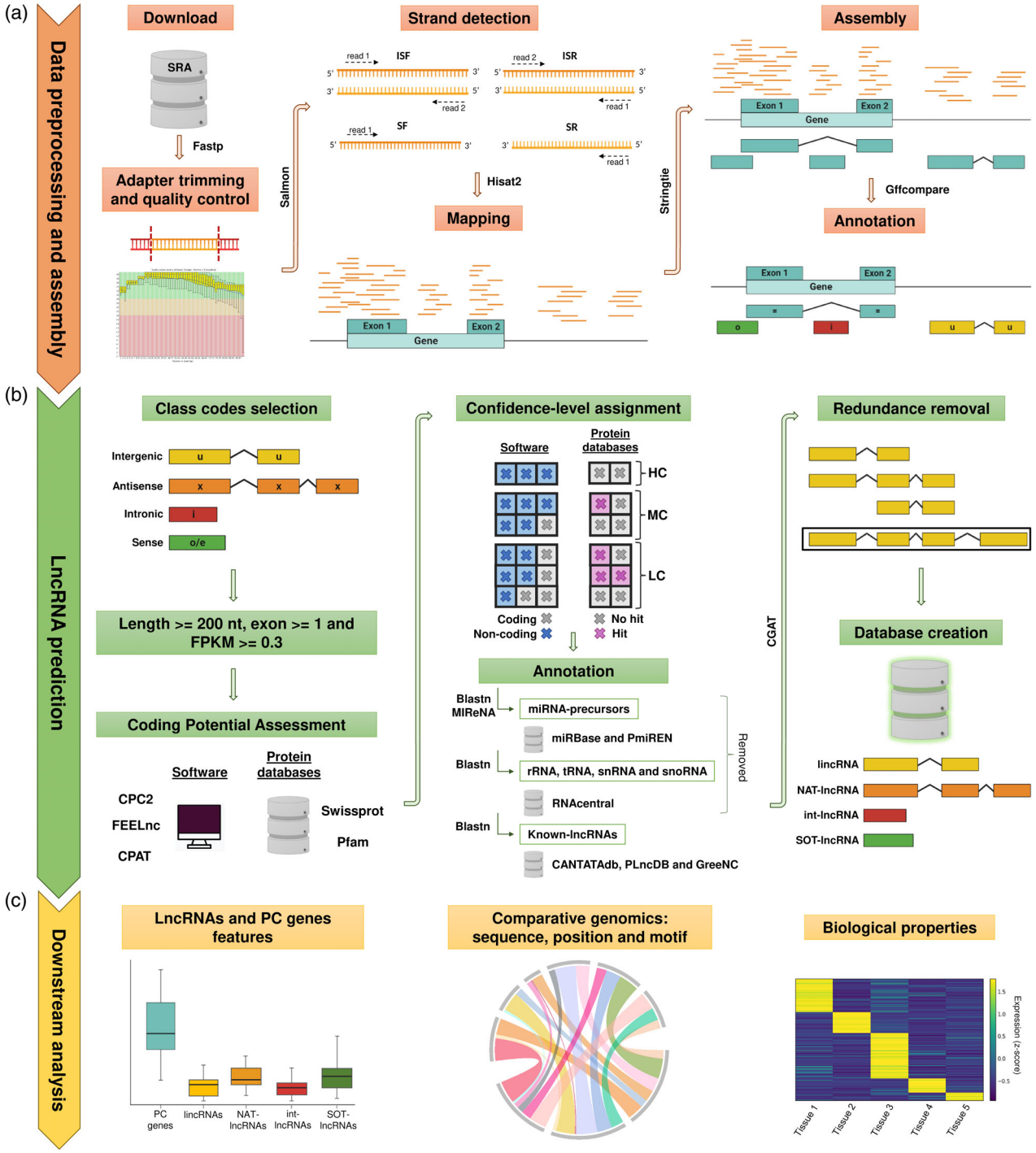


Figure 1. Graphic representation of the bioinformatic workflow used for the prediction, classification, and analysis of lncRNAs.
(a) Data recovery, pre-processing, and transcriptome assembly.
(b) Prediction and categorization of lncRNAs: intergenic (lincRNAs), natural antisense (NAT-lncRNAs), intronic (int-lncRNAs), and sense-overlapping (SOT-lncRNAs) lncRNAs.
(c) Downstream analysis to compare lncRNA features, conservation at three levels (sequence, genomic position, and motifs), differential expression (related to tissue, development, and environment), and expression correlation (influence on nearby/cognate gene expression).

regions (considering intergenic regions to all that is not PC gene) and the percentage of intergenic regions covered by HC-lincRNAs; and (*ii*) the percentage of the genome covered by intergenic regions (considering intergenic regions to all that is not PC gene) and the number of HC-lincRNAs. In both cases, the results obtained support that

Table 1 Detailed information about analyzed data

Species	ID	Genome Size (bases; Mb)	Samples (number)			Projects (number)			Final data size (bytes; Gb)
			Download	Trimming and QC	Strand specific	Download	Trimming and QC	Strand specific	
Cucumis sativus	csa	226.21	1388	1334	360	129	127	35	1167.43
Cucumis melo	cme	357.86	802	776	383	45	44	16	820.87
Citrullus lanatus	cla	365.45	717	711	231	55	54	17	663.61
Lagenaria siceraria	lsi	313.81	92	92	9	7	7	3	27.07
Cucurbita moschata	cmo	273.42	127	126	39	17	16	6	102.73
Cucurbita argyrosperma	car	231.58	10	10	9	2	2	2	30.36
Cucurbita pepo	cpe	263.38	143	142	50	13	13	7	112.61
Cucurbita maxima	cma	279.69	50	50	27	10	10	4	43.36
Momordica charantia	mch	294.01	74	74	8	5	5	2	27.73

the number and relative coverage of lincRNAs identified in the analyzed Cucurbit species show no significant correlation with their genome size (Figure S2).

In order to study the genomic distribution of these regulatory RNAs, we compared the localization patterns of the different types of lncRNAs with those of PC genes. As shown in Figures 2b and S3, the predicted lncRNAs were distributed throughout the genomes of all nine analyzed cucurbits and did not show such a clear pattern as the predominant localization of PC genes in noncentromeric regions. The mean genome covered by the predicted lncRNAs (5.13%) was considerably lower than the observed for PC genes (35.86%) (Table S3 and Figure S4). Considering the categorization by prediction confidence, we observed that the HC-lncRNAs cover a higher percentage of the genome (3.73%) than the MC-lncRNA (0.76%) and LC-lncRNAs (0.64%) (Table S3 and Figure S4). The specie-specific analysis showed that the HC-lncRNAs predicted in *C. sativus* and *C. melo* exhibit the highest values of genome coverage (7.60% and 7).

To have a more quantitative perspective about the spatial distribution, we defined high-density regions (see Experimental Procedures for details) for each type of lncRNA and also for PC genes in the nine species and estimated their overlap (Figure S5). Our results show that the major overlap of high-density regions was observed between NAT-lncRNAs and PC genes (median of 31.53%), while values below 9% were observed for the remaining lncRNA types (Figure S5a). Comparable results were obtained when the number of overlapping high-density regions was individually determined in the nine cucurbit species (Figure S5b).

Predicted lncRNAs show significant sequence differences with coding RNAs

To provide a more detailed view of their molecular properties, we analyzed the GC content, exon number, length, expression, and repeat content of the predicted lncRNAs, using as references PC genes and random intergenic regions (when this comparison was possible, see Experimental Procedures for more details). In general, the values observed in lncRNAs were significantly lower than those in PC genes, except for the content of repetitive regions, where the value obtained for lncRNAs was significantly higher (Figure 3a and Table S4). The most evident difference was found for the expression level estimated as Transcript Per Million (TPM), where lncRNAs (median of 0.40 TPM) were expressed at about 20-fold lower levels than PC genes (median of 8.21 TPM). Considering the differences with intergenic regions, while lncRNAs showed a significantly higher GC content (median of 39.02% and 34.80% for lncRNAs and intergenic regions, respectively), no significant difference was found in the content of repetitive sequences. It is worthy to note that similar differences were observed when this comparison was individually performed for each one of the nine cucurbit species (Figure S6 and Table S4).

Finally, considering the confidence of the lncRNA prediction, the observed differences with PC genes and intergenic regions were more evident in lncRNAs classified as HC than in those identified as MC and LC (Figure 3b and Table S4). Considering that HC-lncRNAs were the most reliably predicted and the predominant class in all the species analyzed, further analyses were restricted to this lncRNA category.

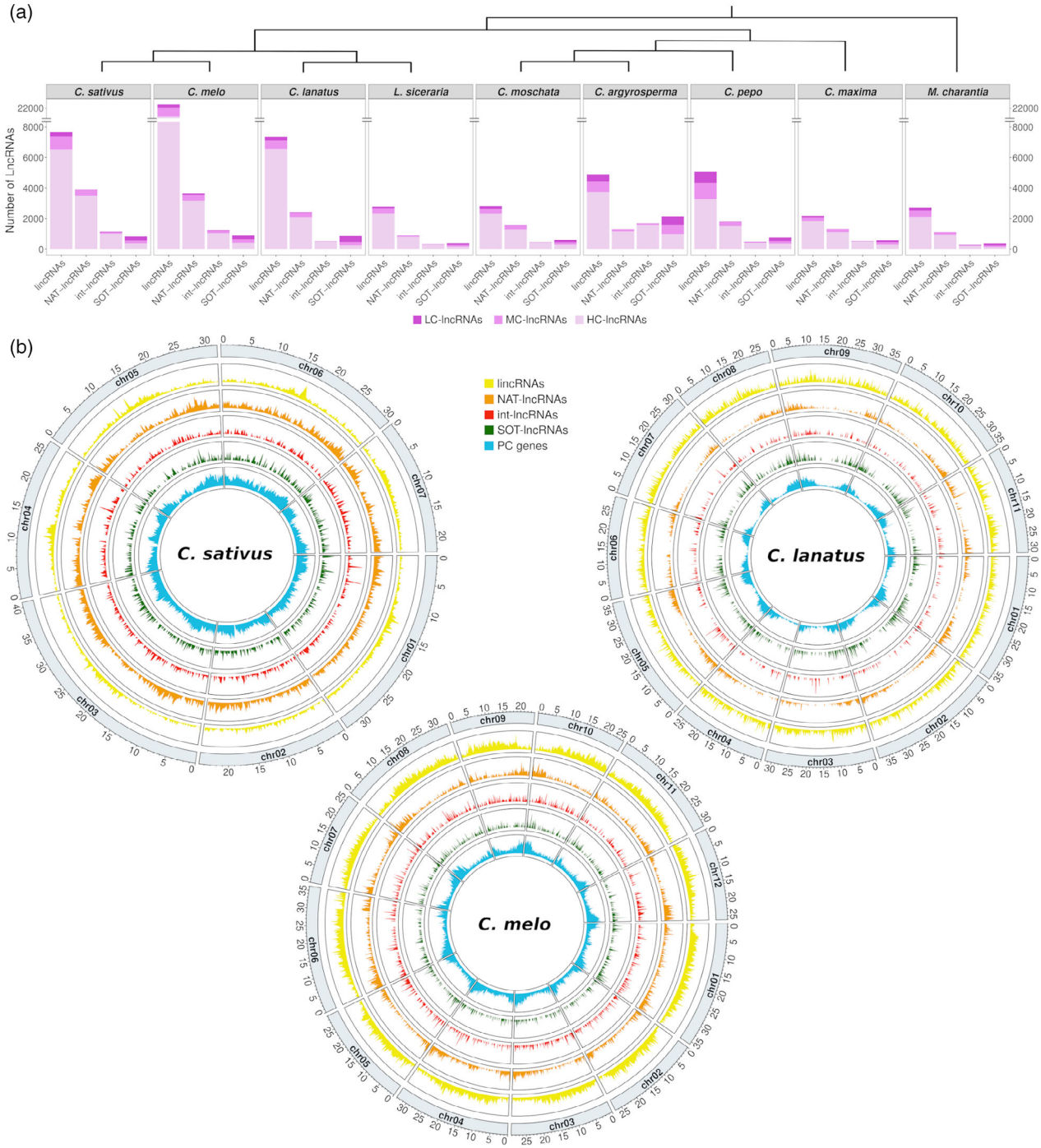


Figure 2. Detailed description and global landscape of the predicted lncRNAs in cucurbits.
(a) Bar plot representing the number of lncRNAs identified in the nine species analyzed in the family *Cucurbitaceae*. These are classified into intergenic (lincRNAs), natural antisense (NAT-lncRNAs), intronic (int-lncRNAs), and sense-overlapping (SOT-lncRNAs) lncRNAs, and categorized according to the confidence level of the prediction into: High-Confidence (HC-lncRNAs), Medium-Confidence (MC-lncRNAs) and Low-Confidence (LC-lncRNAs) lncRNAs. The schematic tree on the upper part of the plot indicates phylogenetic relationships among the species (Chomicki, Schaefer, & Renner, 2020).
(b) Circos plots representing the genomic distribution of the four lncRNA classes identified at the three confidence levels and the annotated protein-coding genes (PC genes) in *C. sativus*, *C. melo*, and *C. lanatus*. Density is measured as the number of transcripts per window using a 250 Kb window size.

