Regulatory annotation identifies KAN1, MYB44 and PIF4 as regulators of Arabidopsis lincRNAs expressed in root

Li Liu[1,2], Michel Heidecker[3,4], Thomas Depuydt[1,2], Nicolas Manosalva Perez[1,2], Martin Crespi[3,4], Thomas Blein[3,4*], Klaas Vandepoele[1,2,5*]

(1) Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 71, 9052 Ghent, Belgium

(2) VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

(3) Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif-sur-Yvette, France

(4) Université Paris Cité, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif-sur-Yvette, France

(5) Bioinformatics Institute Ghent, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

* Shared last author

Corresponding author: klaas.vandepoele@psb.vib-ugent.be

Running title: TF regulatory annotation of Arabidopsis root lincRNAs

ABSTRACT

Thousands of long intergenic noncoding RNAs (lincRNAs) have been identified in plant genomes. While some lincRNAs have been characterized as important regulators in different biological processes, little is known about the transcriptional regulation for most plant lincRNAs. Through the integration of eight annotation resources, we defined 6,599 high-confidence lincRNA loci in *Arabidopsis thaliana*. For lincRNAs belonging to different evolutionary age categories, we identified major differences in sequence and chromatin features, as well as in the level of conservation and purifying selection acting during evolution. Spatiotemporal gene expression profiles combined with transcription factor (TF) chromatin immunoprecipitation data were used to construct a TF-lincRNA regulatory network containing 2,659 lincRNAs and 15,686 interactions. We found that properties characterizing lincRNA expression, conservation and regulation differ between plants and animals. Experimental validation confirmed the role of three TFs, KAN1, MYB44, and PIF4, as key regulators controlling root-specific lincRNA expression, demonstrating the predictive power of our network. Furthermore, we identified 58 lincRNAs, regulated by these TFs, showing strong root cell-type specific expression or chromatin accessibility, which are linked with GWAS genetic associations related to root system development and growth. The multi-level genome-wide characterization covering chromatin state information, promoter conservation, and ChIP-based TF binding, for all detectable lincRNAs across 769 expression samples, permits to rapidly define the biological context and relevance of lincRNAs in Arabidopsis regulatory networks.

One-line sentence:

A multi-level Arabidopsis gene regulatory network identifies novel regulators controlling root-specific lincRNA expression, offering a promising strategy to identify lincRNAs involved in plant biology.

Introduction

Genomes are widely transcribed and produce thousands of long non-coding RNAs (lncRNAs), which are an abundant class of transcripts longer than 200 nucleotides with low protein coding capacity (Wu et al., 2017). LncRNAs are generally transcribed by RNA polymerase II (Pol II), are poly-adenylated and may contain introns as coding RNA. LncRNAs modulate gene expression through a wide range of mechanisms, including chromatin structure remodeling, transcription regulation in cis/trans, fine-tuning of miRNA activity, alternative splicing (AS) regulation, and the control of mRNA stability and translation (Sanchita et al., 2020; Bhogireddy et al., 2021; Lucero et al., 2021). One class of lncRNAs is the primary mRNAs containing miRNA precursors or pri-miRNAs. These are rapidly and generally processed into miRNAs (Jones-Rhoades et al., 2006; Shafiq et al., 2016). However, the large majority of lncRNAs are able to act without being further processed such as the lncRNAs controlling the epigenetic regulation the *Flowering Locus C* (*FLC*) gene expression and mediating plant vernalization, i.e., *COOLAIR*, *COLDAIR* and *COLDWRAP* (Liu et al., 2010; Heo and Sung, 2011; Kim and Sung, 2017). A subgroup of lncRNAs derived from intergenic regions is defined as long intergenic non-coding RNAs (lincRNAs) and have been identified in a wide range of eukaryotes including model and non-model plant species (Wu et al., 2017; Chen et al., 2021). In contrast to antisense lncRNAs, whose sequence evolution is constrained by the overlapping coding genes, the transcription and evolution of lincRNAs are independent of the surrounding genes.

While the low expression levels and tissue-specific expression patterns of lincRNAs in plants initially raised concerns (Liu et al., 2012; Bu et al., 2015), increasing experimental evidence supports the functional activity of lincRNAs. In plants, few lincRNAs have been experimentally validated (Chen et al., 2021), showing their involvement in various biological contexts such as in regulating flowering time (Chen et al., 2020) and root growth and development (Roule et al., 2022a), or influencing germination (Kramer et al., 2022). For example, the Arabidopsis *FLINC* lincRNA has been reported to regulate ambient temperature-

mediated flowering time (Severing et al., 2018). Arabidopsis lateral root development is regulated by the *Alternative Splicing COmpetitor* (*ASCO*) lincRNA, which modulates AS by interacting with the multiple splicing factors (Bardou et al., 2014; Rigo et al., 2020), and the *AUXIN-REGULATED PROMOTER LOOP* (*APOLO*) lincRNA, which influences the local chromatin conformation and the activity of several Auxin-Responsive genes (Ariel et al., 2020). Plants lacking *CONSERVED INBRASSICA RAPA1* (*lncCOBRA1*) were found to show a delayed germination and were overall smaller than wild-type plants (Kramer et al., 2022). Many lincRNAs are differentially expressed in various stress responses, including drought (Qi et al., 2013; Shuai et al., 2014; Li et al., 2017; Qin et al., 2017), cold (Li et al., 2017; Zhao et al., 2018; Shea et al., 2019), salinity (Deng et al., 2018; Fukuda et al., 2019), and nutrient deficiency (Fukuda et al., 2019), implying that lincRNAs may be involved in plant stress resilience (Jha et al., 2020). Interestingly, some of the confirmed functional lincRNAs interact with transcription factors (TFs) to activate or repress the expression of associated genes, such as *APOLO* that interacts with WRKY42 to form a regulatory hub that controls the activity of *RHD6* and promotes the expansion of root hair cell at low temperatures (Moison et al., 2021; Pacheco et al., 2021).

Recently, a lot of attention has been placed on the evolutionary conservation of lincRNAs, which is generally associated with functionality (Ulitsky, 2016; Szczesniak et al., 2021). The conservation of noncoding transcripts can be considered at the level of the primary sequence, position, splice sites, RNA structure, and transcriptional level (Ulitsky, 2016; Szczesniak et al., 2021). However, most lincRNA sequences undergo rapid evolution and are poorly conserved (Ransohoff et al., 2018). In a study by Wang et al., only 5% of 2,281 rice lincRNAs had sequence similarity to maize lincRNAs. It was also found that the positional conservation of lincRNAs was much higher than their sequence conservation (Wang et al., 2015). Nelson and co-workers reported that 22% of 1180 Arabidopsis lincRNA loci were conserved in 10 Brassicaceae genomes (Nelson et al., 2016). These conserved lincRNAs exhibited higher expression levels, stress-responsiveness and their gene body overlapped with conserved

noncoding sequences (CNSs), suggesting a role of their conserved sequence in a genomic context (Nelson et al., 2016).

While different studies have reported on the identification and expression of lincRNAs in plants (Wang et al., 2015; Nelson et al., 2016; Ke et al., 2019; He et al., 2021), a comprehensive overview of the different genomic features contributing to the expression, regulation and evolutionary conservation of plant lincRNAs is missing. How lincRNAs are embedded in transcriptional networks controlling different biological processes remains largely unknown. Furthermore, prioritizing lincRNAs for downstream functional analysis is not straightforward without knowing the regulatory network where they are involved in. Here, we integrated different Arabidopsis lincRNAs gene annotations and explored various functional genomics datasets to characterize lincRNA expression in a biologically relevant context. We leveraged large-scale expression datasets and protein-DNA interaction data to study the molecular networks controlling lincRNA gene activity. Combined with evolutionary conservation analysis, we explored the global transcriptional regulatory properties of different evolutionary age categories and, through regulatory network analysis, identified specific TFs controlling lincRNA regulation in roots.

RESULTS

Integration of lincRNA annotations in *Arabidopsis*

A substantial number of lncRNA transcripts in *Arabidopsis* have been identified and several publicly available resources for the annotation of lncRNAs have been developed (Jha et al., 2020). In contrast to antisense lncRNAs which generally regulate the overlapping coding gene, much less is known about the potential targets of lincRNAs, so we focus our study on these transcripts. Indeed, lincRNAs transcription and evolution are independent of the surrounding genes, in contrast to antisense lncRNAs that are constrained by the coding genes they overlap with. To integrate and unify previously identified lincRNAs, annotations based on transcriptome information from 10 different tissues and various

environmental conditions were collected from different resources including Araport11 (Cheng et al., 2017), CANTATAdb (Szczesniak et al., 2016), NONCODEv5 (Fang et al., 2018), PLNlncRbase (Xuan et al., 2015), key lncRNA research articles (Liu et al., 2012; Nelson et al., 2017; Zhao et al., 2018), and new predictions based on root related stranded RNA-seq (Material and Methods). Next, a pipeline was designed to define a unified set of high-confidence lincRNAs by discarding transcripts with length below 200 bp, removing transcripts that overlapped with protein-coding genes (antisense lncRNAs), re-evaluating the coding potential of the transcripts, and merging the remaining transcripts from various resources (see Materials and Methods, Supplemental Figure S1A-B). In total, we identified 7,706 lincRNA transcripts covering 6,599 high-confidence lincRNA loci (see Supplemental Data Set S1). To explore the overlap of this high-confidence lincRNA gene set with the individual annotations, we assessed the overlap between the different resources (Figure 1A). A total of 4,955 (75.1%) lincRNA loci were supported by only one resource and the remaining 1,644 (24.9%) lincRNA loci were derived from two or more resources (Figure 1B). Araport11 contained the highest number of shared loci and Liu et al. (2012) contained the highest number of unique loci (Chen et al., 2021). Next, the genomic features of the high-confidence lincRNA transcripts were compared to those of protein-coding transcripts. 6,428 (83%) lincRNA transcripts and 6,250 (13%) protein-coding transcripts contained single exons, while 1,278 (17%) lincRNA transcripts and 42,109 (87%) protein-coding transcripts contained multiple exons. Furthermore, a higher frequency of multi-exon transcripts was found in lincRNAs supported by two or more resources (974, 36%) than in those supported by a unique resource (304, 6%) (Figure 1C). The transcript length distribution for lincRNAs showed a U-shape curve with the majority of transcripts being 200-300bp long (Figure 1D).