A more exhaustive analysis of HC-lncRNA by considering individually lincRNAs, NAT-lncRNAs, int-lncRNAs, and SOT-lncRNAs reveals that regarding the GC content, the four types of lncRNAs showed values significantly inferior to those observed in PC genes (median of 43.22%), but superior to intergenic regions (median of 34.80%) (Figure 3c).

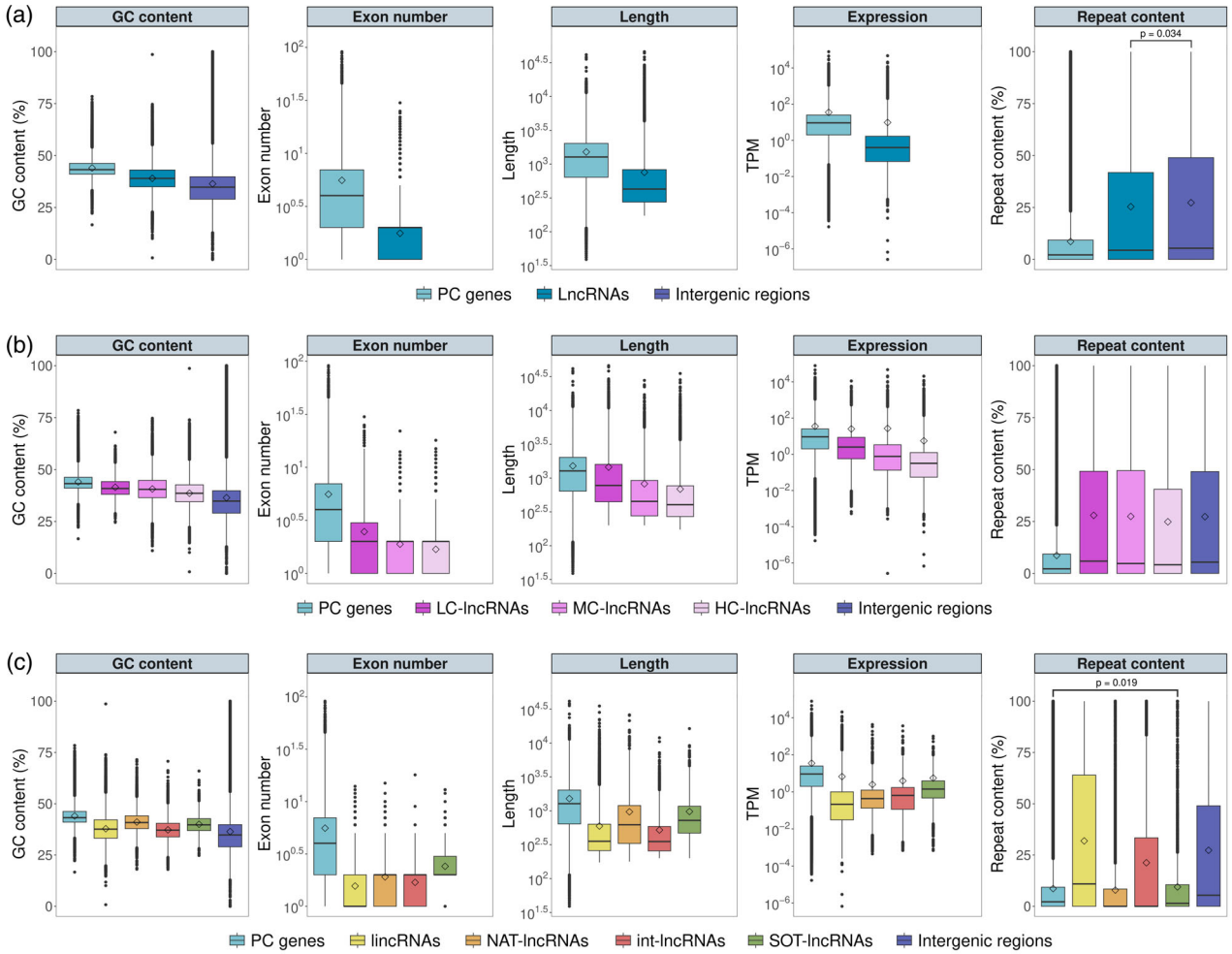


Figure 3. Molecular properties of the identified lncRNAs of cucurbits.
Box plots showing the comparisons of GC content, number of exons, transcript length, expression level, and repetitive content features between lncRNAs, protein-coding genes (PC genes), and random intergenic regions (when corresponding), considering: (a) all the identified lncRNAs; (b) all the identified lncRNAs separated according to the confidence level of the prediction into: High-Confidence (HC-lncRNAs), Medium-Confidence (MC-lncRNAs), and Low-Confidence (LC-lncRNAs) lncRNAs; (c) HC-lncRNAs separated into the four classes: intergenic (lincRNAs), natural antisense (NAT-lncRNAs), intronic (int-lncRNAs), and sense-overlapping (SOT-lncRNAs) lncRNAs. Differences between pairs of box plots within each facet are statistically significant (Wilcoxon rank sum test *P*-value <0.01), unless indicated otherwise. In each box plot, the mean value is represented by a square. Information on box colors is detailed below each panel.

Among the four types, generally, NAT-lncRNAs (median of 40.84%) and SOT-lncRNAs (median of 39.73%) have higher GC content than lincRNAs (median of 37.61%) and int-lncRNAs (median of 37.12%). Similarly, the exon number of the four lncRNA types was significantly lower than that of PC genes (median of 4) (Figure 3c and Table S4). SOT-lncRNAs always showed more than 1 exon in all their transcripts, in contrast to the other classes of lncRNAs, being the only ones with an average of more than 2 exons (mean of 2.41).

It is worthy to note that in general, the length of SOT-lncRNAs (median of 722 bp) and NAT-lncRNAs (median of 623 bp) was higher than the other types of lncRNAs (median of 355 and 352 bp for lincRNAs and int-lncRNAs, respectively), although not as much as PC genes (median of 1275 bp). Consistently with other plant species, the expression level of the four types of cucurbit lncRNAs (median of 0.21, 0.43, 0.64, and 1.21 TPM for lincRNAs, NAT-lncRNAs, int-lncRNAs, and SOT-lncRNAs, respectively) was significantly lower than that observed for PC genes (median of 8.21 TPM) (Figure 3c and Table S4). When the repeat content was analyzed, we observed that lincRNAs and int-lncRNAs have higher content than PC genes, surpassing, in some cases, the value estimated for random intergenic regions. These differences with PC genes with respect to GC content, exon number, length, expression, and repeat content were consistently observed for all analyzed cucurbits (Figure S7).

Evolutionary conservation of lncRNAs

We used sequence similarity and synteny as parameters to analyze the potential relatedness and explore the evolutionary aspects of the HC-lncRNAs predicted in the nine cucurbit species. To infer sequence conservation, a BLAST pairwise alignment between species was performed with all the predicted HC-lncRNAs. Then, putative lncRNA orthologous families were identified by the Markov Cluster (MCL) algorithm using the reciprocal best hits (RBH) from BLAST results. Additionally, a syntenic approach (Hovhannisyan & Gabaldón, 2021) based on 1:1 high-quality orthologous PC genes was employed to classify all the potential HC-lncRNAs into clusters referred to hereafter as syntenic families (see Experimental Procedures for details). These families include lncRNAs from different species that share the same genomic context, meaning that they are surrounded by orthologous PC genes. Considering as conserved the lncRNAs identified in at least two species, we observed that the positional conservation or synteny (mean of 50.97%) was significantly higher than the sequence conservation (mean of 35.41%) (Figure 4a and Table S5). In addition, as expected, the percentage of lncRNAs exhibiting primary sequence conservation was much lower than in PC genes (Figure S8a).

Next, we compared the proportion of lncRNAs included in 3 categories of conservation: low (conserved in 2–3 species), medium (in 4–6 species), and high (in 7–9 species). The obtained results revealed that approximately 27.23% of the lncRNAs predicted in cucurbits were identified as highly conserved at the positional level, while sequence conservation in this category was lower than 5.0% (Figure 4b and Table S5). Further analysis showed that NAT-lncRNAs were the type of lncRNAs with the highest percentage of positional conservation (mean of 54.48%), whereas conservation values obtained for lincRNAs, int-lncRNAs, and SOT-lncRNAs were each less than 25% (Figure 4c and Table S5). Moreover, as in the comparison between sequence and positional conservation, the differences in sequence conservation between PC genes and lncRNAs identified as highly conserved were much greater (Figure S8b). Regarding the relationship between the different cucurbit species, we observed that the syntenic relationships were comparable among all the species analyzed, except for *M. charantia* which showed significantly lower synteny levels (Figures 4d and S9). Positionally conserved lncRNAs were, in general, regularly distributed throughout the genome of the analyzed cucurbits (Figure 4d). Moreover, as expected from their closer evolutionary relationships, the majority of the syntenic relationships in lncRNAs with low and medium conservation were detected between species of the same genus (Figures S10 and S11, respectively).