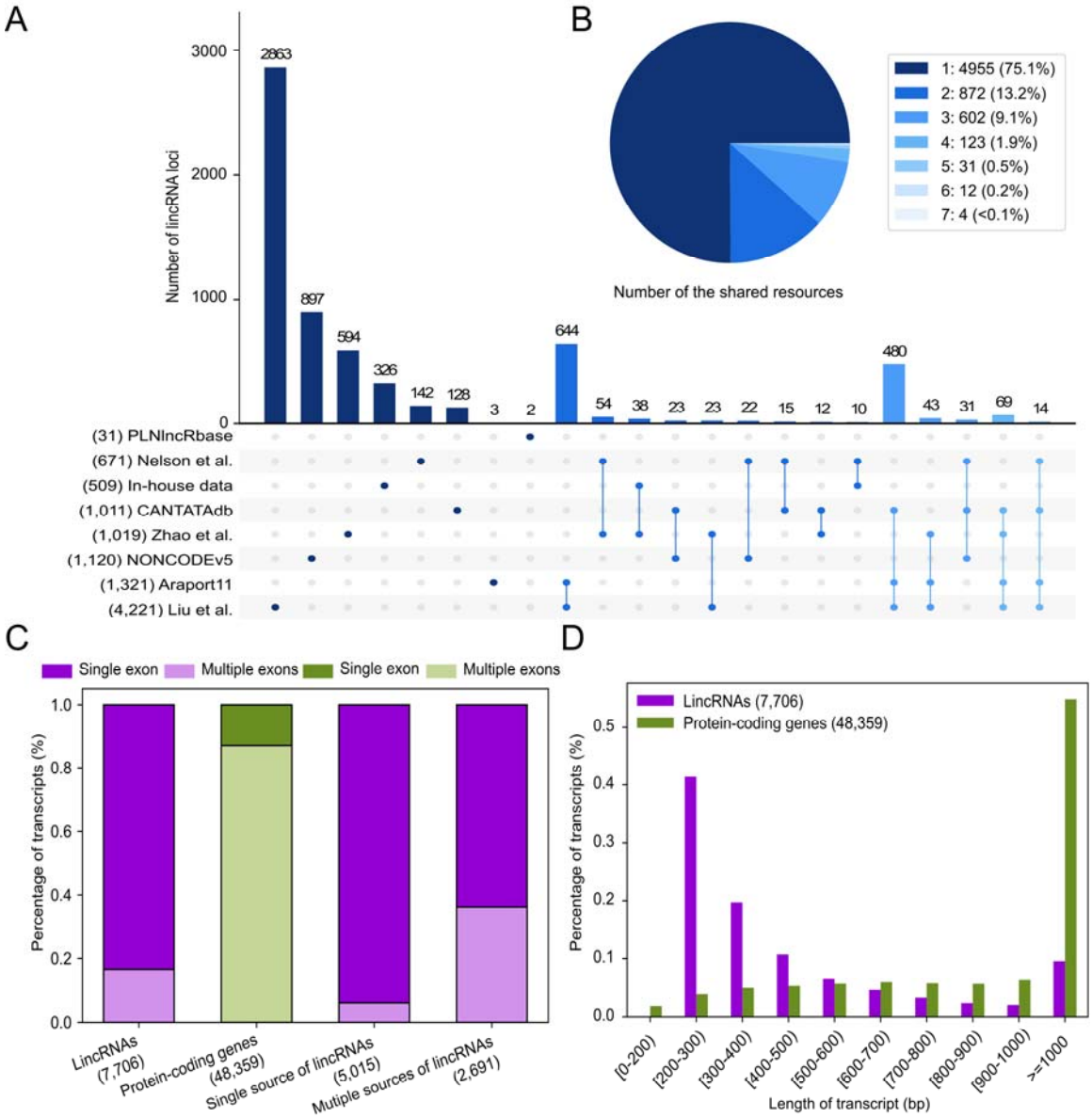

Figure 1. Overlap and gene features of Arabidopsis lincRNA annotations. (A) Upset plot showing the intersection of lincRNA annotation in the eight resources. Each row represents a resource, reporting in parenthesis its total number of lincRNA transcripts before merging. LincRNA annotations unique to a single resource are represented as a single circle while circles connected by lines represent the intersection of lincRNA loci shared between various resources. The bar chart indicates the number of unique lincRNA loci and intersectional lincRNA loci, displaying only intersections that contain at least ten lincRNA loci. More complex overlapping patterns are not shown. **(B)** The pie chart shows the

proportion of lincRNA loci supported by one or more resources. **(C)** The distribution of exon number for all lincRNA transcripts (purple), protein-coding transcripts (green), transcripts of lincRNAs supported by single resource (purple) and multiple resources (purple). Single exon and multiple exons are shown in dark and light colors, respectively. **(D)** The distribution of transcript length for lincRNAs (purple) and protein-coding genes (green).

Contrasting patterns of sequence conservation for lincRNAs belonging to different evolutionary age categories

To assess the evolutionary conservation of Arabidopsis lincRNAs within flowering plants, DNA sequence similarity searches were performed by comparing our set of lincRNAs with the genomes of 40 plant species (see Materials and Methods, Supplemental Table S1). Among the 6,599 lincRNAs, 2,480 lincRNAs were restricted to Arabidopsis and named Arabidopsis-specific lincRNAs. The other lincRNAs were classified into four evolutionary age categories according to the presence of homologs in closely and more distantly related species (Figure 2A). We found 81 lincRNAs with at least one homolog in eudicots and in monocots and therefore conserved during 180 million years (MY) of evolution (Beilstein et al., 2010; Zhang et al., 2020), defined as angiosperm-conserved lincRNAs. Forty-two lincRNAs were conserved in eudicots with at least one homolog in rosids and asterids, but without homologs in monocots. Similarly, 44 lincRNAs were identified only having homologs in rosids species, outside the Brassicaceae family. As the lincRNAs conserved in eudicots and rosids showed highly similar conservation patterns for gene body and promoter ($P<0.376$, Mann–Whitney *U* test), we combined these genes in one category, called Eudicot/rosid-conserved lincRNAs (86 genes). 1,671 Brassicaceae_I_II-conserved lincRNAs were present in the common ancestor of Brassicaceae lineages I and II, without homologs outside the Brassicaceae. Lastly, 2,281 lincRNAs were identified only having homologs in the Brassicaceae I lineage (Brassicaceae_I-conserved lincRNAs).

The monotonous decrease in the number of lincRNAs for the older age categories suggests there is a continuous birth of lincRNA loci at the species level, with only a small fraction showing deep conservation in other plant families or orders.

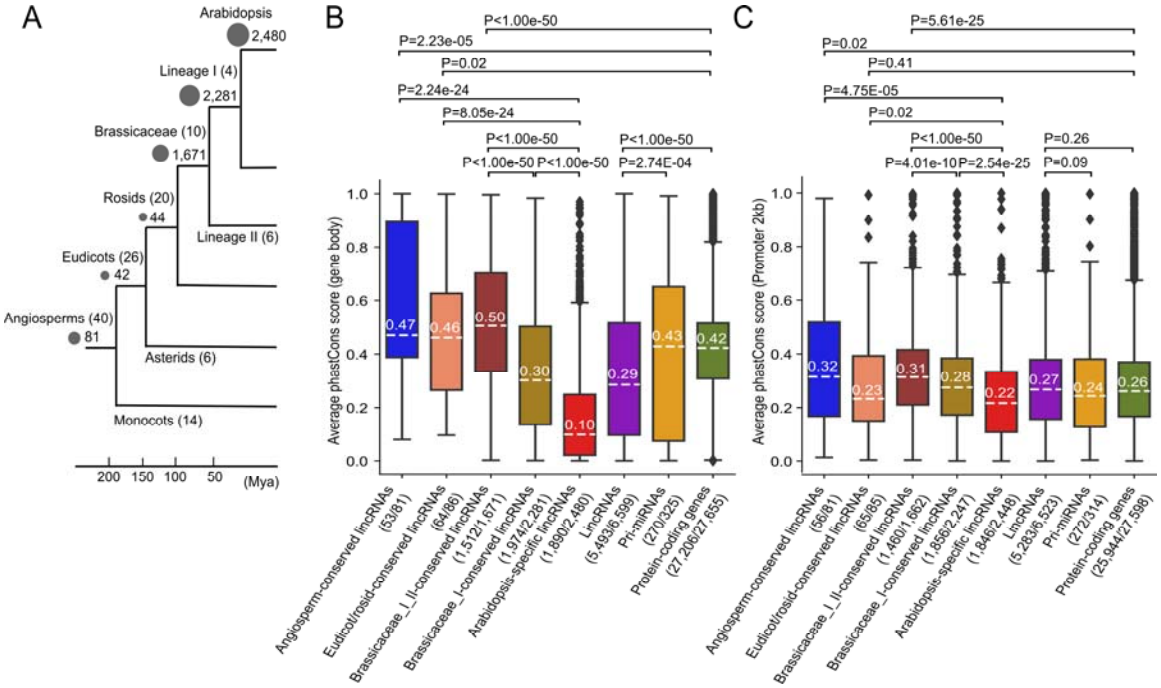

Figure 2. Sequence conservation analysis for lincRNAs of different evolutionary age categories. (A) Simplified species tree reporting the number of lincRNAs found for the different age categories, which are also indicated by the grey circle sizes. Numbers in parenthesis report the number of genomes included per clade to assess sequence similarity and define a lincRNA's age category. Boxplot showing the average phastCons score for **(B)** gene bodies and **(C)** promoter regions (2kb upstream of transcription start site) of different lincRNA age categories and gene types (lincRNAs, pri-miRNAs, protein-coding genes). The numbers in parentheses report the number of gene bodies and promoters with at least 50% of informative nucleotides over the total number of gene bodies and promoters in that category, respectively. PhastCons score ranges from 0 (not under selection) to 1 (strong negative selection). P-values for pairwise Mann–Whitney *U* test are shown using the horizontal lines connecting the series.

Apart from detecting lincRNA homologs, we also evaluated evolutionary selection of lincRNA loci, pri-miRNAs, where globally the 21-24 miRNA sequence is the only conserved region at the nucleotide level, and protein-coding genes using phastCons scores (Siepel et al., 2005). This score reports the probability for each nucleotide to evolve neutrally or under negative, or purifying, selection (low or high phastCons score, respectively). In contrast to the age categories, which are based on finding similar homologous sequences in other plant genomes and do not give information about the mode of selection, phastCons works by fitting a two-state hidden Markov model to a genome-wide multiple sequence alignment and predicting, based on the pattern of nucleotide substitutions, conserved elements representing sites under purifying selection (Siepel et al., 2005). For lincRNA, pri-miRNA, and protein-coding gene loci, we compared phastCons scores for the gene body and the promoter region (2kb upstream of transcription start site) (Supplemental Data Set S2). In general, for gene bodies, lincRNAs are significantly less conserved than pri-miRNAs and protein-coding genes (Figure 2B, purple, yellow, and green series). However, we found that Angiosperm- and Eudicot/rosid-conserved lincRNAs were as conserved as protein-coding genes. Brassicaceae_I_II-conserved lincRNAs showed even significantly higher conservation levels than protein-coding genes, indicating high levels of purifying selection. Arabidopsis-specific lincRNAs show the lowest phastCons scores.

In contrast to the differences in gene body, the promoter scores are comparable for lincRNA, pri-miRNA, and protein-coding genes (Figure 2C, purple, yellow and green series). The level of purifying selection on promoter regions is similar in Angiosperm- and Eudicot/rosid-conserved lincRNAs as well as in protein-coding genes. We observed that conservation levels were again significantly higher in the Brassicaceae_I_II-conserved lincRNA promoters than protein-coding gene promoters. Taken together, these results indicate that Brassicaceae_I_II-conserved lincRNAs stand out in the level of purifying selection acting on these loci, both in the gene body and in their promoter, suggesting that the primary

sequence of the RNA and its promoter regulation is a critical element for lincRNA function.

Large-scale transcriptome analysis reveals highly-specific lincRNA expression in roots

Increasing evidence supports the tissue or cell-type specific role of lincRNA gene functions (Liu et al., 2012; Li et al., 2016; Cheng et al., 2017; Rai et al., 2019). To characterize spatiotemporal Arabidopsis gene expression patterns also considering lincRNAs, we curated, processed, and integrated 791 RNA-seq samples to construct a genome-wide gene expression atlas covering a wide range of tissues, developmental stages, and stress conditions (Supplemental Table S2 and Supplemental Data Set S3). To find a threshold of detectable expression above background, we normalized all data to establish detectable expression levels for protein-coding genes and lincRNAs (Ramskold et al., 2009; Li et al., 2016) (see Materials and Methods, Supplemental Figure S2A). A normalized transcripts per million (TPM) value >= 0.2 was considered to define 5,586 (84.6%) expressed lincRNAs and 284 (87.4%) pri-miRNAs, whereas a TPM >= 2 was used to define 26,254 (94.9%) expressed protein-coding genes. Globally, the expression breadth, defined as the number of samples in which a gene is expressed, was lower for lincRNAs (median: 12/791 samples) compared to pri-miRNAs (median: 35.5/791 samples) and protein-coding genes (median: 648/791 samples) (Figure 3A). Although the identification of lincRNAs is known to be impacted by the sequencing depth (Cabili et al., 2011; Liu et al., 2012; Sun et al., 2017), we did not observe a clear correlation between the expression breadth and sequencing depth (Pearson correlation coefficient, r=0.04, p=0.86) (Supplemental Figure S2B). This result indicates the reported expression values are not strongly biased by study-specific technical parameters. As previously reported, the expression level of lincRNAs is generally lower than that of pri-miRNAs and protein-coding genes (Figure 3B). The lincRNA loci part of the two “older” categories (Angiosperm and Eudicot/rosid-conserved lincRNAs) showed

significantly higher expression levels compared with the three “younger” categories of lincRNAs (Brassicaceae_I_II-conserved, Brassicaceae_I-conserved and Arabidopsis-specific lincRNA). We also observed a six-fold difference in the fraction of expressed genes depending on their annotation source: lincRNAs annotated in only one resource where more frequently found as undetectable compared to lincRNAs annotated in two or more resources (20% versus 3%, respectively).

To identify groups of samples with similar expression patterns, 769 RNA-Seq samples were split into 22 expression clusters using meta-data curation (22 samples were discarded, see Materials and Methods). Globally, these clusters group Arabidopsis (Col-0) samples according to the organ or tissue considered and the stress, either biotic or abiotic, applied (Supplemental Data Set S3). As lincRNAs are proposed to exert their role in a tissue-specific manner (Liu et al., 2015), the tissue-specificity index (tau) (Yanai et al., 2005) for each lincRNA, pri-miRNA and protein-coding gene was calculated to estimate expression specificity between clusters (Figure 3C). LincRNAs were more specifically expressed than pri-miRNAs and protein-coding genes, with the median tau scores of 0.97, 0.95 and 0.74, respectively ($P<0.001$, Mann–Whitney U-test). Forty-six percent of all expressed lincRNAs (2,573/5,563) with a tau score ≥ 0.97 were defined as lincRNAs showing highly-specific expression (Supplemental Data Set S4). Even though we did not observe a uniform distribution of these highly-specific lincRNA over the 22 clusters (Figure 3D), especially cluster 19 (root), 18 (cell line light induction) and 3 (seedling heat) have the largest number of highly-specific lincRNAs. Between 28 and 42% of the highly-specific expressed lincRNAs were present in the different evolutionary categories, with the highest fraction in the Arabidopsis-specific lincRNAs (Figure 3E). Noteworthy, expression cluster 19, containing 66 samples covering both whole root and specific root cell types (Li et al., 2016), contains the highest fraction (14-17%) of highly-specific lincRNAs in the younger age categories (Figure 3E), hinting towards a role of these lincRNAs in root tissues.

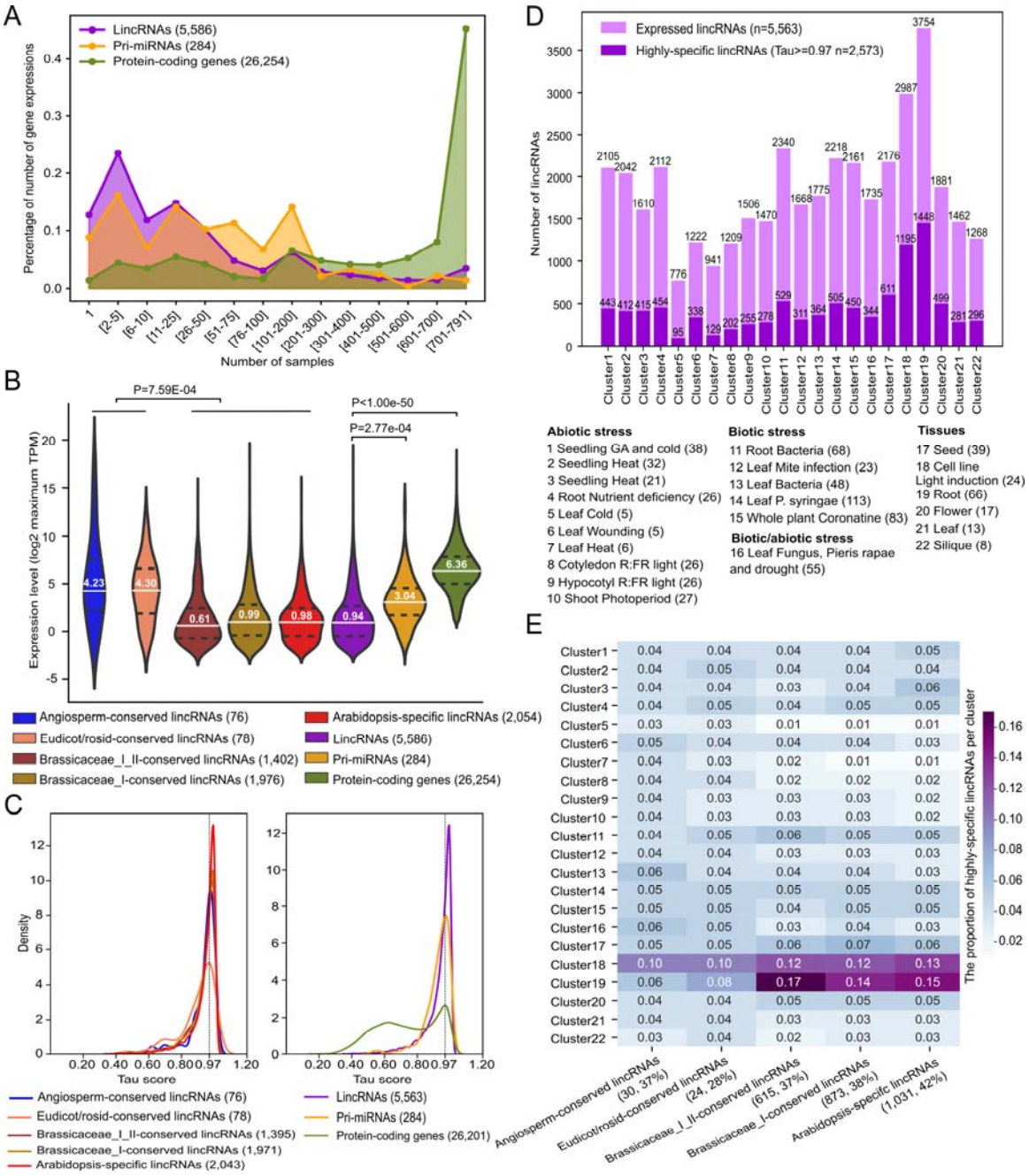

Figure 3. Expression analysis of lincRNAs. (A) Line chart showing the distribution of expression breadth for lincRNAs, pri-miRNAs and protein-coding genes across all samples **(B)** Distribution of the maximum TPM expression levels for different lincRNA age categories and gene types (lincRNAs, pri-miRNAs and protein-coding genes). **(C)** Distribution of tissue specificity tau scores for different lincRNA age categories and gene types. Tau scores range from zero to one, where zero means widely expressed, and one means very specifically expressed

(detectable in only one cluster). The black dotted line represents a tau score of 0.97. **(D)** Bar chart reporting the number of expressed and highly-specific expressed lincRNAs for the different expression clusters. Cluster numbers and descriptions are shown below the chart, with numbers in parenthesis indicating the number of samples present per cluster. **(E)** Heatmap showing the proportion of highly-specific expressed lincRNAs in each cluster for each lincRNA age category. Cluster descriptions are the same as in panel D. Numbers in parenthesis report the number of genes per age category together with the fraction of lincRNAs showing highly-specific expression.

To validate these tissue-specific expression patterns, we verified the expression of previously characterized Arabidopsis lincRNAs (Supplemental Table S3). The lincRNA *SVALKA*, which was identified in a cold-sensitive region of the Arabidopsis genome, was maximally expressed in the cold stress-related cluster 5 in our study (Kindgren et al., 2018). In addition, the lincRNA *MARS*, which is involved in the response to abscisic acid, seed germination and root growth under osmotic stress (Roule et al., 2022b), was found to be expressed at high levels in root cluster 19 as well as cluster 11 (root bacteria) and cluster 17 (seed). The lincRNA *ASCO*, reported to be involved in lateral root formation and response to pathogens (Bardou et al., 2014; Rigo et al., 2020), was found to be widely expressed, including cluster 11, which contained samples reporting bacterial flagellin stress responses. The Arabidopsis lncCOBRA1, also conserved in *Brassica rapa* (Kramer et al., 2022) and playing a role in seed germination, was found in our set of Brassicaceae_I_II-conserved lincRNA and showed the highest expression in cluster 17 (seed) and cluster 3 (seedling heat).

Among the highly-specific lincRNAs, the vast majority (66%) were derived from one resource, with lincRNAs identified in Liu et al. 2012 being most abundant (30%, Supplemental Figure S3A-B). Taken together, these findings indicate that the expression clusters offer a good starting point for the context-specific characterization of known and novel lincRNAs. Furthermore, young age

categories, including the Brassicaceae-conserved and Arabidopsis-specific lincRNAs, show a bias for expression specificity in the root samples, with more than half (1,448/2,573) of the highly-specific lincRNAs active in root, suggesting a potentially diverse role of lincRNA networks in root growth and development.