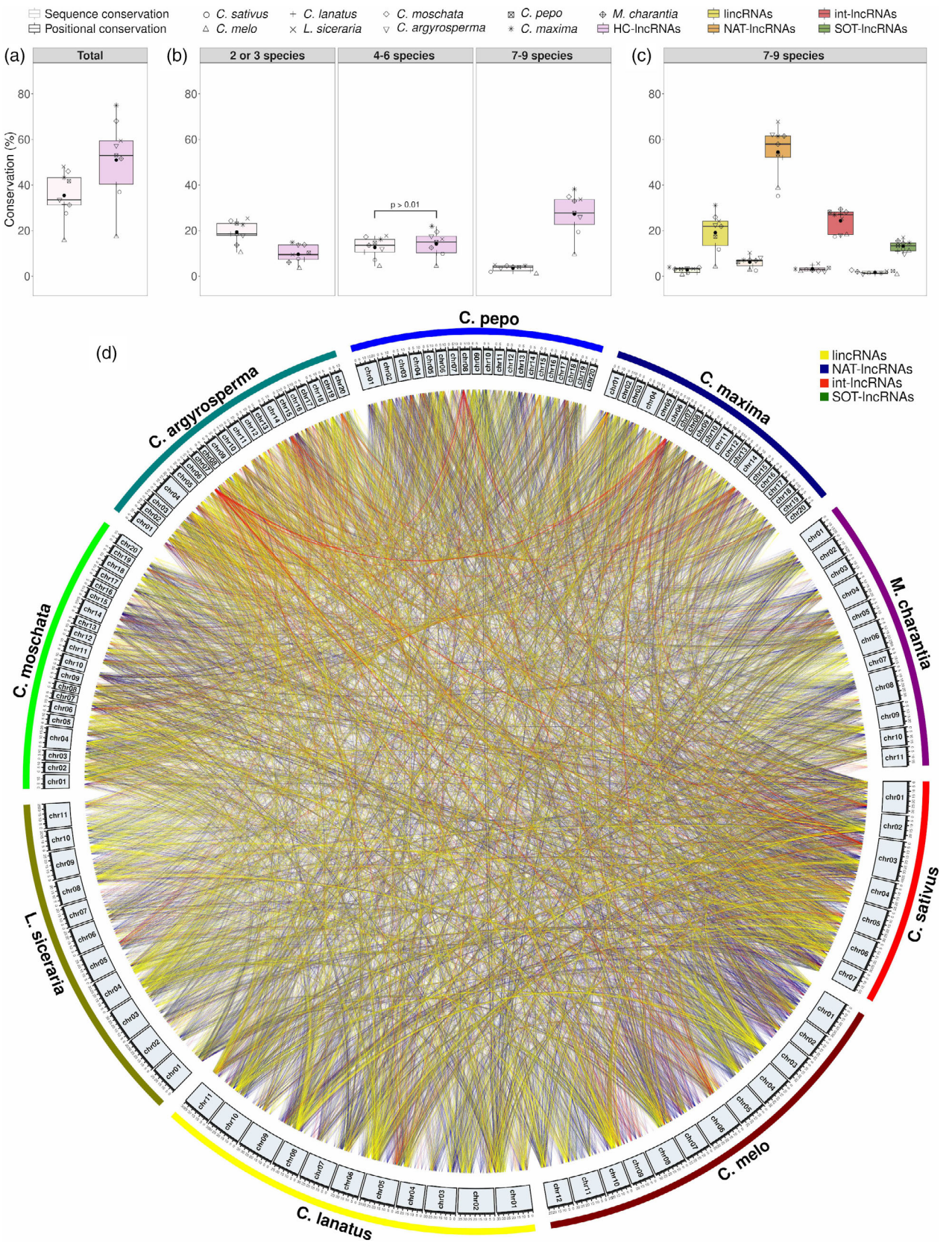
Finally, we analyzed the presence of conserved sequence motifs shared by lncRNAs within syntenic families, and their statistical significance was calculated by generating datasets of randomized syntenic families (see Experimental Procedures for details).

We identify 982 syntenic families (30.13%) with at least one shared conserved motif. This proportion of syntenic families was significantly higher than what could be expected by chance, as assessed in the simulated data (mean of 6.93%) (Figure S12a, upper panel). Although comparable results were observed when each class of lncRNAs was considered individually, the higher differences in shared conserved motifs were detected in SOT-lncRNAs and int-lncRNAs (Figure S12a, lower panel). We also compared the length (in nucleotides) and the statistical robustness (estimated by considering *e*-values) of the predicted motifs. The results showed that motifs identified in syntenic families were significantly longer and more reliable than those predicted in the simulated data (upper panels in Figure S12b,c, respectively). These significant differences were also observed when the different classes of lncRNAs were considered individually (Figure S12b,c, lower panels).

Expression of cucurbit lncRNAs is tissue-specific, development-dependent, and responsive to environmental changes

To explore the potential biological properties of the predicted lncRNAs, we specifically analyzed projects included in our dataset in which different plant tissues, developmental phases, and/or changing environmental conditions have been considered. According to the criteria detailed in Experimental Procedures, 13 projects comprising eight cucurbit species were selected to analyze the tissue-specific expression of HC-lncRNAs and PC genes (see

Figure 4. Conservation of cucurbit lncRNAs at the sequence and positional levels.
(a) Box plots showing the percentage of lncRNAs identified as conserved at sequence (light color) and positional (dark color) levels.
(b) Detailed information considering if lncRNAs are shared by two or three (left), four to six (center), and seven to nine species (right).
(c) Values obtained for highly conserved lncRNAs (shared by 7–9 species) according to lncRNA classes (lincRNAs, NAT-lncRNAs, int-lncRNAs, and SOT-lncRNAs). The nine species and the four lncRNA classes are represented by different shapes and colors, respectively (detailed on the top). Differences between pairs of box plots, sequence versus positional conservation, are statistically significant (Wilcoxon signed rank test *P*-value <0.01), unless indicated otherwise. In each boxplot, the mean value is represented by a black point.
(d) Circos plot showing the relationships of the highly conserved lncRNAs (shared by 7–9 species) at the positional level (syntenic relationships). Each lncRNA class (lincRNAs, NAT-lncRNAs, int-lncRNAs, and SOT-lncRNAs) is represented by a different color (detailed on the top right). In all sections, only lncRNAs identified as High-Confidence lncRNAs (HC-lncRNAs) were considered. Intergenic (lincRNAs), natural antisense (NAT-lncRNAs), intronic (int-lncRNAs), and sense-overlapping (SOT-lncRNAs) lncRNAs.



The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2024), **120**, 799–817

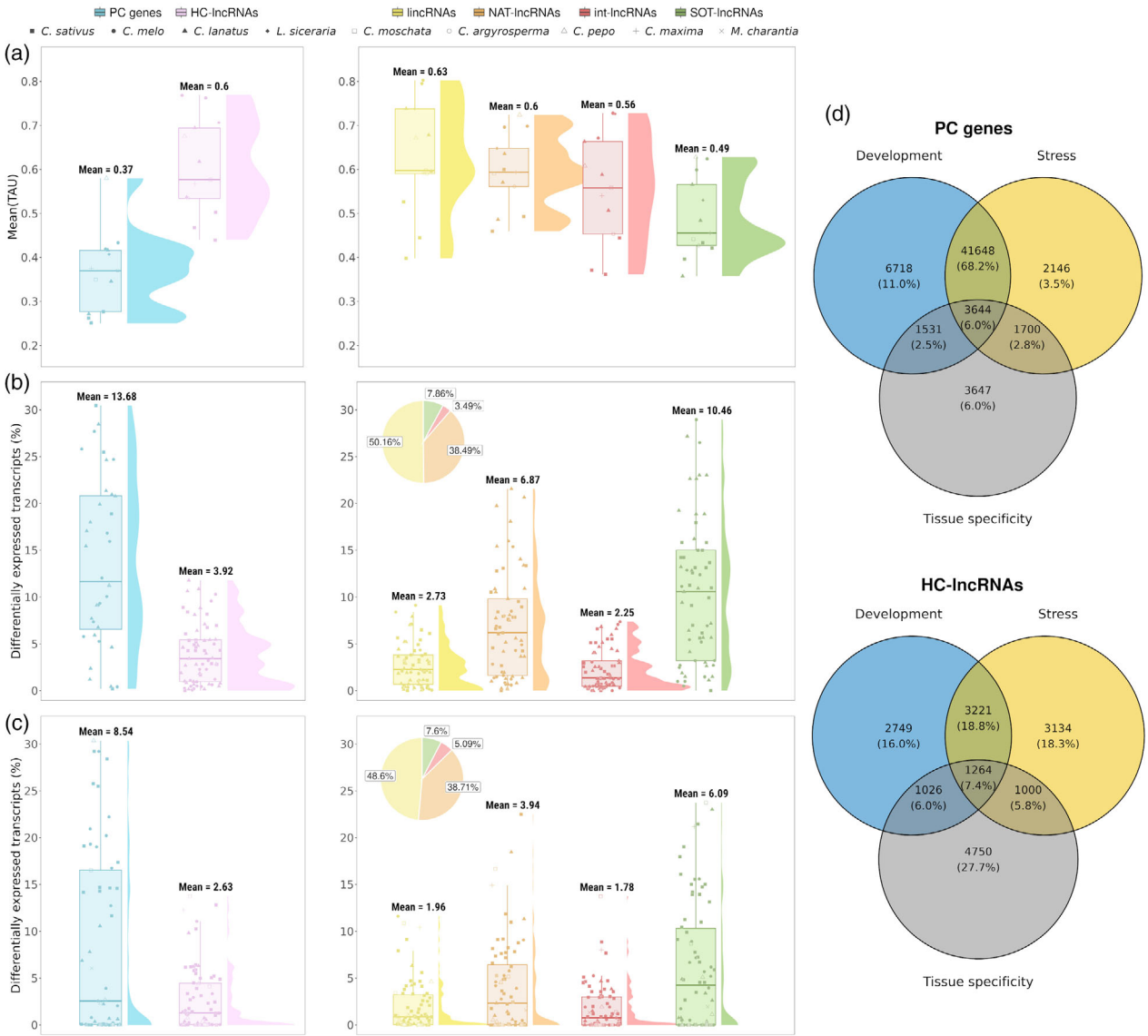


Figure 5. Expression of cucurbit lncRNAs considering different tissues, developmental stages, and stress conditions.
Box plots on the left side represent the comparison between protein-coding genes (PC genes) and lncRNAs, while those on the right represent the comparison between the four lncRNA classes: intergenic (lincRNAs), natural antisense (NAT-lncRNAs), intronic (int-lncRNAs), and sense-overlapping (SOT-lncRNAs).
(a) Comparison of tissue-specificity calculated as the mean of TAU values (0 means broadly expressed and 1 is the most specific), where each point refers to a study with at least 3 different tissues.
(b) Comparison of the percentage of differentially expressed (DE) transcripts in different developmental stages, where each point refers to values obtained for each independent comparison (late development stage vs. early development stage).
(c) Comparison of the percentage of DE transcripts in response to different stress conditions, where each point refers to values obtained for each independent comparison (treatment vs. control). In sections b and c, pie charts represent the percentage of DE lncRNAs in each class over the total number of DE lncRNAs. The nine species and the four lncRNA classes are represented by different shapes and colors, respectively (detailed on the top). Differences between pairs of box plots on the left side, lncRNAs vs. PC genes, are statistically significant (Wilcoxon signed rank test *P*-value <0.01), unless indicated otherwise.
(d) Venn diagrams showing the overlap between tissue-specific transcripts (TAU above 0.8) and those differentially expressed in specific developmental stages or in response to stress conditions. The upper diagram compares PC genes, while the lower one compare lncRNAs. In all sections, only lncRNAs identified as High Confidence (HC-lncRNAs) were considered.

Tables S6 and S7 for detailed information). Tissue-specificity was estimated using the TAU value, which varies from 0 to 1, where 0 means broadly expressed and 1 is specific (Yanai et al., 2005). Our results showed that the tissue-specific expression ratio (TAU values) observed for the totality of the cucurbit lncRNAs (mean of TAU 0.60) was significantly higher than the obtained for protein-coding genes (mean of TAU 0.37) (Figure 5a, left panel). These significant differences in the mean TAU values were also observed when each lncRNA class was

considered individually. LincRNAs (mean of TAU 0.63) and NAT-lncRNAs (mean of TAU 0.60) were in general more specific in their expression in comparison to int-lncRNAs and SOT-lncRNAs, with TAU ratios of 0.56 and 0.49, respectively (Figure 5a, right panel). To obtain more detailed information on the tissue specificity of the predicted HC-lncRNAs, we performed an additional analysis in which the expression across tissues of the lncRNAs identified as tissue-specific (TAU values above 0.8) was analyzed. For this purpose, we selected seven studies performed in five cucurbits (*C. melo*, *C. lanatus*, *C. moschata*, *C. maxima*, *and L. siceraria*) in which four or more tissues had been analyzed. The results obtained showed that, in general, the highest TAU values were observed in flower, fruit, and apical tissues (Figure S13). In contrast, and with the exception of one study conducted in *C. maxima*, the lowest tissue-specific expression ratio was found in leaf samples.

The correlation between different developmental stages and the expression of the predicted lncRNAs was analyzed in three cucurbit species, comprising a total of 65 developmental events in 13 projects (see Tables S6 and S7 for detailed information). The obtained results showed that the proportion (estimated as the mean value for the total of the studies) of lncRNAs with expression associated with development (3.92%) was significantly lower than the obtained (13.68%) when protein-coding genes were considered (Figure 5b, left panel). Further analysis showed that, although SOT-lncRNA (with a mean of 10.46%) was the class with the highest percentage of members with development-associated expression (Figure 5b, box plots in the right panel), considering the absolute number of lncRNA transcripts, 50.16% of the lncRNAs with development-associated expression corresponded to lincRNAs (Figure 5b, pie chart in the right panel).

Regarding environment-dependent expression, we analyzed 75 stress events (with both biotic and abiotic origins) performed in seven cucurbit species (see Tables S6 and S7 for detailed information). Our analysis demonstrated that the mean of the relative ratio of stress-responsive lncRNAs in cucurbits (2.63%) was also significantly lower than observed for PC genes (8.54%) in the same studies (Figure 5c, left panel). Consistent with what was observed in development, transcripts classified as SOT-lncRNA showed the highest percentage of reactive members (mean 6.09%) (Figure 5b, box plots in the right panel), and lincRNAs were the predominant type (48.6% of the total) of stress-responsive lncRNA (Figure 5c, pie chart in the right panel).

Finally, we selected the cucurbit species (*C. lanatus*, *C. sativus*, and *C. melo*) with available data in the three selected conditions to compare the specificity of the expression. Results shown in Figure 5d demonstrate that predicted lncRNAs have (in comparison to PC genes) a higher specific expression rate (defined as the percentage of specific transcripts) in the three analyzed conditions: tissue (27.7% vs. 6.0%), development (16.0% vs. 11.0%), and stress (18.3% vs. 3.5%) (Figure 5d). These differences in specificity were also observed when each one of the three species was analyzed individually (Figure S14).

Expression of a group of lincRNAs and NAT-lncRNAs in cucurbits is correlated with that of near PC genes

There is increasing evidence that many functionally characterized plant lncRNAs can act as regulators of gene expression, either in *cis* or in *trans* (Palos et al., 2023). A characteristic of *cis*-regulatory lncRNAs is correlation in expression relative to nearby (for lincRNAs) (Palos, Nelson Dittrich, Yu, et al., 2022) or cognate (for NAT-lncRNAs) (Zhao et al., 2018) genes. To identify cucurbit lncRNAs with potential *cis*-regulatory roles, we analyzed the correlation between lincRNAs/NAT-lncRNAs and their nearby/cognate PC genes (see Experimental Procedures for details). In addition, only stress- and development-associated studies from *C. melo*, *C. sativus*, and *C. lanatus* species (i.e., those with the most studies) were included in this analysis.

As for lincRNAs, we identified 603, 161, and 168 unique pairs of lincRNAs correlated with nearby PC genes, where 73.96%, 81.37%, and 67.86% exhibited strong correlation ($r \geq 0.5$ or $r \leq -0.5$) in *C. melo*, *C. sativus*, and *C. lanatus*, respectively. Considering the mean Pearson correlation coefficients (PCC) for each dataset, the correlation of lincRNA-PC gene pairs (mean PCC of 0.162) was significantly more positive than that observed for PC gene-PC gene and random gene pairs (mean PCC of 0.066 and 0, respectively) (Figure 6a and Table S8). When the distance between the analyzed lincRNA and its neighboring PC gene was examined, we observed that the positive correlation values decreased as the distance increased (Figure 6b and Table S8). A comparable situation was observed for the lincRNA-PC gene pairs when these were analyzed individually in each cucurbit species (Figure S15a,b and Table S8).

When NAT-lncRNAs were analyzed, we identified 385, 242, and 118 unique pairs of NAT-lncRNAs correlated with cognate PC genes, where 84.16%, 77.27%, and 73.73% showed strong correlation ($r \geq 0.5$ or $r \leq -0.5$) in *C. melo*, *C. sativus*, and *C. lanatus*, respectively. Similarly to lincRNAs, we found that the positive correlation values for overlapping NAT-lncRNA/sense gene pairs (mean PCC of 0.278) were significantly higher than those obtained for random sense/antisense gene pairs (mean PCC of 0.005) (Figure 6C and Table S8). This significant positive correlation between the expression of NAT-lncRNAs and their cognate PC genes was maintained in each species analyzed (Figure S15c and Table S8).

To investigate the potential functional role of the pairs of NAT-lncRNAs and associated PC genes, we selected a

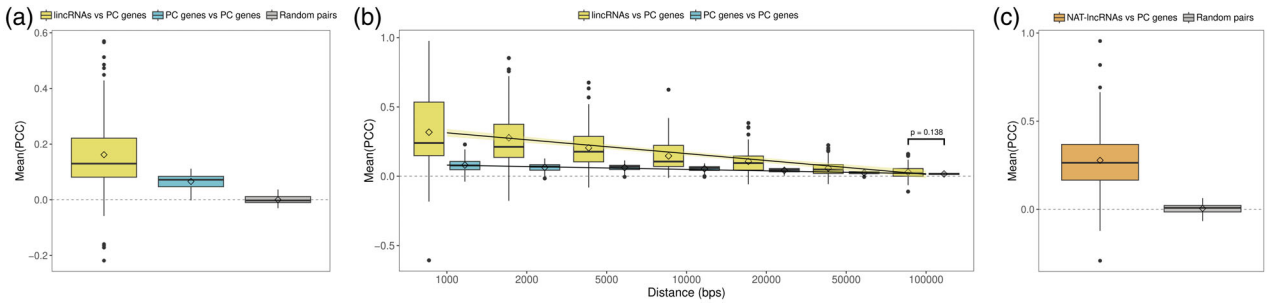


Figure 6. Cucurbit lincRNA/NAT-lncRNA influence on nearby/cognate gene expression.
Expression correlation analysis was used in all experiments related to the development and stress response of *C. sativus*, *C. melo*, and *C. lanatus* species.
(a) Box plots showing the expression correlation between a gene and its nearby gene. The mean of the Pearson correlation coefficients (PCC) obtained for each gene pair group (lincRNA-PC gene pairs, PC gene-PC gene pairs, and random gene pairs) and experiment is represented.
(b) Box plots showing the expression correlation between a gene and its nearby genes within defined distances (in bps). The mean of the PCC obtained for each gene pair group (lincRNA-PC gene pairs and PC gene-PC gene pairs) and experiment is represented. Note that all pairs within smaller distances are contained within larger distances.
(c) Box plots showing the expression correlation between a gene and its cognate gene. The mean of the PCC obtained for each gene pair group (NAT-lncRNA-PC gene pairs and random gene pairs) and experiment is represented. Different groups of gene pairs are represented by different colors (detailed on the top). Differences between pairs of box plots within each section are statistically significant (Wilcoxon rank sum test *P*-value <0.01), unless indicated otherwise. In all sections, only lncRNAs identified as High Confidence (HC-lncRNAs) were considered.

syntenic NAT-lncRNA conserved in *C. argyrosperma*, *C. melo*, *C. sativus*, *and C.* lanatus with differential expression in diverse studies performed in the three last species (Figure S16a). Detailed analysis revealed that in *C. melo*, *C. sativus*, *and C. lanatus*, this syntenic NAT-lncRNA and its cognate PC gene showed differential expression associated predominantly with development (mean lncRNA and PC gene 54.04% and 46.02% of the studies, respectively) than that stress response (mean lncRNA and PC gene 8.05% and 8.12% of the studies, respectively) (Figure S16b). The expression of both NAT-lncRNA and its associated PC gene showed a significant positive correlation ($R = 0.51$, $p = 1.2e^{-4}$) when the totality of the development-related studies was considered (Figure S16C). Interestingly, this correlation value was higher than the observed for the totality of the NAT-lncRNA–PC gene pairs ($R = 0.278$) analyzed in cucurbits (Figure 6c). The PC gene associated with the NAT-lncRNA was identified to encode one of the subunits of the homologous in cucurbits to the SIEVE ELEMENT OCCLUSION protein of *A. thaliana* (AthSEO), a conserved phloem protein with biological processes associated with phloem development (https://www.uniprot.org/uniprotkb/Q9SS87/entry).

DISCUSSION

The tremendous achievements made over the last few years in the computational prediction and functional biology of plant lncRNA have prompted the creation of detailed datasets containing descriptions and information about this emerging type of regulator in diverse model and nonmodel species (Palos et al., 2022; Palos et al., 2023). However, the lack of systematic analyses performed from a more global point of view has meant that a significant number of economically relevant crops and closely related species can be poorly represented in these repositories. In this sense, members of the *Cucurbitaceae* family constitute a paradigmatic case of worldwide relevant crops in which global information about lncRNAs is limited. Here, we have analyzed a large amount of available transcriptomic studies performed in nine representative species to generate a comprehensive inventory of lncRNAs in cucurbits.

Importantly, our analysis is not restricted to lincRNAs, as it happens in most studies of different plant species (Palos et al., 2022; Xu et al., 2016). The use of directional RNA-seq data enabled the prediction of the other lncRNA categories: NAT-lncRNAs, int-lncRNAs, and SOT-lncRNAs. This is especially relevant since the two latter categories are considerably understudied in plants, while NAT-lncRNAs have been associated with important physiological roles by regulating the expression of their cognate sense genes by different mechanisms (Csorba, Questa, Sun, & Dean, 2014; Wang et al., 2014; Zhao et al., 2018), but their annotation in cucurbits was lacking.

To provide robustness to our inventory, we established three confidence levels for the prediction: High (HC), Medium (MC), and Low (LC). Results obtained during our study revealed that HC-lncRNAs (beside being predominant) showed the highest structural and/or potential biological differences with PC genes. Consequently, the majority (except when the contrary is specified) of the analyses shown in this work were performed with this lncRNA type. In coincidence with what has been observed in other plants, most of the lncRNAs predicted in the nine species are derived from intergenic regions. Considering the number of lncRNAs identified in each species, *C. melo*, *C. sativus*, and *C. lanatus* have the largest number of predicted lncRNAs, while *C. maxima*, *M. charantia*, and *L. siceraria*

have the smallest. Although diverse reasons could be associated with this difference, it is important to note that both groups are composed by the species with the highest and lowest numbers of analyzed samples, respectively. In this sense, it has been proposed that the achievement of multiple transcriptome analyses under diverse environmental and physiological conditions in specific cell types can contribute to the discovery of more functional lncRNAs (Mattick et al., 2023; Yamada, 2017).

The analysis of the main molecular features of the predicted lncRNAs revealed significant differences between these regulatory transcripts and PC genes and/or a set of randomly generated intergenic regions. Consistent with observations in other plants (Chen, Zhu, & Kaufmann, 2020; Huang et al., 2023; Wierzbicki et al., 2021; Yang, Cui, et al., 2023), lncRNAs were shorter and had a lower GC content and exon number than PC genes. In addition, and as expected, cucurbit lncRNAs also showed low expression levels. In this sense, a recent study supports the idea that, in mice, lower expression can be essential for the functional role of lncRNAs by ensuring specific recognition of their regulated targets, suggesting that low accumulation may be an essential feature of how lncRNAs work (Jachowicz et al., 2022).