Experimental TF-lincRNA regulatory networks reveal active and complex gene regulation of Arabidopsis lincRNA genes

To integrate lincRNAs into epigenetic and transcriptional networks, we compared the chromatin states inferred by (Liu et al., 2018; Hazarika et al., 2022) for the different lincRNA gene sets delineated in our study. These chromatin states report a combination of multiple epigenetic marks along the genome and offer detailed insights in the locations and functions of regulatory regions and genes. Globally, lincRNAs show enrichment for vastly different chromatin states compared to protein-coding genes (Figure 4A). Chromatin states 33-34, typically associated with DNA methylation, repressive histone modifications and transposable elements, were strongly overrepresented in angiosperm-conserved lincRNAs, and to a lesser extent in lincRNAs showing highly-specific expression patterns. Chromatin states 31-32, also associated with DNA methylation and repressive histone modifications were strongly overrepresented in Arabidopsis-specific lincRNAs. Interestingly, these repressive chromatin marks were less enriched in the Brassicaceae-I_II-conserved lincRNAs, where chromatin states 13, 19 and 20 were most strongly overrepresented. State 13 refers to Polycomb group mediated deposition of trimethylation of the lysine 27 of histone H3 (H3K27me3) while states 19-20 denote accessible chromatin and TF ChIP binding. Chromatin states 19-20, but also states 31-32 and 34, were enriched in highly-specific lincRNAs. Although we found no evidence of a correlation between lincRNA age categories and chromatin states, our results revealed that a significant number of Brassicaceae_I_II-conserved lincRNAs have epigenetic signatures associated with Polycomb regulation and TF binding in accessible chromatin.

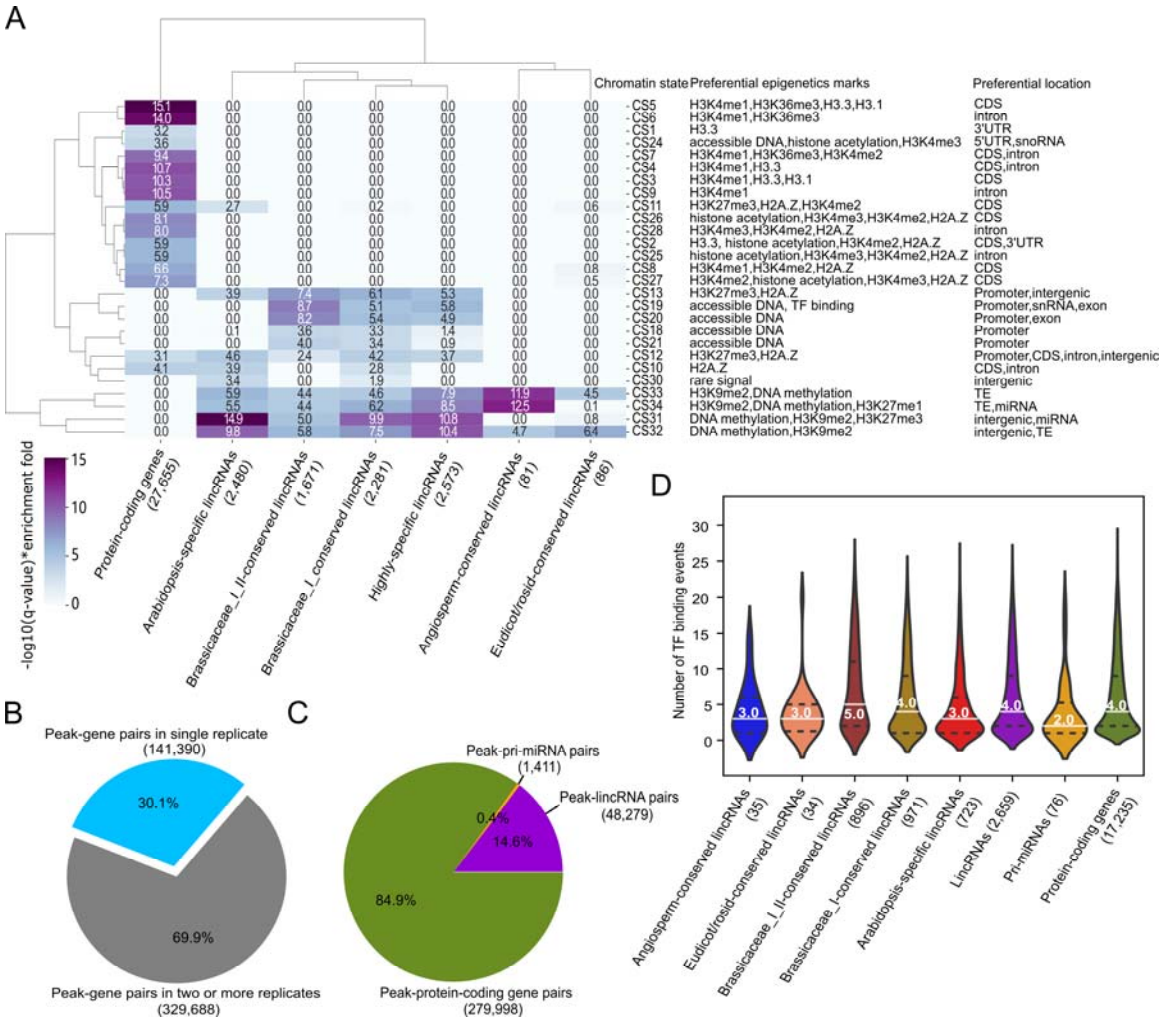

Figure 4. Chromatin state and TF ChIP-Seq peak annotation for different gene types (A) Dendrogram showing the enrichment for different lincRNA gene sets (x-axis) towards different chromatin states (CS) (y-axis). The values report the product of -log10(q-value) and the enrichment fold. Only significant enrichment values are reported (q-value < 0.05). **(B)** The proportion of peak-gene pairs present in single replicate (blue) and two or more ChIP-Seq replicates (grey). **(C)** The percentage of three gene types assigned to peaks in two or more ChIP-Seq replicates. **(D)** Distributions of the number of TF binding events for lincRNA evolutionary age categories and gene types (lincRNAs, pri-miRNAs and protein-coding genes).

Based on the specific expression profiles for different lincRNA genes, the chromatin state information, as well as the high levels of promoter conservation, we next integrated TF chromatin immunoprecipitation (ChIP) data to further identify the regulators controlling lincRNAs. Before we used TF ChIP data to characterize the organization, complexity, and evolution of TF binding for protein-coding genes (Heyndrickx et al., 2014), hence we here re-processed publicly available ChIP-Seq to identify TF binding events potentially controlling lincRNA gene expression. A total of 114 TF ChIP-Seq datasets, covering 45 TFs with at least two biological replicates, were reprocessed (Supplemental Table S4). Starting from our genome annotation containing protein-coding genes, pri-miRNAs and the high-confidence lincRNAs, ChIP-Seq peaks were assigned to the closest genes. A gene was defined as a potential target gene for a profiled TF if it was the closest to at least one ChIP-Seq peak of that TF (see Material and Methods). Based on all 471,078 TF peak-target gene pairs identified from the 114 ChIP-Seq datasets, 329,688 (70%) of the peak-gene pairs were confirmed by two or more replicates and were kept to construct a robust TF- target gene regulatory network (Figure 4B). While most TF peaks were associated with protein-coding genes, we identified 48,279 (14.6%) peaks that were associated with lincRNAs (Figure 4C). Given the strong localization bias for TF binding in Arabidopsis (Heyndrickx et al., 2014; Yu et al., 2016), we only considered expressed target genes localized relative to the peak midpoint within a 2kb window (Supplemental Figure S4A), retaining 93.0% of the binding events (15,686 TF-lincRNA interactions, see Supplemental Data Set S5). Globally, 2,659 expressed lincRNAs had one or more TF binding events. Twenty-two out of the 45 TFs, belonging to the bHLH, HD-Zip, bZIP, C2H2, GRAS, MYB, NAC, NF-YB and NF-YC TF families, were each associated with at least 200 lincRNAs (Supplemental Table S5).

(Haudry et al., 2013) identified over 90,000 conserved non-coding sequences (CNS) in Arabidopsis that show evidence of transcriptional and post-transcriptional regulation. Comparing these CNS with the ChIP-Seq peaks of target genes revealed that most of the TF binding events close to lincRNAs

(78.1%) and protein-coding genes (73.7%) overlapped with a CNS (Supplemental Figure S4B). This fraction was much lower for pri-miRNAs (57.0%). The highest fraction of ChIP-Seq peaks containing a CNS was detected for Brassicaceae_I_II-conserved lincRNAs loci (91.8%), indicating that these TF binding events are evolutionary constrained and potentially functional. Considering the different gene types, the median number of TF binding events per locus was higher for protein-coding genes and lincRNAs (median of 4 TFs) compared to pri-miRNAs, suggesting these genes are differently controlled (Figure 4D). Brassicaceae-conserved lincRNAs have more TF binding events than lincRNAs in any other age categories (Figure 4D). Furthermore, we observed a positive correlation between TF binding frequency and the conservation of lincRNA gene body (r=0.27, P=8.11e-92) (Supplemental Figure S4C). More precisely, a large faction of lincRNAs without any TF binding event also has very low phastCons scores, close to zero, indicating that these genes are not under purifying selection. We observed a neglectable correlation between the number of TF binding events and tissue specificity (r=0.02, p=1.69e-01). Altogether, lincRNAs experiencing stronger levels of purifying selection are bound, and potentially regulated, by more TFs, independent of their tissue specificity (Supplemental Figure S4D). Taken together, we generated a multi-level genome-wide characterization covering chromatin state information, promoter conservation, ChIP-based TF binding and CNSs, for all detectable lincRNA across >700 expression samples, permitting to rapidly define the biological context and relevance of lincRNAs in Arabidopsis regulatory networks.

MYB44, PIF4 and KAN1 regulate Arabidopsis lincRNAs in different root cell types

To identify TF-lincRNA regulatory interactions active in specific cellular contexts, we used the previously defined expression clusters to combine TFs regulation and lincRNA expression in specific organs, tissues, or stress conditions. Considering all expressed lincRNAs and all TF peak-based regulatory

interactions described above, we tested if specific expression clusters are overrepresented for lincRNAs controlled by specific TFs (Y-axis and X-axis in Figure 5A-B, respectively). We observed a significant overrepresentation for KANADI1 (KAN1) binding to lincRNA loci in 11 clusters, of which clusters 19 (root), 18 (cell line light induction), 14 (leaf *P. syringae*) and 6 (leaf wounding) showed the most significant overlaps (Figure 5A). Comparing the different clusters revealed that root cluster 19 contained numerous enriched TFs, including Arabidopsis Zinc-Finger protein 1 (AZF1), JAGGED (JAG), Phytochrome-Interacting Factor 5 (PIF5), Repressor of GA (RGA), PIF1, MYB3, KAN1 and HAT22. The observed patterns of TF-binding in different expression clusters were unique for lincRNAs and highly dissimilar compared to TF binding enrichment for protein-coding genes (Supplemental Figure S5A).