Another distinctive characteristic of the predicted lncRNAs was their high content of repetitive regions in comparison to PC genes, but significantly smaller than those in random intergenic regions. Interestingly, an increased level of repetitive regions with respect to PC genes has been described as a common characteristic for lncRNAs in several organisms (Mattick et al., 2023). These significant differences with PC genes are generally more evident in lincRNAs, suggesting the existence of certain differential characteristics for this type of lncRNAs, may be related to their relevant regulatory role (Franco-Zorrilla et al., 2007; Wang et al., 2014; Yamada, 2017).

Considering the evolutionary relationships of the predicted lncRNAs, lower sequence conservation was observed in comparison to the positional conservation of the syntenic families. This conservation bias (higher at positional than sequence level) was particularly evident in lncRNAs conserved in most species (7 or more). This suggests that during the evolution of cucurbits, their lncRNAs have mostly diverged in their sequences but not that much in their genomic organization. Interestingly, syntenic lncRNAs also showed a significant tendency to contain more conserved and longer sequence motifs than randomly generated transcripts. Overall, these results suggest that despite large evolutionary distances (evidenced by low sequence conservation), the analyzed cucurbits possess syntenic lncRNAs that share relatively conserved sequence motifs that can be assumed to be modules. This observation is in consonance with previous studies (Graf & Kretz, 2020; Guttman & Rinn, 2012; Hovhannisyan & Gabaldón, 2021) supporting that, in general, lncRNAs tend to acquire a modular structure and are rich in repeats, implying that small sequence elements can also be key determinants of lncRNA function (Kirk et al., 2018).

It is widely accepted that lncRNAs exhibit distinctive biological features such as tissue-specific expression, differential temporal regulation, and expression related to environmental changes (Chen et al., 2020; Datta & Paul, 2019; Mattick et al., 2023; Roulé, Crespi, & Blein, 2021; Yang, Ariel, & Wang, 2023). The detailed analysis of our dataset demonstrated that lncRNAs predicted in cucurbits fulfilled these functional conditions. The obtained results revealed that lncRNAs identified in the eight analyzed species (having studies with multiple tissues) showed significantly higher TAU values than PC genes, supporting that their expression is mostly tissue-dependent, in accordance with the observed for other plant lncRNAs (Chen et al., 2020; Wierzbicki et al., 2021). The proportion of lncRNAs with differential expression was lower than that obtained for PC genes. However, these differential lncRNAs showed (when compared with PC genes) a specific response pattern highly dependent on environment and development (see Figure 5D). Interestingly, this high specificity in the expression pattern was also observed when tissue-dependent lncRNAs were analyzed. However, it cannot be excluded that, at least partly, this apparent specificity in the expression of lncRNAs could also be attributed to the generally low expression level of lncRNAs (Chen et al., 2020).

Under functional viewing, plant lncRNAs have been shown to participate in the regulation of developmental processes and response to biotic/abiotic stress conditions (Chen et al., 2020; Yang, Ariel, & Wang, 2023). In this regard, a positive expression correlation with neighboring genes in multiple datasets has been established as a signature to identify cis-regulatory lincRNAs and NAT-lncRNAs (Palos et al., 2022; Zhao et al., 2018). Consequently, here we investigate the putative role that stress- and development-associated lincRNAs and NAT-lncRNAs identified in cucurbits can play in the *cis*-regulation of PC genes. Our results support the existence of a clear positive correlation between the expression of these lncRNAs and their neighboring (for lincRNAs) or cognate (for NAT-lncRNAs) PC genes. Interestingly, in coincidence with the observations for *cis*-regulators lncRNAs in general (Gil & Ulitsky, 2020; Mattick et al., 2023) and plant lncRNAs in particular (Palos et al., 2022), the highest correlation values obtained for lincRNAs corresponded to the pairs with closer neighboring genes (Palos et al., 2022). Thus, as demonstrated for distinct animal and plant species (Gil & Ulitsky, 2020), the putative *cis*-regulatory effects of cucurbit lincRNAs should act within fairly short genomic distances. The detailed analysis of a syntenic NAT-lncRNA conserved in the nine analyzed species and its cognate PC gene shows that both transcripts exhibit significant co-expression,

predominantly altered in development-associated studies, in comparison to stress response. The observation that the PC gene associated with this conserved NAT-lncRNA encodes for a subunit of a protein (SEO) associated with developmental function (Froelich et al., 2011) suggests that certain pairs of NAT-lncRNA and their cognate PC gene identified here can constitute regulatory modules controlling diverse aspects of cucurbits biology, as previously described in arabidopsis (Meena et al., 2023; Zhao et al., 2018), maize (Huanca-Mamani et al., 2018), rice (Liu et al., 2018), tomato (Cui, Luan, Jiang, Bao, & Meng, 2017) or white lupin (Aslam et al., 2022).

Both the differential structural and functional features support that the cucurbit lncRNAs constitute a subset of RNAs significantly distinctive of the protein coding and/or randomly generated intergenic regions. The population of lncRNAs found in the *Cucurbitaceae* family exhibits the main characteristics (low accumulation, tissue-specific expression, development-related regulation, lower sequence conservation, conservation of structural domains, and expression related to environmental changes) shown by this type of regulatory RNAs in other plant species, from the model plant *A. thaliana* to crops such as cotton, wheat, rice, and maize (Domínguez-Rosas, Hernández-Oñate, Fernandez-Valverde, & Tiznado-Hernández, 2023; Traubenik, Charon, & Blein, 2024). Thus providing additional support to the emergent notion that this heterogeneous group of transcripts can constitute one of the key regulators in plant biology (Numan, Sun, & Li, 2024; Yang, Ariel, & Wang, 2023).

Therefore, in this study, we provide a comprehensive catalog of the different types of lncRNAs in the family *Cucurbitaceae*, describing the confidence of their prediction, their molecular features, and conservation (Table S9 - available at https://github.com/ncRNA-lab/Cucurbit_lncRNAs_landscape/Tables/Final_summary). Although we recognize that this is only an initial step and that additional experimental approaches and cucurbit species should be considered in the future, it is expected that this extensive inventory will constitute a valuable resource for further research lines focused on elucidating the basis of the regulation mediated by lncRNAs in cucurbits. In addition, considering that lncRNAs are emerging as target materials for genetic engineering in plant improvement (Yang, Ariel, & Wang, 2023), this detailed collection paves the way for the ncRNAs described here to become a valuable genetic resource for cucurbit breeding in the near future.

EXPERIMENTAL PROCEDURES

Data preprocessing and assembly

With the aim to perform a comprehensive study, we selected all cucurbit species with a fully sequenced genome and annotation file available (at 24th of January 2022) in the Cucurbit Genomics Database (CuGenDB) (Zheng et al., 2019). Then, all RNA-Seq data publicly available (at 24th of January 2022) at the Sequence Read Archive database (SRA) (Leinonen et al., 2011) were retrieved using the commands prefetch and fasterq-dump provided by SRA Toolkit v.2.11.2 (https://github.com/ncbi/sra-tools). Only species comprising initially more than five RNA-seq samples were retained for quality control filtering and strandedness identification: *C. sativus*, *C. melo*, *C. lanatus*, *L. siceraria*, *B. hispida*, *C. moschata*, *C. argyrosperma*, *C. pepo*, *C. maxima*, *L. aegyptiaca*, *M. charantia*, and *S. grosvenorii*. First, we removed adapters and filtered reads by quality (average quality >20, window size = 4 bp) and length (length >49 bp) using fastp v.0.23.2 (Chen, Zhou, Chen, & Gu, 2018) to provide clean data for downstream analysis. FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Multiqc v.1.11 (Ewels, Magnusson, Lundin, & Käller, 2016) were used to perform quality control of clean data. Second, we identified the strandedness of the data and discarded nonstrand-specific or nonstranded RNA-seq samples. Strand-specific or stranded RNA-seq samples provide a more accurate estimate of transcript expression and allow the identification of more types of lncRNAs, such as natural antisense lncRNAs. To perform this step, we ran rsem-prepare-reference from the software package RSEM v.1.3.3 (Li & Dewey, 2011) to extract the transcriptomes of the analyzed species, and subsequently executed the pseudoaligner salmon v.1.6.0 (–l A) (Patro, Duggal, Love, Irizarry, & Kingsford, 2017) in mapping-based mode, which identifies the strandedness of the data. Then, only species comprising more than five cleaned and strand-specific RNA-seq samples were included in the subsequent analysis.

Once the species and their samples were preprocessed and selected, we used the splice-aware sequence alignment program HISAT2 v2.2.1 (Kim, Paggi, Park, Bennett, & Salzberg, 2019) for mapping the clean data to the corresponding reference genomes. In this step, strandedness information was considered (--rna-strandness <Strandedness salmon code>), and maximum intron length was set to 10 000 bp (--max-intronlen 10 000), which corresponds to approximate maximum intron length in plant genomes. Moreover, –-dta parameter was used to help transcript assemblers significantly improve computation and memory usage. Next, we reconstructed the transcriptome of each sample individually using the StringTie2 v2.2.0 assembler (Pertea et al., 2015). In particular, we performed a genome-guided assembly approach taking into account the strandedness information (--rf/--fr). All individual assembly gtf files produced by StringTie2 in the previous step were merged into a single and unified transcriptome for each species using the merge option of StringTie2 (-g 50 -F 0.3 -T 0). Finally, we compared the transcriptomes to the reference annotation files of each species to identify novel transcripts using the software GffCompare v.0.12.6 (Pertea & Pertea, 2020). According to their genomic location and referring to the neighboring protein-coding (PC) genes, this software classified the assembled transcripts and assigned them a class code.

Computational prediction of lncRNAs

Based on the class code annotation, we selected transcripts annotated as "u" (intergenic), "x" (antisense), "i" (intronic), and "o" or "e" (sense). Specifically, class code "u" refers to transcripts that come from intergenic regions of both genomic strands, class code "x" refers to transcripts that overlap with the exons of a PC gene on the opposite genomic strand, class code "i" refers to transcripts that are completely contained within the intron of a PC gene on the same genomic strand, and class codes "o" or "e" refer to transcripts that overlap with the exons of a PC gene on

the same genomic strand. Next, transcripts were filtered by length (length >200 bp), exon number (exon number ≥ 1) and expression level (Fragments Per Kilobase Million >0.3).

Further, we assessed the coding potential of the predicted transcripts using three alignment-free computational tools: CPC2 v.1.0.1 (Kang et al., 2017), FEELnc v.0.2 (Wucher et al., 2017), and CPAT v.3.0.2 (Wang et al., 2013) based on various intrinsic properties. For CPC2, transcripts assigned with the "noncoding" label were considered noncoding transcripts. For FEELnc, two training datasets, known PC genes and known lncRNAs, were required, and the shuffled mode was used to generate the later. The FEELnc cutoff to consider a transcript as coding or noncoding was defined using a tenfold cross-validation. For CPAT, we used the *Arabidopsis thaliana* logit model and hexamer frequency table obtained from the CREMA tool (Simopoulos, Weretilnyk, & Golding, 2018). The CPAT cutoff to consider a transcript as coding or non-coding was the default CREMA cutoff (0.5). To improve the robustness of the results, homologies were searched (on 18th of February 2022) with Swissprot (Anon, 2023) and Pfam-A databases (on 29th of May 2022) (Mistry et al., 2021). On the one hand, we ran the blastx command (−−strand plus --more-sensitive --top 5 --evalue 1e-5) from the program DIAMOND v.2.0.14 (Buchfink, Reuter, & Drost, 2021) to align the predicted transcripts to the protein sequences present in Swissprot. On the other hand, the TransDecoder.LongOrfs script (−m 20 -S) from the program transdecoder v.5.5.0 (https://github.com/TransDecoder/TransDecoder) was run to extract the longest Open Reading Frame (ORF) of each transcript and after that, the hmmsearch command (−E 1e-5 --domE 1e-5) from the software package HMMER v.2.0.14 (Eddy, 2011) was used to align these longest ORFs to the protein domains present in Pfam-A database. Then, transcripts were classified into three confidence levels according to the fulfillment of the following criteria: (i) High (HC): predicted as lncRNA by the three software (CPC2, CPAT, and FEELnc) and lacking homology with ORFs in the two protein databases (SwissProt and Pfam-A), (ii) Medium (MC): predicted as lncRNA by the three software but have similarity to one protein database or lack similarity to the two protein databases but are only predicted by two software and (iii) Low (LC): predicted as lncRNA by two software and have similarity to at least one protein database or lack similarity to the two protein databases but are only predicted by one software. Those transcripts that do not meet any of the present scenarios were not classified.