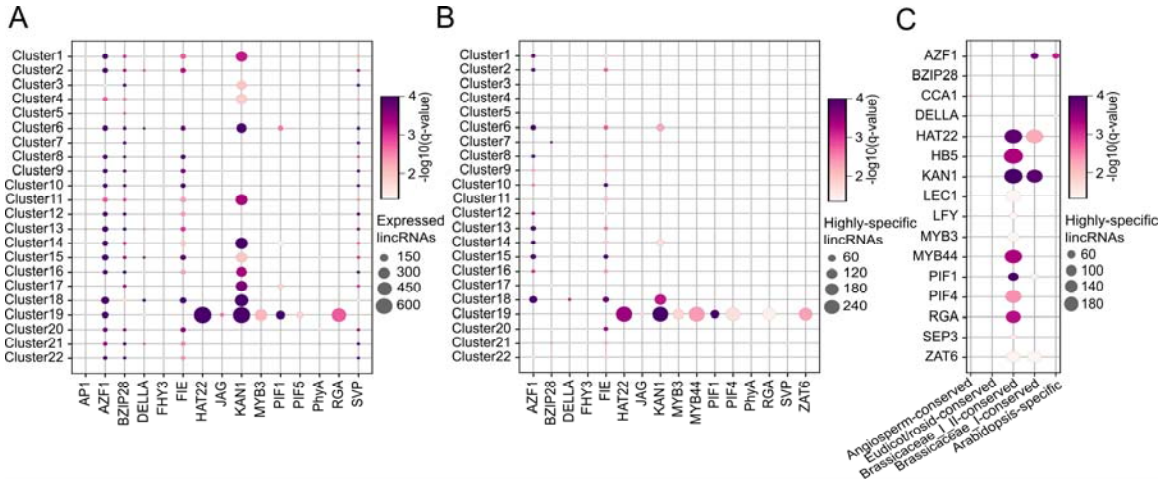

Figure 5. Overview of TF-lincRNA regulatory interactions in different expression clusters and age categories. The dot sizes represent the number of the lincRNAs while the color represents the statistical significance. Cluster descriptions are the same as in Figure 3D. **(A)** Bubble chart showing the enrichment of TF binding for expressed lincRNAs in different expression clusters. TFs lacking significant enrichment in any of the 22 clusters are not shown. **(B)** Bubble chart showing the enrichment of TF binding for highly-expressed lincRNAs in different expression clusters. TFs lacking significant enrichment in

any of the 22 clusters are not shown. **(C)** Bubble chart showing the enrichment of TF binding for expressed lincRNAs in different age categories.

Focusing on the set of 2,573 lincRNAs showing highly-specific expression, most of the TF enrichments for root cluster 19 remained (Figure 5B). PIF1, PIF4, KAN1, HAT22, MYB3, MYB44, ZAT6 and RGA showed the largest overlap and all these TFs, apart from MYB3 and PIF1, contained >160 lincRNA target genes (Figure 5B and Supplemental Table S6). We confirmed that these eight TFs were all expressed in one or more samples of the root expression cluster 19 (Supplemental Figure S5B) and earlier studies have reported that these TFs, apart from HAT22, are involved in root development in Arabidopsis (Hawker and Bowman, 2004; Devaiah et al., 2007; Moubayidin et al., 2016; Tominaga-Wada and Wada, 2016; Zhao et al., 2016; Li et al., 2022). Comparing TF binding for the different age categories revealed that the Brassicaceae_I_II-conserved lincRNAs were most strongly enriched for these TFs (Figure 5C).

To validate the ChIP-based regulatory interactions for the identified TFs, we used reverse transcription quantitative PCR (RT-qPCR) analysis in roots of lines affected in TF expression such as overexpression or T-DNA inactivation (called TF perturbation lines). We used an inducible line for KAN1 (*KAN1-GR*), overexpression lines for MYB4 and PIF4 (*35S::MYB4*, 35S*::PIF4*) or quadruple mutant of the *pif1*, *pif2*, *pif3*, and *pif4* (*pifq*), and a knockout mutant for RGA (*rga28* T-DNA line). We then selected 27 potentially regulated lincRNAs showing high expression levels in the root cluster 19 and that were targeted by several of the selected TFs. For example, *LincRNA5331* and *LincRNA1119* were predicted to be regulated by the four TFs. We could detect deregulation for *LincRNA5331* and *LincRNA1119* in *35S::MYB44* and *pifq* roots, whereas the former was also deregulated in *KAN1-GR* and the latter in *35S::PIF4* (Figure 6A, Supplemental Figure S6). Overall, out of the 74 inferred regulatory interactions investigated, 36 were confirmed, meaning that for the tested TFs, a significant deregulation of the lincRNA was found in comparison to the control (“Has a peak and is DE”, in

Table 1, Supplemental Table S7). For 23 out of the 27 tested lincRNAs, we confirmed one or more regulatory interactions (Figure 6A). The precision (i.e. the positive predictive value), varied between 27-65% while the average accuracy (i.e. the proportion of correct predictions among all genes examined) was 59%. For eight interactions the TF peak annotation to the lincRNA was unclear ("Has a putative peak and is DE" in Table 1), meaning these deregulated lincRNAs might also resemble confirmed interactions. Lastly, while the 10 interactions where we observed deregulation in the absence of a peak could indicate false predictions, they might also represent cases of indirect regulation controlled by the perturbed TF, influencing the expression of the profiled lincRNAs.

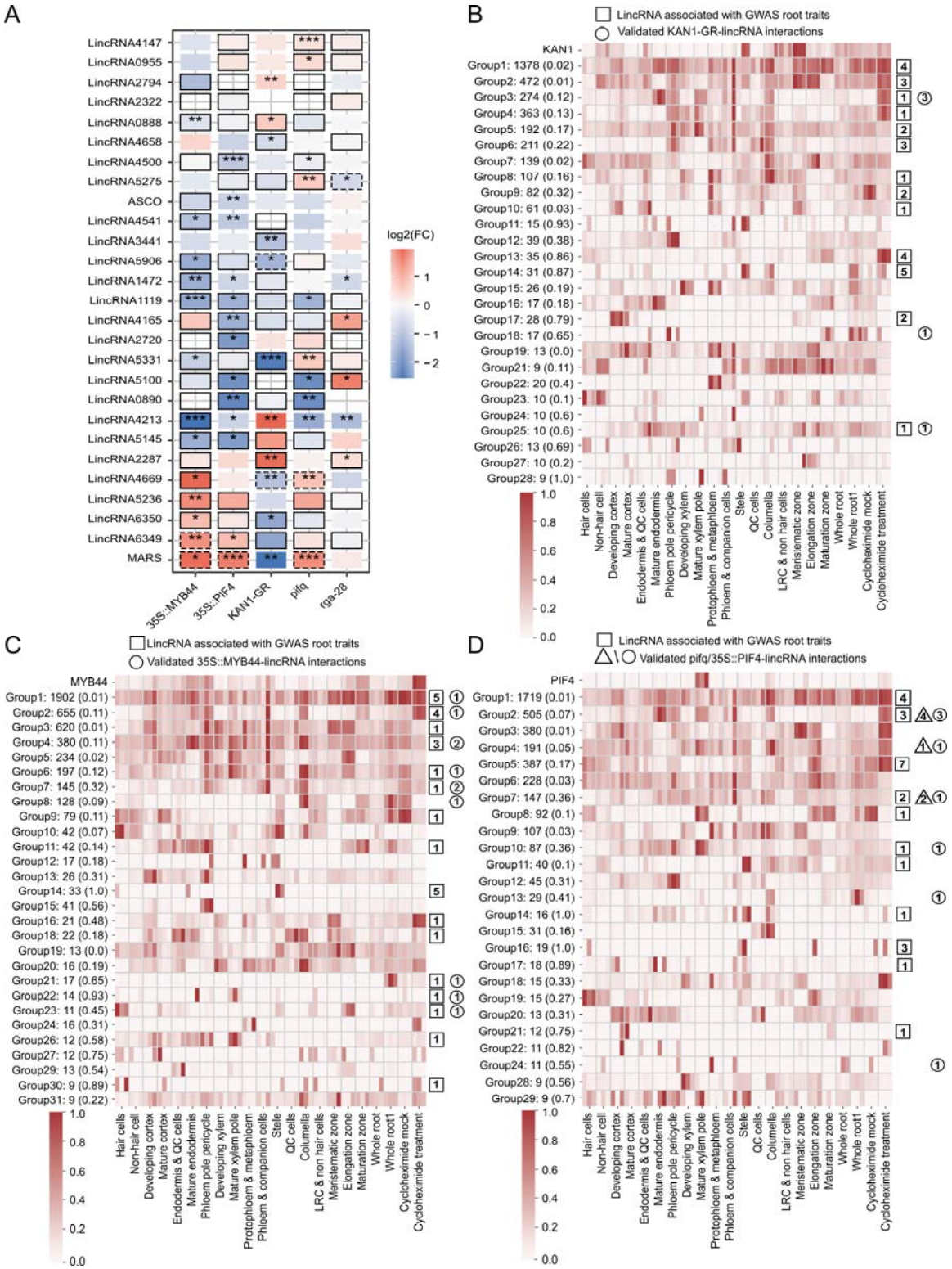

Figure 6. Experimental validation and characterization of TF-lincRNA regulatory interactions. (A) Heatmap showing log2 fold change (FC) of lincRNA relative expression levels in transcription factor (TF) overexpressing lines

(35S::*MYB44*, 35S::*PIF4* and *KAN1-GR*) or TF knockout lines (*rga28* and *pifq* (*pif1/pif2/pif3/pif4* quadruple mutants)) vs. control wild-type lines in 14-day old roots. Expression values were determined by RT-qPCR. Asterisks indicate statistically significant differences (*p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001) in an unpaired two-tailed Student's t-Test (n =3). Solid boxes indicate TF ChIP peaks <2kb from the lincRNA gene. Dashed boxes indicate TF ChIP peaks were identified only in one ChIP-Seq replicate or the lincRNA is not the closest gene to the TF peak. **(B)** Heatmap showing KAN1 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. The first line shows the expression profile of the TF under investigation while the other lines show groups of target genes with similar expression. Next to the group number, the number of target genes is reported and in parenthesis the fraction of lincRNAs per group is indicated. The numbers in squares and circles report the number of GWAS root-related and RT-qPCR experimentally validated lincRNA genes, respectively. **(C)** Heatmap showing MYB44 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. Group information is the same as in panel B. **(D)** Heatmap showing PIF4 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. Group information is the same as in panel B, but here triangles and circles report the RT-qPCR experimentally validated lincRNA genes for the *pifq* and *35S::PIF4* lines, respectively.

Based on the experimental validation results confirming that several of the root-specific lincRNAs are controlled by the TFs inferred using the ChIP-based peaks, we next integrated genotype-phenotype relationships from the AraGWAS catalog. The aim of this analysis was to verify if root-related phenotypes have been reported for regions containing lincRNAs in genome-wide association studies (GWAS). After processing the significant associations from all GWAS studies present in the catalog and only retaining associations overlapping with lincRNA gene bodies or 2kb promoter regions (see Material and Methods), we identified 2,865 single nucleotide polymorphisms (SNPs) overlapping with 1,124 lincRNAs,

covering 147 different studies. While 61.6% of the lincRNAs had only significant associations via SNPs in the promoter region, 38.4% had associations via SNPs in the gene body (194 and 238 lincRNAs with significant associations only in the gene body and in both gene body and promoter, respectively). After parsing and summarizing phenotype information from the available trait ontology annotations, we could link 23 lincRNAs to abiotic stress-related traits, 133 lincRNAs to flower-related traits, 33 lincRNAs to leaf-related traits, 376 lincRNAs to root-related traits, and 27 lincRNAs to seed-related traits (see Supplemental Data Set S6). The number of root-related GWAS annotations was much larger for lincRNAs targeted by TFs showing significant enrichment for binding in root cluster 19, compared to TFs lacking this enrichment (Supplemental Figure S7), confirming the functional relevance of these lincRNAs in roots. Overall, the reprocessed GWAS data indicate that hundreds of regions overlapping with lincRNAs loci are significantly associated with different plant traits and can be used to prioritize lincRNAs likely involved in specific biological processes.

To further investigate the relevance of the identified regulatory network, we focused on KAN1, MYB44, and PIF4, as these regulators had several root-expressed lincRNAs target genes that showed differential expression in the TF perturbation lines (Table 1). Starting from the protein-coding genes and lincRNAs expressed in root cluster 19 and annotated with a TF ChIP-seq peak, we first grouped target genes with similar expression in the root expression atlas. Secondly, for each group we determined the fraction of lincRNAs and verified if any of these were overlapping with our RT-qPCR experimentally validated lincRNAs or had root-related traits in the GWAS catalog. Finally, we verified if the tissue- or cell-type specific expression of the different groups was in agreement with the expression of the regulatory TF. For KAN1, we identified 14 groups that fulfilled these criteria (Figure 6B), of which eight contained two or more lincRNAs with root-related traits and several showed strong cell-type specific expression. Group 14, showing high expression in stele contained 27 lincRNAs (87%) of which five were associated with GWAS root traits. The high expression of KAN1 in the meristematic zone agreed with groups 1-2, 8-10, and 25, which together

contained 12 confirmed lincRNAs, either through root-related traits or our experimental validations. For MYB44, we identified multiple groups containing GWAS root-related lincRNAs showing specific expression patterns in different cell types: group 4 contained multiple confirmed target genes expressed in mature endodermis while group 7 contained 3 confirmed lincRNAs (one GWAS root-related and two experimentally validated genes) showing high expression in phloem and companion cells (Figure 6C). For PIF4, which showed highest expression in mature xylem pole, groups 8 and 10 contained confirmed lincRNA target genes showing similar cell-type specific expression (Figure 6D). While other groups contained GWAS root-related lincRNAs showing strong expression in stele (groups 11, 14 and 16), our expression data does not confirm that PIF4 is actively expressed in the root stele. All GWAS root-related (58) or experimentally validated (21) lincRNAs regulated by KAN1, MYB44 or PIF4 can be found in Supplemental Table S8.

To confirm the root and cell-type specificity of the TF-lincRNA regulatory interactions, we integrated publicly available chromatin accessibility datasets of Arabidopsis roots based on Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq): three bulk datasets (Maher et al., 2018; Potter et al., 2018; Tannenbaum et al., 2018) and one single-cell dataset (Dorrity et al., 2021). Assessing the tissue specificity of our TF-lincRNA regulatory network revealed that 93% of the TF peaks associated with lincRNAs were detected in root accessible chromatin regions (ACRs), of which 21% were specifically detected in cell-type specific root ACRs. Randomly shuffling the TF peaks associated with lincRNAs showed that the observed overlap of TF-lincRNA regulatory interactions with bulk and single-cell root ACRs was 10-fold and 36-fold higher than expected by chance, respectively, confirming the high specificity of the inferred regulatory interactions. Of the 105 regulatory interactions covering GWAS root-associated or experimentally validated lincRNAs regulated by KAN1, MYB44, or PIF4, 97 were detected in the bulk root ACRs, of which 22 were detected in root cell-type specific ACRs. For several lincRNAs the cell-type specificity identified using the cell-type gene expression profiles were confirmed

by the single-cell ATAC data (Supplemental Figure S8, Supplemental Table S8). For example, *LincRNA1736*, regulated by KAN1, and *LincRNA1296* and *LincRNA6434*, regulated by MYB44, were expressed in stele (xylem/phloem) and confirmed by xylem-specific ACRs. *LincRNA4180*, regulated by PIF4, was expressed in endodermis and cortex, which was confirmed by endodermis- and cortex-specific ACRs.

This analysis further supports that KAN1, MYB44 and PIF4 are controlling lincRNA genes showing highly specific expression in different root tissues and cell types. Additionally, the overlap with genetic associations hints to a role for many of these TF-controlled lincRNA loci in root growth and development.

DISCUSSION

Comprehensive annotation of Arabidopsis lincRNAs using transcriptomics and evolutionary genomics

In contrast to protein-coding genes, the characterization of lincRNAs is more challenging as we lack highly curated annotations and extensive experimental observations. Furthermore, the low levels of sequence conservation for the majority of lincRNA loci makes it difficult to translate biological knowledge learned in one species to another. Apart from co-expression network analysis, reporting putative (in) direct associations between lincRNAs and other genes, information about TF regulation of lincRNAs is scarce. Embedding lincRNAs in biological networks has great potential to define lincRNAs linked to specific cellular or morphological phenotypes.

Through the integration of eight lincRNA annotation resources, as well as mapping various conservation, chromatin, and expression features, we presented a global view on gene regulation for 5,586 expressed Arabidopsis lincRNAs. We strongly focused on using replicated samples when processing high-throughput datasets to obtain high-confidence gene information (Ponting and Haerty, 2022). Combined with comparative genome analysis yielding information about age

categories and selection acting on gene bodies and promoter regions, we found that different subsets of lincRNAs have distinct molecular properties. While the high tissue-specificity and low levels of primary sequence conservation corroborate previous findings about lincRNAs (Necsulea et al., 2014; Ma et al., 2019; Palos et al., 2022), the analysis of lincRNA expression using a genome-wide gene expression atlas covering 769 samples revealed that lincRNA expression is widespread in different organs and not restricted to stress conditions, which is in agreement with previous studies (Jha et al., 2020; Corona-Gomez et al., 2022). In a recent study, Corona-Gomez and colleagues reconstructed a co-expression network to annotate lncRNAs with associated protein-coding genes. They identified several modules associated with root development or root-related stress functional annotation (Corona-Gomez et al., 2022). These results are consistent with the high representation of lincRNAs exhibiting highly-specific expression in the root expression cluster (unique set of 1448 lincRNAs), which is the highest number among all expression clusters we studied. When intersecting the different age categories with the expression clusters, Brassicaceae_I_II-conserved lincRNAs covered a large fraction of these root-expressed lincRNAs. While recent studies also identified a large number of context-specific lincRNAs expressed in root tip or meristem (Corona-Gomez et al., 2022; Palos et al., 2022), our age category analysis revealed that many lincRNAs that originated in the common ancestor of Brassicaceae lineages I and II, showed specific expression in various root cell types.

We observed a clear trend of increasing levels of lincRNA gene expression when going from young to old age categories. However, quantifying selection levels acting on different loci revealed that the oldest age categories, showing the most wide-spread and highest expression, are not undergoing the highest levels of purifying selection. While older categories like angiosperm-conserved and Eudicot/rosid-conserved lincRNAs had similar median phastCons scores as protein-coding genes, Brassicaceae_I_II-conserved lincRNAs showed the highest levels of purifying selection, both in their gene body and promoter. In animals, a significant increase of dynamically expressed genes and higher levels

of purifying selection has been reported for older age categories (Necsulea et al., 2014; Sarropoulos et al., 2019). Furthermore, massively parallel reporter assays surveying thousands of human promoters revealed that tissue-specific lincRNAs had fewer TF motifs compared to ubiquitously expressed genes (Mattioli et al., 2019). In contrast to Arabidopsis protein-coding genes, this pattern was not found for lincRNAs in our analysis (correlation between tau score and TF binding frequency = 0.02). These results suggest, based on the genome-wide TF binding data available in Arabidopsis, that the complexity in TF control of lincRNAs is different between plants and animals and that the observed pattern of highly-specific expression and purifying selection for Brassicaceae_I_II-conserved lincRNAs deviates from the global trends observed in animals. Therefore, these unique properties suggest that these lincRNAs, which only emerged 42 million years ago, are better integrated in plant networks when compared to young lincRNAs in animals.

Integrative regulatory annotation of Arabidopsis lincRNAs

The regulatory annotation using genome-wide chromatin state information revealed that protein-coding genes and lincRNAs have distinct chromatin signatures. The enriched chromatin states for lincRNAs were largely variable between, and sometimes within, the different age categories. While states associated with DNA methylation and repressive histone modifications were most strongly overrepresented in the youngest (Arabidopsis-specific) and oldest (angiosperm-conserved) lincRNA age categories, states denoting polycomb group mediated deposition of H3K27me3 and accessible chromatin were most enriched for Brassicaceae_I_II-conserved lincRNAs. H3K27me3 is a repressive covalent histone modification resulting from the activity of Polycomb repressive complexes. It was recently shown that a reduction in H3K27me3 levels leads to a decrease in the interactions within Polycomb-associated repressive domains, resulting in a global reconfiguration of chromatin architecture and transcriptional reprogramming during plant development (Huang et al., 2021). Chromatin

accessibility is a hallmark of regulatory DNA as it allows sequence-specific binding of TFs, key components of transcriptional regulatory networks (Schmitz et al., 2022). The association of these chromatin states with specific sets of lincRNAs strongly indicates active transcriptional regulation.

To identify context-specific TF regulation potentially driving the highly-specific expression observed for many lincRNAs, we reprocessed, filtered and annotated 114 ChIP-Seq experiments covering 45 TFs, yielding a TF-lincRNA gene regulatory network containing 2,659 lincRNAs and 15,686 interactions. To assess the potential functionality of these inferred regulatory interactions, we overlapped CNSs identified using nine Brassicaceae genomes, which confirmed that TF peaks close to lincRNAs show similarly high levels of sequence constraint compared to protein-coding genes (74-78%). Furthermore, TF binding events close to Brassicaceae_I_II-conserved lincRNAs showed the highest levels of CNS conservation (92%), which agrees with the very high phastCons promoter scores we observed for this age category. While Palos and co-workers reported that CNSs significantly correlated with gene bodies of Brassicaceae-conserved lincRNAs (Palos et al., 2022), our results revealed that also lincRNA promoters and TF binding sites are strongly conserved for Brassicaceae_I-II-conserved lincRNAs. Ultraconserved CNSs, frequently associating with TF binding sites for key plant regulators controlling essential biological processes, have been identified for thousands of protein-coding genes (Van de Velde et al., 2016). Our results suggest that such deep conservation of cis-regulatory elements is extremely rare for lincRNAs, as only a small number of lincRNA genes show deep evolutionary conservation. Such deeply conserved CNSs for protein-coding genes frequently occur in divergent gene pairs, where they form mini-regulons representing conserved transcriptional units of co-regulated and co-expressed neighboring genes (Van de Velde et al., 2016). It is currently unclear if such conserved transcriptional regulons also exist for lincRNA loci and could explain the observed patterns of positionally-conserved but sequence-diverged lncRNAs (Mohammadin et al., 2015).