The next step of the procedure was performed to filter out housekeeping noncoding RNA classes such as ribosomal RNA (rRNA), transport RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and miRNA precursors (pre-miRNAs). To remove these housekeeping ncRNAs, we ran the blastn command (−strand plus -evalue 1e-5) from the NCBI-BLAST software v.2.13.0+ (Camacho et al., 2009) to align the remaining transcripts from the confidence-level classification step to RNAcentral (on 15th of June 2022) (Anon, 2021). Only ncRNAs from the *Cucurbitaceae* family were considered in the alignment. To remove pre-miRNAs, we repeated the previous step against miRBAse (on 1st of June 2022) (Kozomara, Birgaoanu, & Griffiths-Jones, 2019) and PmiREN (on 15th of June 2022) (Guo et al., 2020) using only pre-miRNAs from the analyzed species. Subsequently, significant hits were aligned to miRNA sequences from the analyzed species and from miRBase and PmiREN using blastn (−dust no -evalue 0.05 -word_size 7 -strand plus). Then, the program MIReNA v.2.0 (Mathelier & Carbone, 2010) was used to validate potential pre-miRNAs with information on potential miRNAs. In order to provide additional information, we annotated filtered transcripts using potential known lncRNAs databases such as CANTATAdb v.2.0 (Szcześniak et al., 2019), PLncDB v.2.0 (Jin et al., 2021), and GreeNC v.2.0 (Di Marsico et al., 2022), last access on 15th of June 2022. Only lncRNAs from the analyzed species were downloaded, and the program used to align the filtered transcripts to the different lncRNAs databases was blastn (−strand plus -evalue 1e-5). Only significant hits covering more than 50% of both aligned sequences were really considered hits. After that, when several isoforms were present, we kept only the longest one using the gtf2gtf command from the software CGAT v.1.0 (Sims et al., 2014). Finally, we created the database in table format as well as gtf annotation file format, and the different classes of potential lncRNAs were renamed from intergenic, antisense, intronic, and sense to lincRNA, NAT-lncRNA, int-lncRNA, and SOT-lncRNA, respectively.

Genomic distribution

Genome-wide distribution of potential lncRNAs and annotated PC genes was analyzed. Considering genomic density as the number of transcripts, we calculated the density per genomic window (size of 250 Kb) for the four classes of lncRNAs and PC genes using the R package circlize v.0.4.15. This information was displayed on a circular density plot using the same R package. After that, the percentage of the genome covered by each category (lncRNAs and PC genes) was obtained using bedtools genomecov v.2.27.1 (Quinlan & Hall, 2010). In addition, high-density regions (HDR) of the four classes of lncRNAs and PC genes were identified considering a window size of 100 Kb. In each case, the threshold was defined as (i) the mean genomic density across all windows, (ii) multiplied by 1.5, and (iii) rounded up. All windows above their corresponding threshold were considered HDRs. Finally, we calculated the percentage of HDRs of PC genes that overlap with HDRs of each class of lncRNA.

Molecular properties comparison

Potential lncRNAs and PC genes were compared using five features: GC content, exon number, length, expression level, and repeat content. The first three features were obtained previously during the creation of the database. Expression level was calculated using the pseudoaligner salmon in mapping-based mode to quantify both transcript categories considering strandedness information (−l <Strandedness salmon code>). The relative transcript abundance was estimated in units of Transcripts Per Million (TPM), normalized by transcript length and library size. To analyze, the repeat content repeat calling was performed using RepeatModeler v.2.0.3 (Flynn et al., 2020) and RepeatMasker v.4.1.3-p1 (Flynn et al., 2020) on the cucurbit genomes. Then, bedtools intersect was used to calculate the proportion of each transcript in both categories covered by the repeat regions found. In addition, a third category of sequences corresponding to random intergenic genome regions (500 bp in length) was included to assess the GC content and repeat content. This new category was obtained using bedtools random (-l 500 -n 25 000) on the cucurbit genomes, and bedtools intersected to select those random sequences that do not match any of the other two categories.

Evolutionary conservation analysis

We assessed evolutionary relationships between predicted lncRNAs across species at the sequence and positional level. To analyze the conservation at the sequence level, pairwise alignments between species were performed using BLASTn. To do this, we built a custom BLAST database for the set of lncRNAs of each species. Then, each set of lncRNAs was aligned against each database with BLASTn (−strand plus) using a relatively

nonstringent e-value threshold of 1e-5, and the best-matched query-target pair was selected. After that, orthologous gene-pairs were identified based on Reciprocal Best Hits (RBH), and subsequently we ran OrthoFinder v.2.5.4 (Emms & Kelly, 2019) based on the Markov Cluster (MCL) algorithm to infer the putative lncRNA orthologous families across species using RBHs information. To analyze the conservation at the positional level, a syntenic approach developed and validated in Pegueroles et al. (2019) was implemented. This computational strategy identifies lncRNAs from different species that share the same genomic context, that is, those surrounded by 1:1 orthologous PC genes. Therefore, we first ran OrthoFinder to infer high-quality 1:1 orthologous PC genes in the analyzed species using the peptide sequences downloaded from CuGenDB. Next, pairwise comparisons between species were performed to search for syntenic relationships between lncRNAs. Once the orthologous genes were established, we used a previously developed and validated approach (Pegueroles et al., 2019) to identify lncRNAs from different species that share a genomic context comparable using the previously identified 1:1 orthologous PC genes. The code was automated so that we could compare as many species as we wanted to, and avoid any problem related to species names. Regarding the parameters, we kept the same as previously to compare the genomic context of lncRNAs between two different species: (i) Consider three transcripts on each side of a given lncRNA, (ii) a minimum of overall three shared PC genes (orthologs) and, (iii) a minimum of one shared PC gene on each side of a given lncRNA. Finally, pairwise syntenic lncRNAs were classified into clusters across species called syntenic families.

In addition, MEME v.5.5.1 from the MEME Suite (Bailey, Johnson, Grant, & Noble, 2015) was used to identify conserved sequence motifs within the previously identified syntenic families. As parameters, we selected classic objective function, oops distribution, and a motif width between 6 and 50 nucleotides. Only motifs with an e-value <0.05 were considered significant. As previously (Hovhannisyan & Gabaldón, 2021), we assessed the differences between the identified number of syntenic families with shared motifs and the random expectation. To do this, we randomly chose lncRNAs to generate 50 simulated datasets with the same number and size of lncRNA families as the real dataset, preserving the number of lncRNAs per species observed in the real syntenic families. Then, we scanned the sequences of lncRNAs of these simulated families for motifs as described above.

Tissue-specificity analysis

The tissue-specificity analysis was performed individually per RNA-seq experiment. Therefore, we first selected those experiments that had at least three different tissues in order to increase the robustness of the analysis. Next, we generated a table corresponding to the mean expression level (in TPM) of each potential lncRNA and PC gene per tissue. To be stringent, those transcripts that did not have more than 1 TPM in any of the tissues were filtered out. Finally, we used the Python API of the tissue-specificity calculator tool tspex (https://github.com/apcamargo/tspex) to obtain the tissue-specificity metric Tau (Kryuchkova-Mostacci & Robinson-Rechavi, 2017), which describes in a single value how tissue-specific or ubiquitous is a gene across all tissues, ranging from 0 (broadly expressed genes) to 1 (highly tissue-specific genes).

Differential expression analysis in stress response and development

Given the regulatory role of lncRNAs in various developmental processes and stress responses (biotic and abiotic), all the projects present in this study related to both cases were selected for differential expression analysis (DEA). We first prepared the metadata to be able to carry out all pairwise comparisons (stress vs. control and late development stage vs. early development stage). Comparisons in which a condition had less than 2 replicates were discarded. Then, DEA was conducted by the R package DEseq2 (Love, Huber, & Anders, 2014), testing PC genes and potential lncRNAs. All p-values were adjusted by the false discovery rate (FDR), and only transcripts with an adjusted *P*-value ≤ 0.05 were considered differentially expressed.

Analysis of the correlation in the expression between PC genes and lncRNAs

To analyze if predicted lncRNAs and neighboring genes are co-expressed, we calculated the correlation in the expression of lincRNAs and PC genes with their closest upstream and/or downstream PC gene, and of NAT-lncRNAs with their cognate PC gene. In order to do this, we based our approach on the strategy used by Palos et al. (2022). Only gene expression datasets from the three most studied cucurbit species (*C. melo*, *C. sativus*, and *C. lanatus*) were used. These gene expression datasets, which contain all samples from a study's control group and its associated noncontrol samples, were filtered by low variance; that is, genes with a high median absolute deviation score were retained (top 25%). After that, genes were variance-stabilized in DESeq2 and filtered again to keep only genes present in the adjacent pairs previously created using bedtools closest v.2.27.1. Then, Pearson correlation coefficients (PCC) of expression were calculated using the R package stats. As a control, two sets of random gene pairs ($n = 1500$) from each gene expression dataset were generated from all pairwise correlations using the slice_sample function from the R package dplyr v.1.1.4. The first set of random gene pairs was composed of lincRNAs and PC genes, while the second set was composed of NAT-lncRNAs and PC genes of different strands to avoid strand bias.

To check if the distance of the gene pairs has an influence on the correlation, we generated all possible gene pairs (lincRNA-PC gene and PC gene-PC gene) within defined distances using bedtools intersect v.2.27.1 (Palos et al., 2022). The distances tested in this analysis were 1000, 2000, 5000, 10 000, 20 000, 50 000, and 10 000 bp of each other, and the Pearson correlation coefficients were calculated in the same way as in the previous analysis. In this case, random gene pairs were not created to avoid distance biases.

AUTHOR CONTRIBUTIONS

PVB Searched, downloaded, and curated the data. PVB Designed/adapted the pipeline and performed the bioinformatic analysis. PVB, JMM, and GG Designed the strategy, analyzed/discussed the results, and wrote the paper. GG Conceived the general idea and drafted the manuscript. All authors read, revise and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported by the Agencia Estatal de Investigacion (AEI) (co-supported by FEDER - EU) Grants PID2022-1393930B-I00. The funders had no role in the experiment design, data analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

All the authors declare no conflict interests.

DATA AVAILABILITY STATEMENT

All necessary components, including scripts, software versions, and additional files needed to replicate the study findings, can be found on our GitHub page (https://github.com/ncRNA-lab/Cucurbit_lncRNAs_landscape). The gtf annotation files of the identified lncRNAs are also available (https://github.com/ncRNA-lab/Cucurbit_lncRNAs_landscape/Tables/LncRNAs). Accessions to the analyzed RNA-seq datasets are detailed as Supporting material (Table S1).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Genome coverage of the assembled transcripts in the family *Cucurbitaceae*.

Figure S2. Analysis of linear correlation between intergenic regions and lincRNAs.