The integrated TF binding information showed that promotors of lincRNAs differ strongly from those of protein-coding genes, but also revealed high levels of heterogeneity among the different age categories. In animals, promoters of protein-coding genes contain more TF binding sites than those of lincRNAs, suggesting a stronger and more complex transcriptional regulation of the former (Necsulea et al., 2014; Sarropoulos et al., 2019). When comparing the number of TF binding events for the different gene types in Arabidopsis, we observed no difference in TF binding frequency for protein-coding genes and lincRNAs, indicating that lincRNAs are also regulated in a complex manner in plants. While no correlation between lincRNA expression tissue-specificity and TF binding frequency was found, a positive correlation between TF binding frequency and the level of purifying selection was observed. Again, Brassicaceae_I_II-conserved lincRNAs stood out having the highest number of binding TFs, suggesting that neither the age nor the expression of a lincRNA, but its importance for plant fitness, is a major factor in determining its regulatory complexity. While our results confirm that broadly expressed protein-coding genes, showing high expression breadth, are positively correlated with the number of regulating TFs (Heyndrickx et al., 2014), we did not observe this global trend for lincRNAs, indicating that the regulatory properties of TF control for plant protein-coding genes and lincRNAs are different. Although the number of Arabidopsis TFs profiled using ChIP-Seq may be considered as limited, future research will have to address whether the complexity of TF control varies for lincRNAs, as well as protein-coding genes, active in different organs, tissues, or stress conditions.

A TF-lincRNA gene regulatory network identifies KAN1, MYB44 and PIF4 as regulators controlling root lincRNAs

Through integration of our spatiotemporal expression clusters and the TF-lincRNA gene regulatory network, we identified eight TF regulators showing a significant enrichment for TF binding close to lincRNAs specifically expressed in

roots. While the overlap between TF ChIP-Seq peaks and CNSs gave an indirect indication of the potential functionality of TF binding sites close to hundreds of lincRNA loci, we experimentally validated a set of inferred regulatory interactions, focusing on 27 root-expressed lincRNAs and 5 TF perturbation lines. The number of confirmed regulatory interactions as well as the positive prediction values found for the tested TFs and lincRNAs here (27-65%) are 3 to 4 times higher than the fraction of TF-bound protein-coding genes also showing deregulation reported for a set of TF regulators involved in flowering (7-22%) (O'Maoileidigh et al., 2014). Compared to the discovery rates obtained for large-scale phenotypic screens of insertional lines (1.3%) (Ransbotyn et al., 2015), our discovery rates for deregulation are 20-50 fold higher. Globally, for 85% of the profiled lincRNAs one or more confirmed regulatory interactions were found. These findings indicate that the ChIP peak-based inference of TF regulation is a promising approach to characterize TF regulation of lincRNAs. Additional deregulated lincRNAs lacking a TF peak were also identified, which might be due to an indirect effect caused by crosstalk between different regulators in the Arabidopsis root as well as the type of mutations chosen in the perturbed TF lines (e.g. partners lacking in overexpressing lines, compensatory effects in gene families) (Heyndrickx et al., 2014).

While detailed functional characterizations of lincRNA genes are scarce, the integration of GWAS information allowed us to identify genetic associations for 1,124 lincRNAs covering various traits. While this number, corresponding to a frequency of 9%, is slightly lower compared to that of protein-coding genes associated with a specific trait in the AraGWAS catalog (3,030/27,655 = 11%), it does confirm the great potential of this largely untapped resource to biologically characterize lincRNAs potentially controlling different plant traits. As shown for KAN1, MYB44 and PIF4, multiple groups of co-expressed lincRNAs were identified bound by one of these TFs. Most of these groups showed strong cell-type specific expression and contained lincRNAs that were annotated with root-related traits. While most regulatory interactions were confirmed by TF peaks overlapping with root bulk ACRs, for several lincRNAs cell-type specificity was

confirmed based on single-cell ATAC data, despite the high sparsity associated with this data type.

While co-expression networks and modules containing lincRNA genes cannot differentiate between direct and indirect regulatory interactions and lack functional information about individual lincRNAs (Corona-Gomez et al., 2022; Palos et al., 2022), our complementary approach relying on TF regulation and GWAS information overcomes some of these shortcomings. Taken together, the integration of different gene annotations combined with information about evolutionary conservation, selection, expression, TF regulation and GWAS data yielded new insights on the biological relevance of hundreds of Arabidopsis lincRNAs and offers a promising strategy to identify lincRNAs involved in different aspects of plant biology.

MATERIALS AND METHODS

Prediction of lincRNAs from in-house dataset

Paired-end RNA-seq datasets with high sequencing depth conducted in previous projects in the group at the Institute of Plant Sciences Paris-Saclay were used to predict additional lincRNAs. All data came from experiments carried out in the Col0 ecotypes and involved nsra/b mutant seedlings in response to NPA/NAA treatment GSE65717 and GSE116923 (Tran Vdu et al., 2016; Bazin et al., 2018), seedlings with modified expression of the *ASCO* lincRNA GSE135376 (Rigo et al., 2020), root tip submitted to a short phosphate starvation kinetic GSE128250 (Blein et al., 2020) and a lateral root initiation kinetic from a binding essay of five time points without replicates (6h, 12h, 24h, 36h and 48h after binding, unpublished). All reads were quality trimmed using Trimmomatic. For each library independently, cleaned reads were mapped on TAIR9 genome sequence with STAR (version 2.7.2a) using Araport11 as a guided annotation with the following additional parameters: --alignIntronMin 20 --alignIntronMax 3000. For each alignment file, StringTie (version 2.1.4) was used to predict transcripts using

Araport11 annotations as a guide (additional parameters: -c 2.5 -j 10). GFFcompare (v0.12.6) was then used to isolate the new transcripts in comparison to Araport11 gene annotation (removing transcripts with class code =, c, e or s). The different transcripts prediction were then combined using StringTie in merge mode (additional parameters: -F 0 -T 0 -c 0 -f 0 -g 0 -i). The final set of transcripts was compared against Araport11 annotation with GFFcompare removing all transcripts with a class code of =, c, e, s or m. Transcript were then associated with their already annotated gene or to newly defined genes in case they were predicted in unannotated portion of the genome. Coding potential was then assets using COME (Hu et al., 2017), Coding Potential Calculator CPC (v0.9-r2), CPC2 (Kang et al., 2017) and infernal (v1.1.2) (Nawrocki and Eddy, 2013) against Rfam v14.1 (Kalvari et al., 2021) with default parameters. Non-coding transcripts were the ones predicted by CPC, CPC2 and COME as non-coding and having no hits against tRNA, rRNA, snRNA or snoRNA genes in Rfam.

Integration of lincRNA gene annotations

Arabidopsis lincRNAs were collected from public databases including Araport11 (Cheng et al., 2017), CANTATAdb (Szczesniak et al., 2016), NONCODEv5 (Fang et al., 2018), PLNlncRbase (Xuan et al., 2015), obtained from publications (Liu et al., 2012; Nelson et al., 2017; Zhao et al., 2018) or shared by Andrew D. L. Nelson from the Boyce Thompson Institute at Cornell using their previously published method, and the in-house dataset. These collections contain lincRNA annotations based on transcriptomic information coming from a wide variety of organs including seedling, root, pollen, rosette leaf, endosperm, seed, siliques, inflorescence, flower, floral buds, as well as abiotic stress treatments. The pipeline used for the identification of putative lincRNA is described in Supplemental Figure S1A. (1) LincRNA transcripts with a length of at least 200 bp were retained. (2) Only transcripts that were at least 500 bp away from any protein-coding gene were retained and considered as intergenic (Liu et al., 2012; Yamada, 2017). (3) Transcripts lacking strand information were discarded. (4)

The coding potential of transcripts was assessed using the Coding-Non-Coding Index (CNCI; Version 2) (Sun et al., 2013), CPC2, and Pfam-scan (PFAM) (Finn et al., 2016), and only those transcripts fulfilling the CPC2 (cutoff < 0), CNCI (cutoff < 0) and PFAM (E-value 1e-5) criteria were retained. (6) All candidate intergenic transcripts were assigned to lincRNA loci using the GFFcompare program (Pertea and Pertea, 2020). The number of transcripts retained after each filtering step is reported in Supplemental Figure S1B.

Evolutionary conservation analysis

Our set of Arabidopsis lincRNAs was classified into distinct evolutionary age categories based on sequence similarity. The sequences of 6599 lincRNA loci were extracted using BEDTools getfasta v2.30.0 (Quinlan and Hall, 2010). Genome sequence data for 40 plant species were obtained from PLAZA 5.0 Dicots and PLAZA 5.0 Monocot (Van Bel et al., 2021), representing 26 eudicots (20 rosids and 6 asterids) and 14 monocots. The 20 rosids contain 10 Brassicaceae species. LincRNA homologs were identified using sequence similarity searches against these 40 genomes using BLASTn (Altschul et al., 1990) and applying an E-value cutoff of 1e–10 (Nelson et al., 2017). Classification rules were defined to construct five evolutionary age categories. A lincRNA was deemed conserved in the Angiosperm evolutionary age category when at least one homolog was found in eudicots and one homolog in monocots. LincRNA was defined as a eudicot/rosid-conserved lincRNA with at least one homolog in rosids, one homolog in asterids, and no homolog in monocots, in addition with one homolog only in rosids. The lincRNA was assigned as a Brassicaceae_I_II-conserved lincRNA with at least one homolog in Brassicaceae lineage I, one homolog in Brassicaceae lineage II, and no homologs outside the Brassicaceae species. A lincRNA that had at least one homolog in a Brassicaceae lineage I species, apart from Arabidopsis, was defined as Brassicaceae_I-conserved lincRNA. The last category, defined as Arabidopsis-specific lincRNA, were restricted to Arabidopsis lincRNAs without homologs in any of the other species.

We downloaded the GFF annotation file for 27,655 protein-coding genes and 325 pri-miRNAs from the Araport11 genome release (Cheng et al., 2017) and extracted gene body sequences using BEDTools getfasta v2.30.0. We also extracted the promoter region of 2kb upstream of the transcription start site for lincRNAs, pri-miRNAs, and protein-coding genes. Noted that the promoter of a gene was shortened when it overlapped with a nearby gene sequence.