Figure S3. Genomic distribution of protein-coding genes and the predicted lncRNAs in cucurbits.

Figure S4. Genome coverage of protein-coding genes and the predicted lncRNAs in the family *Cucurbitaceae*.

Figure S5. High-density regions of protein-coding genes and the predicted lncRNAs across cucurbit genomes.

Figure S6. Molecular properties of the lncRNAs identified in each cucurbit species.

Figure S7. Molecular properties of the four lncRNA classes identified in each cucurbit species.

Figure S8. Predicted lncRNAs exhibit lower sequence conservation.

Figure S9. Detailed information about syntenic relationships of lncRNAs positionally conserved in seven to nine cucurbit species.

Figure S10. Syntenic relationships of lncRNAs positionally conserved in two or three cucurbit species.

Figure S11. Syntenic relationships of lncRNAs positionally conserved in four to six cucurbit species.

Figure S12. Syntenic lncRNAs exhibit modular structure.

Figure S13. Expression of tissue-specific lncRNAs across tissues.

Figure S14. Expression of the identified lncRNAs per cucurbit species considering different tissues, developmental stages, and stress conditions.

Figure S15. LincRNA/NAT-lncRNA influence on nearby/cognate gene expression in each cucurbit species.

Figure S16. Positive correlation between a syntenic HC-NAT-lncRNA and its cognate protein-coding gene related to phloem development.

Table S1. Detailed description of the downloaded cucurbit dataset during data recovery and pre-processing steps.

Table S2. Detailed description and global landscape of the lncRNAs predicted in cucurbits.

Table S3. Detailed information about the genome covered by predicted lncRNAs and protein-coding genes in the family *Cucurbitaceae*.

Table S4. Detailed information about the molecular properties of the lncRNAs identified in cucurbits.

Table S5. Detailed information about lncRNA conservation at sequence, positional and motif level in the family *Cucurbitaceae*.

Table S6. Details of the dataset used for expression analysis in different tissues, developmental stages and stress conditions.

Table S7. Detailed information about the cucurbit lncRNA expression associated to plant tissues, developmental stages, and stress conditions.

Table S8. Detailed information about the expression correlation between neighboring genes.

Table S9. Summary about all the characteristics of the lncRNAs predicted in the nine species of cucurbits.

REFERENCES

Anon. (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, **49**, D212–D220. Available from: https://doi.org/10.1093/nar/gkaa921

Anon. (2023) UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, **51**, D523–D531. Available from: https://doi.org/10.1093/nar/gkac1052

Aslam, M.M., Waseem, M., Xu, W., Ying, L., Zhang, J. & Yuan, W. (2022) Global identification of white lupin lncRNAs reveals their role in cluster roots under phosphorus deficiency. *International Journal of Molecular Sciences*, **23**, 9012.

Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. (2015) The MEME suite. *Nucleic Acids Research*, **43**, W39–W49. Available from: https://pubmed.ncbi.nlm.nih.gov/25953851

Bhowmick, B.K. & Jha, S. (2015) Dynamics of sex expression and chromosome diversity in Cucurbitaceae: a story in the making. *Journal of Genetics*, **94**, 793–808. Available from: https://doi.org/10.1007/s12041-015-0562-5

Buchfink, B., Reuter, K. & Drost, H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, **18**, 366–368. Available from: https://pubmed.ncbi.nlm.nih.gov/33828273

Budak, H., Kaya, S.B. & Cagirici, H.B. (2020) Long non-coding RNA in plants in the era of reference sequences. *Frontiers in Plant Science*, **11**, 276. Available from: https://pubmed.ncbi.nlm.nih.gov/32226437

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421. Available from: https://pubmed.ncbi.nlm.nih.gov/20003500

Chen, L., Zhu, Q.-H. & Kaufmann, K. (2020) Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. *Planta*, **252**, 92. Available from: https://doi.org/10.1007/s00425-020-03480-5

Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Chomicki, G., Schaefer, H. & Renner, S.S. (2020) Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *The New Phytologist*, **226**, 1240–1255. Available from: https://doi.org/10.1111/nph.16015

Crespi, M.D., Jurkevitch, E., Poiret, M., d'Aubenton-Carafa, Y., Petrovics, G., Kondorosi, E. *et al.* (1994) enod40, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. *The EMBO Journal*, **13**, 5099–5112. Available from: https://doi.org/10.1002/j.1460-2075.1994.tb06839.x

Csorba, T., Questa, J.I., Sun, Q. & Dean, C. (2014) Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 16160–16165. Available from: https://pubmed.ncbi.nlm.nih.gov/25349421

Cui, J., Luan, Y., Jiang, N., Bao, H. & Meng, J. (2017) Comparative transcriptome analysis between resistant and susceptible tomato allows the identification of lnc RNA 16397 conferring resistance to Phytophthora infestans by co-expressing glutaredoxin. *The Plant Journal*, **89**, 577–589.

Datta, R. & Paul, S. (2019) Long non-coding RNAs: fine-tuning the developmental responses in plants. *Journal of Biosciences*, **44**, 1–11.

Dey, S.S., Sharma, P.K., Munshi, A.D., Jaiswal, S., Behera, T.K., Kumari, K. *et al.* (2022) Genome wide identification of lncRNAs and circRNAs having regulatory role in fruit shelf life in health crop cucumber (Cucumis sativus L.). *Frontiers in Plant Science*, **13**, 884476. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2022.884476

Di Marsico, M., Paytuvi Gallart, A., Sanseverino, W. & Aiese Cigliano, R. (2022) GreeNC 2.0: a comprehensive database of plant long non-coding

RNAs. *Nucleic Acids Research*, **50**, D1442–D1447. Available from: https://doi.org/10.1093/nar/gkab1014

Domínguez-Rosas, E., **Hernández-Oñate, M.Á.**, **Fernandez-Valverde, S.-L.** & **Tiznado-Hernández, M.E.** (2023) Plant long non-coding RNAs: identification and analysis to unveil their physiological functions. *Frontiers in Plant Science*, **14**, 1275399.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195. Available from: https://doi.org/10.1371/journal.pcbi.1002195

Emms, D.M. & **Kelly, S.** (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 238. Available from: https://doi.org/10.1186/s13059-019-1832-y

Engreitz, J.M., **Ollikainen, N.** & **Guttman, M.** (2016) Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nature Reviews. Molecular Cell Biology*, **17**, 756–770. Available from: https://doi.org/10.1038/nrm.2016.126

Ewels, P., **Magnusson, M.**, **Lundin, S.** & **Käller, M.** (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048. Available from: https://pubmed.ncbi.nlm.nih.gov/27312411

Flynn, J.M., **Hubley, R.**, **Goubert, C.**, **Rosen, J.**, **Clark, A.G.**, **Feschotte, C.** ***et al.*** (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, **117**, 9451–9457. Available from: https://pubmed.ncbi.nlm.nih.gov/32300014

Franco-Zorrilla, J.M., **Valli, A.**, **Todesco, M.**, **Mateos, I.**, **Puga, M.I.**, **Rubio-Somoza, I.** ***et al.*** (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, **39**, 1033–1037.

Froelich, D.R., **Mullendore, D.L.**, **Jensen, K.H.**, **Ross-Elliott, T.J.**, **Anstead, J.A.**, **Thompson, G.A.** ***et al.*** (2011) Phloem ultrastructure and pressure flow: Sieve-element-Occlusion-related agglomerations do not affect translocation. *Plant Cell*, **23**, 4428–4445. Available from: http://www.jstor.org/stable/41433932

Gao, C., **Sun, J.**, **Dong, Y.**, **Wang, C.**, **Xiao, S.**, **Mo, L.** ***et al.*** (2020) Comparative transcriptome analysis uncovers regulatory roles of long non-coding RNAs involved in resistance to powdery mildew in melon. *BMC Genomics*, **21**, 125. Available from: https://doi.org/10.1186/s12864-020-6546-8

Gil, N. & **Ulitsky, I.** (2020) Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews. Genetics*, **21**, 102–117. Available from: https://doi.org/10.1038/s41576-019-0184-5

Gómez, G., **Torres, H.** & **Pallás, V.** (2005) Identification of translocatable RNA-binding phloem proteins from melon, potential components of the long-distance RNA transport system. *The Plant Journal*, **41**, 107–116.

Graf, J. & **Kretz, M.** (2020) From structure to function: route to understanding lncRNA mechanism. *BioEssays*, **42**, 2000027. Available from: https://doi.org/10.1002/bies.202000027

Guo, J., **Xu, W.**, **Hu, Y.**, **Huang, J.**, **Zhao, Y.**, **Zhang, L.** ***et al.*** (2020) Phylotranscriptomics in Cucurbitaceae reveal multiple whole-genome duplications and key morphological and molecular innovations. *Molecular Plant*, **13**, 1117–1133. Available from: https://www.sciencedirect.com/science/article/pii/S1674205220301465

Guo, Z., **Kuang, Z.**, **Wang, Y.**, **Zhao, Y.**, **Tao, Y.**, **Cheng, C.** ***et al.*** (2020) PmiREN: a comprehensive encyclopedia of plant miRNAs. *Nucleic Acids Research*, **48**, D1114–D1121. Available from: http://www.bioinformatics.babraham

Guttman, M. & **Rinn, J.L.** (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346. Available from: https://pubmed.ncbi.nlm.nih.gov/22337053

He, X., **Guo, S.**, **Wang, Y.**, **Wang, L.**, **Shu, S.** & **Sun, J.** (2020) Systematic identification and analysis of heat-stress-responsive lncRNAs, circRNAs and miRNAs with associated co-expression and ceRNA networks in cucumber (Cucumis sativus L.). *Physiologia Plantarum*, **168**, 736–754. Available from: https://doi.org/10.1111/ppl.12997

Hovhannisyan, H. & **Gabaldón, T.** (2021) The long non-coding RNA landscape of Candida yeast pathogens. *Nature Communications*, **12**, 7317. Available from: https://doi.org/10.1038/s41467-021-27635-4

Huanca-Mamani, W., **Arias-Carrasco, R.**, **Cardenas-Ninasivincha, S.**, **Rojas-Herrera, M.**, **Sepúlveda-Hermosilla, G.**, **Caris-Maldonado, J.C.** ***et al.*** (2018) Long non-coding RNAs responsive to salt and boron stress in the hyper-arid Lluteno maize from Atacama Desert. *Genes (Basel)*, **9**, 170.

Huang, J., **Zhou, W.**, **Zhang, X.** & **Li, Y.** (2023) Roles of long non-coding RNAs in plant immunity. *PLoS Pathogens*, **19**, e1011340. Available from: https://doi.org/10.1371/journal.ppat.1011340

Jachowicz, J.W., **Strehle, M.**, **Banerjee, A.K.**, **Blanco, M.R.**, **Thai, J.** & **Guttman, M.** (2022) Xist spatially amplifies SHARP/SPEN recruitment to balance chromosome-wide silencing and specificity to the X chromosome. *Nature Structural & Molecular Biology*, **29**, 239–249.

Jha, U.C., **Nayyar, H.**, **Jha, R.**, **Khurshid, M.**, **Zhou, M.**, **Mantri, N.** ***et al.*** (2020) Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation. *BMC Plant Biology*, **20**, 466. Available from: https://doi.org/10.1186/s12870-020-02595-x

Jin, J., **Lu, P.**, **Xu, Y.**, **Li, Z.**, **Yu, S.**, **Liu, J.** ***et al.*** (2021) PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. *Nucleic Acids Research*, **49**, D1489–D1495. Available from: https://doi.org/10.1093/nar/gkaa910

Kang, Y.J., **Yang, D.C.**, **Kong, L.**, **Hou, M.**, **Meng, Y.Q.**, **Wei, L.** ***et al.*** (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, **45**, W12–W16.

Kęska, K., **Szcześniak, M.W.**, **Adamus, A.** & **Czernicka, M.** (2021) Waterlogging-stress-responsive LncRNAs, their regulatory relationships with miRNAs and target genes in cucumber (Cucumis sativus L.). *International Journal of Molecular Sciences*, **22**, 8197.

Kim, D., **Paggi, J.M.**, **Park, C.**, **Bennett, C.** & **Salzberg, S.L.** (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915. Available from: https://doi.org/10.1038/s41587-019-0201-4

Kirk, J.M., **Kim, S.O.**, **Inoue, K.**, **Smola, M.J.**, **Lee, D.M.**, **Schertzer, M.D.** ***et al.*** (2018) Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, **50**, 1474–1482.

Kozomara, A., **Birgaoanu, M.** & **Griffiths-Jones, S.** (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Research*, **47**, D155–D162. Available from: https://doi.org/10.1093/nar/gky1141

Kryuchkova-Mostacci, N. & **Robinson-Rechavi, M.** (2017) A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, **18**, 205–214. Available from: https://pubmed.ncbi.nlm.nih.gov/26891983

Leinonen, R., **Sugawara, H.**, **Shumway, M.** & **Collaboration, I.N.S.D.** (2011) The sequence read archive. *Nucleic Acids Research*, **39**, D19–D21. Available from: https://pubmed.ncbi.nlm.nih.gov/21062823

Li, B. & **Dewey, C.N.** (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323. Available from: https://doi.org/10.1186/1471-2105-12-323

Liu, X., **Li, D.**, **Zhang, D.**, **Yin, D.**, **Zhao, Y.**, **Ji, C.** ***et al.*** (2018) A novel antisense long noncoding RNA, TWISTED LEAF, maintains LEAF blade flattening by regulating its associated sense R2R3-MYB gene in rice. *The New Phytologist*, **218**, 774–788.

Lough, T.J. & **Lucas, W.J.** (2006) Integrative plant biology: role of phloem long-distance macromolecular trafficking. *Annual Review of Plant Biology*, **57**, 203–232. Available from: https://doi.org/10.1146/annurev.arplant.56.032604.144145

Love, M.I., **Huber, W.** & **Anders, S.** (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550. Available from: https://doi.org/10.1186/s13059-014-0550-8

Lucero, L., **Ferrero, L.**, **Fonouni-Farde, C.** & **Ariel, F.** (2021) Functional classification of plant long noncoding RNAs: a transcript is known by the company it keeps. *The New Phytologist*, **229**, 1251–1260. Available from: https://doi.org/10.1111/nph.16903

Marquardt, S., **Raitskin, O.**, **Wu, Z.**, **Liu, F.**, **Sun, Q.** & **Dean, C.** (2014) Functional consequences of splicing of the antisense transcript COOLAIR on FLC transcription. *Molecular Cell*, **54**, 156–165. Available from: https://www.sciencedirect.com/science/article/pii/S1097276514002597

Martínez, C. & **Jamilena, M.** (2021) To be a male or a female flower, a question of ethylene in cucurbits. *Current Opinion in Plant Biology*, **59**, 101981. Available from: https://www.sciencedirect.com/science/article/pii/S1369526620301394

Martinez, G., **Castellano, M.**, **Tortosa, M.**, **Pallas, V.** & **Gomez, G.** (2014) A pathogenic non-coding RNA induces changes in dynamic DNA methylation of ribosomal RNA genes in host plants. *Nucleic Acids Research*, **42**, 1553–1562.

Mathelier, A. & **Carbone, A.** (2010) MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing

data. *Bioinformatics*, **26**, 2226–2234. Available from: https://doi.org/10.1093/bioinformatics/btq329

Mattick, J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genetics*, **5**, e1000459. Available from: https://doi.org/10.1371/journal.pgen.1000459

Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.L. *et al.* (2023) Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews. Molecular Cell Biology*, **24**, 430–447. Available from: https://doi.org/10.1038/s41580-022-00566-8

Meena, S.K., Heidecker, M., Engelmann, S., Jaber, A., de Vries, T., Triller, S. *et al.* (2023) Altered expression levels of long non-coding natural antisense transcripts overlapping the UGT73C6 gene affect rosette size in Arabidopsis thaliana. *The Plant Journal*, **113**, 460–477.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Research*, **49**, D412–D419.

Numan, M., Sun, Y. & Li, G. (2024) Exploring the emerging role of long non-coding RNAs (lncRNAs) in plant biology: functions, mechanisms of action, and future directions. *Plant Physiology and Biochemistry*, **212**, 108797. Available from: https://www.sciencedirect.com/science/article/pii/S0981942824004650

Palos, K., Nelson Dittrich, A.C., Yu, L. *et al.* (2022) Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. *Plant Cell*, **34**, 3233–3260.

Palos, K., Yu, L., Railey, C.E., Nelson Dittrich, A.C. & Nelson, A.D.L. (2023) Linking discoveries, mechanisms, and technologies to develop a clearer perspective on plant long noncoding RNAs. *Plant Cell*, **35**, 1762–1786. Available from: https://doi.org/10.1093/plcell/koad027

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**, 417–419. Available from: https://doi.org/10.1038/nmeth.4197

Pech, J.C., Bouzayen, M. & Latché, A. (2008) Climacteric fruit ripening: ethylene-dependent and independent regulation of ripening pathways in melon fruit. *Plant Science*, **175**, 114–120. Available from: https://www.sciencedirect.com/science/article/pii/S0168945208000095

Pegueroles, C., Iraola-Guzmán, S., Chorostecki, U., Ksiezopolska, E., Saus, E. & Gabaldón, T. (2019) Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis. *RNA Biology*, **16**, 320–329. Available from: https://pubmed.ncbi.nlm.nih.gov/30691342

Pertea, G. & Pertea, M. (2020) GFF utilities: GffRead and GffCompare. *F1000Research*, **9**, ISCB Comm J-304. Available from: https://pubmed.ncbi.nlm.nih.gov/32489650

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. & Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295. Available from: https://doi.org/10.1038/nbt.3122

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ranjan, J.K., Pandey, S., Prgaya, Akhter Ansari, W., Krishna, R., Tarique Zeyad, M. *et al.* (2022) In: Kole, C. (Ed.) *Biotic Stresses in Cucurbits: Status, Challenges, Breeding and Genetic Tools to Enhance Resistance BT - Genomic Designing for Biotic Stress Resistant Vegetable Crops*. Cham: Springer International Publishing, pp. 345–379. Available from: https://doi.org/10.1007/978-3-030-97785-6_9

Roulé, T., Crespi, M. & Blein, T. (2021) Regulatory long non-coding RNAs in root growth and development. *Biochemical Society Transactions*, **50**, 403–412. Available from: https://doi.org/10.1042/BST20210743

Simopoulos, C.M.A., Weretilnyk, E.A. & Golding, G.B. (2018) Prediction of plant lncRNA by ensemble machine learning classifiers. *BMC Genomics*, **19**, 316. Available from: https://pubmed.ncbi.nlm.nih.gov/29720103

Sims, D., Ilott, N.E., Sansom, S.N., Sudbery, I.M., Johnson, J.S., Fawcett, K.A. *et al.* (2014) CGAT: computational genomics analysis toolkit. *Bioinformatics*, **30**, 1290–1291. Available from: https://doi.org/10.1093/bioinformatics/btt756

Szcześniak, M.W., Bryzghalov, O., Ciomborowska-Basheer, J. & Makałowska, I. (2019) In: Chekanova, J.A. & Wang, H.-L.V. (Eds.) *CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs BT - Plant Long Non-Coding RNAs: Methods and Protocols*. New York, NY: Springer New York, pp. 415–429. Available from: https://doi.org/10.1007/978-1-4939-9045-0_26

Tian, J., Zhang, G., Zhang, F., Ma, J., Wen, C. & Li, H. (2022) Genome-wide identification of powdery mildew responsive long non-coding RNAs in Cucurbita pepo. *Frontiers in Genetics*, **13**, 933022. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2022.933022

Tian, Y., Bai, S., Dang, Z., Hao, J., Zhang, J. & Hasi, A. (2019) Genome-wide identification and characterization of long non-coding RNAs involved in fruit ripening and the climacteric in Cucumis melo. *BMC Plant Biology*, **19**, 1–15.

Traubenik, S., Charon, C. & Blein, T. (2024) From environmental responses to adaptation: the roles of plant lncRNAs. *Plant Physiology*, **kiae034**, 232–244.

Wang, H., Chung, P.J., Liu, J., Jang, I.-C., Kean, M.J., Xu, J. *et al.* (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Research*, **24**, 444–453. Available from: https://pubmed.ncbi.nlm.nih.gov/24402519

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P. & Li, W. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research*, **41**, 1–7.

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D. *et al.* (2014) Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proceedings of the National Academy of Sciences*, **111**, 10359–10364.

Wierzbicki, A.T., Blevins, T. & Swiezewski, S. (2021) Long noncoding RNAs in plants. *Annual Review of Plant Biology*, **72**, 245–271.

Wierzbicki, A.T., Haag, J.R. & Pikaard, C.S. (2008) Noncoding transcription by RNA polymerase pol IVb/pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell*, **135**, 635–648.

Wu, J., Okada, T., Fukushima, T., Tsudzuki, T., Sugiura, M. & Yukawa, Y. (2012) A novel hypoxic stress-responsive long non-coding RNA transcribed by RNA polymerase III in Arabidopsis. *RNA Biology*, **9**, 302–313. Available from: https://doi.org/10.4161/rna.19101

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T. *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, **45**, 1–12.

Xu, X.-W., Zhou, X.-H., Wang, R.-R., Peng, W.-L., An, Y. & Chen, L.-L. (2016) Functional analysis of long intergenic non-coding RNAs in phosphate-starved rice using competing endogenous RNA network. *Scientific Reports*, **6**, 20715. Available from: https://doi.org/10.1038/srep20715

Yamada, M. (2017) Functions of long intergenic non-coding (linc) RNAs in plants. *Journal of Plant Research*, **130**, 67–73. Available from: https://doi.org/10.1007/s10265-016-0894-0

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659. Available from: https://doi.org/10.1093/bioinformatics/bti042

Yang, H., Cui, Y., Feng, Y., Hu, Y., Liu, L. & Duan, L. (2023) Long non-coding RNAs of plants in response to abiotic stresses and their regulating roles in promoting environmental adaption. *Cells*, **12**, 729. Available from: https://pubmed.ncbi.nlm.nih.gov/36899864

Yang, J., Ariel, F. & Wang, D. (2023) Plant long non-coding RNAs: biologically relevant and mechanistically intriguing. *Journal of Experimental Botany*, **74**, 2364–2373. Available from: https://doi.org/10.1093/jxb/erac482

Zhao, X., Li, J., Lian, B., Gu, H., Li, Y. & Qi, Y. (2018) Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. *Nature Communications*, **9**, 5056. Available from: https://pubmed.ncbi.nlm.nih.gov/30498193

Zheng, Y., Wu, S., Bai, Y., Sun, H., Jiao, C., Guo, S. *et al.* (2019) Cucurbit genomics database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Research*, **47**, D1128–D1136.

Zhou, X., Cui, J., Cui, H., Jiang, N., Hou, X., Liu, S. *et al.* (2020) Identification of lncRNAs and their regulatory relationships with target genes and corresponding miRNAs in melon response to powdery mildew fungi. *Gene*, **735**, 144403. Available from: https://www.sciencedirect.com/science/article/pii/S037811192030072X

Zhu, J.-K. (2016) Abiotic stress signaling and responses in plants. *Cell*, **167**, 313–324. Available from: https://pubmed.ncbi.nlm.nih.gov/27716505