The sequence conservation of gene body and promoter region per gene type was evaluated using phastCons scores, which were calculated using the alignments of 20 angiosperm plant genomes (Hupalo and Kern, 2013). The phasCons scores were downloaded from the araTha9 genome browser available at genome.genetics.rutgers.edu as a bedgraph file. The bedgraph, consisting of variable width bin of equal phastCons score, was reprocessed to use a fix bin width of 1nt. In case of absence of a phasCons score on a portion of the genome, no bin was created which allow making the difference between nucleotides with a score (informative nucleotides) or absence of score (non-informative nucleotides). The average phastCons score of each annotation or promoter was computed using the BEDTools map v2.30.0 with the "–c4 –o mean" options, giving the average phastCons score using only informative nucleotides. The number of informative nucleotides of the genome proportion with phasCons score was computed using BEDTools intersect v2.30.0 Only for loci with at least 50% of informative nucleotides average phastCons scores were computed and reported in Figure 2 B-C (other loci were discarded).

Expression analysis of lincRNAs

To generate an expression atlas for lincRNAs, pri-miRNAs and protein-coding genes, we used Curse (Vaneechoutte and Vandepoele, 2019) to search and curate relevant RNA-seq experiments. Details of the 18 RNA-seq experiments across all 791 samples are shown in Supplemental Table S2 and Supplemental Data Set S3. Next, we imported the expression metadata to the Prose tool (Vaneechoutte and Vandepoele, 2019), which downloads the raw data from SRA by using the SRA toolkit, performs quality control and adapter detection,

trimmomatic to perform adapter clipping and quality trimming and finally Kallisto (Pertea et al., 2015) for quantifying transcript expression to normalized transcripts per million (TPM) values. We downloaded the transcript FASTA file for 27,655 protein-coding genes and 325 pri-miRNAs from the Araport11 genome release (Cheng et al., 2017) and retrieved transcript sequences for 6,599 lincRNAs using gffread (Pertea and Pertea, 2020). Gene-level atlases were created by summing the TPM values of all transcripts. We retained TPM value per gene across the biological replicates and took an average of TPM values per gene from technical replicates, resulting in an expression atlas covering 791 samples.

We used a simulation experiment (Ramskold et al., 2009; Li et al., 2016) to determine thresholds for detectable expression in TPM for protein-coding genes and lincRNAs. The protein-coding genes were used as true positives and the lincRNAs were used as true negatives. We calculated the false-positive rate and false-negative rate at different TPM thresholds for each sample. The applied cutoff of 0.2 TPM to define lincRNA and Pri-miRNAs expression is a good trade-off to detect lowly expressed lincRNAs and keep false-positives under control (false positive rate<0.10). A more stringent threshold TPM >= 2 was used to define the expression of protein-coding genes.

To identify the clusters containing samples with related gene expression information, a $\log_2$-transformed gene expression matrix (TPM+1) was used for t-SNE clustering with perplexity 30 and n_iter = 1000 (Pedregosa et al., 2011). 22 clusters were retained, the largest of which contained 113 samples and the smallest contained 5 samples. Twenty-two samples were not clustered and removed (Supplemental Data Set S3). Expression specificity per gene type, for the gene expression matrix described above, was measured using tissue-specificity index (tau), defined by the following equation (Yanai et al., 2005):

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1}, \hat{X}_i = \frac{xi}{max_{1\leq i\leq n}(xi)}$$

Where x_i was defined as the average TPM value of per cluster and n corresponds to the number of clusters analyzed.

Groups of genes showing similar expression profiles reported in Figure 6 B-D were identified using CAST clustering and only clusters with at least 10 genes were retained (Vandepoele et al., 2009).

Chromatin states analysis and TF peak annotation

The enrichment of the lincRNA sets in different chromatin states (CSs) of Arabidopsis from (Liu et al., 2018) was done by shuffling the lincRNA genome coordinates 1000 times over the whole Arabidopsis genome. Then, we compared the distribution of the number of lincRNAs expected to overlap by chance with each CS with the real number of overlaps, and we used these values to calculate enrichment statistics: p-value as the number of times the real overlap was higher than the overlap with any of the 1000 shuffled lincRNA sets, and enrichment fold as the real overlap divided by the median of overlap expected by chance (median of the 1000 shuffling events). The p-value was adjusted for multiple testing using Benjamini-Hochberg correction (significance level 0.05). For visualization purposes, the two enrichment metrics were combined into the π-value, which is the -log10(p-value)*enrichment fold.

TF ChIP-Seq peak coordinates were retrieved from the PlantPAN 3.0 database (Chow et al., 2019) and (Song et al., 2016). The original ChIP-Seq of KAN1, MYB44 and PIF4 were derived from these studies (Merelo et al., 2013; Pfeiffer et al., 2014; Song et al., 2016). For each peak the closest gene was identified and only peaks confirmed in two or more replicates and within a 2kb window of the gene body were retained. Starting from all TF – target gene pairs (either a protein-coding gene or a lincRNAs; Supplemental Data Set S5), enrichment analysis was performed to identify enriched TFs in different expression clusters or age categories. For all enrichment analyses the hypergeometric distribution was applied and the q-value of enrichment was determined using the Benjamini–Hochberg correction for multiple hypotheses testing.

Identification of GWAS-associated genes

GWAS data were collected from AraGWAS (Togninalli et al., 2020) and overlapped with the gene body and promoter 2kb sequences of lincRNAs to associate with the phenotype of interest. All significant associations (Permutation threshold <0.05 and FDR <0.05) were retained for screening for minor allele frequency>0.01, resulting in 1,124 lincRNAs, covering 147 different studies. These significantly associated lincRNAs were classified into five main traits by ontological annotation, including root, seed, flower, leaf, and abiotic-related traits (see Supplemental Data Set S6).

Overlap with accessible chromatin regions

ATAC-seq data for Arabidopsis root hair, non-hair and whole roots were collected from three publications (Maher et al., 2018; Potter et al., 2018; Tannenbaum et al., 2018), and scATAC-seq data for Arabidopsis root epidermis, endodermis, stele (pericyle, xylem, phloem), and cortex cells were collected from (Dorrity et al., 2021). Cell type-specific marker peaks were identified by scATAC-seq data (p<0.05 and avg_lofFC>0). BEDtools intersect v2.30.0 (Quinlan and Hall, 2010) was used to detect whether TF ChIP-seq peaks associated with lincRNAs overlapped with ACR peaks. Only overlaps between ACRs and TF ChIP peaks in at least two replicates were retained, requiring that at least of 10% of the ChIP peak was covered by the ACR. BEDtools shuffle v2.30.0 (with parameter -chrom) was used to shuffle the TF ChIP-seq peaks associated with lincRNAs.

Validation of TF-lincRNA regulatory interactions using reverse transcription-quantitative PCR

Total RNA from five to eight 14-day roots grown vertically on solid MS ½ media was extracted using TRI Reagent (Sigma-Aldrich) and digested with RNase-free DNase (Fermentas) following the manufacturer's recommendations. cDNA was synthetized using Maxima Reverse Transcriptase (Thermo Scientific). Expression analysis by RT-qPCR was performed using SYBR Green master I (Roche) and the LightCycler® 96 system following a standard protocol (40 cycles, 60°C

annealing). Data were analyzed using the ΔΔCt method with PP2A (PROTEIN PHOSPHATASE 2A SUBUNIT A3 (AT1G13320)) as reference transcript for normalization of RT-qPCR data. WT Col-0 plants grown at the same time were used as sample reference. For the dexamethasone inducible KAN1-GR expression lines analysis, after 14 days on MS media, the plants were transferred for one day either on dexamethasone containing plates (10µM) or only DMSO (dexamethasone solvent) as control. Primers used are listed in Supplemental Table S9. Three biological replicates were performed per condition. Statistical analyses were performed using the unpaired two-tailed Student's t-Test (GraphPad prism).

SUPPLEMENTAL DATA

Supplemental Figure S1. Integration of lincRNA resources.

Supplemental Figure S2. Definition of expressed genes and influence of sequencing depth.

Supplemental Figure S3. Resource annotation for highly-specific expressed lincRNAs.

Supplemental Figure S4. Regulatory properties of lincRNAs.

Supplemental Figure S5. TF binding and expression.

Supplemental Figure S6. RT-qPCR validation results.

Supplemental Figure S7. GWAS traits associated with lincRNAs confirmed by ChIP-Seq TF binding.

Supplemental Figure S8. Identification of cell type-specific TF-lincRNA regulatory interactions.

Supplemental Table S1. The number of homologous lincRNAs aligned to the genomes of 40 species.

Supplemental Table S2. Summary of RNA sequencing data.

Supplemental Table S3. List of lincRNAs with known functions.

Supplemental Table S4. Overview of TFs ChIP-Seq data used in this study.

Supplemental Table S5. TF and the absolute number of target genes per gene type.

Supplemental Table S6. Overview of TFs showing significant enrichment for binding to highly-specific expressed lincRNAs for the different expression clusters (corresponds to Figure 5B).

Supplemental Table S7. Overview of lincRNA peak annotation and RT-qPCR validation.

Supplemental Table S8. RT-qPCR confirmed/associated with GWAS root traits lincRNAs expression information for the different samples of expression cluster 19 (root) (corresponds to Figure 6B-D) and regulatory interactions detected in ACRs (corresponds to Supplemental Figure S8).

Supplemental Table S9. RT-qPCR primer sequences used in this study.

Supplemental Data Set S1. Full list of putative lincRNAs integrated from different resources.

Supplemental Data Set S2. PhastCons Scores for lincRNA loci and evolutionary age category.

Supplemental Data Set S3. Metadata for 791 samples of RNA-seq and cluster information.

Supplemental Data Set S4. The list of the highly-specific lincRNAs.

Supplemental Data Set S5. The table of TF-lincRNA interactions.

Supplemental Data Set S6. The lincRNA loci containing significant GWAS hits.

AUTHOR CONTRIBUTIONS

L.L., M.C., T.B., and K.V., designed the research. L.L., T.D., and T.B., performed data analysis. N.M.P. performed chromatin state analysis. M.H. performed experimental work. L.L., T.B., and K.V., wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Mariek Dubois, Andrés Ritter, Tereza Vavrdova, Freya De Winter, Rebecca De Clercq for kindly providing the Arabidopsis lines. We thank Olivier Martin and Federico Ariel for providing comments on the manuscript. The IPS2 is benefited from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). L.L. is supported by the China Scholarship Council for a PhD fellowship (201808530499). M.H. is supported by a PhD fellowship from the Fondation pour la Recherche Médicale (ECO202106013730).

FIGURE LEGENDS

Figure 1. Overlap and gene features of Arabidopsis lincRNA annotations. (A) Upset plot showing the intersection of lincRNA annotation in the eight resources. Each row represents a resource, reporting in parenthesis its total number of lincRNA transcripts before merging. LincRNA annotations unique to a single

resource are represented as a single circle while circles connected by lines represent the intersection of lincRNA loci shared between various resources. The bar chart indicates the number of unique lincRNA loci and intersectional lincRNA loci, displaying only intersections that contain at least ten lincRNA loci. More complex overlapping patterns are not shown. **(B)** The pie chart shows the proportion of lincRNA loci supported by one or more resources. **(C)** The distribution of exon number for all lincRNA transcripts (purple), protein-coding transcripts (green), transcripts of lincRNAs supported by single resource (purple) and multiple resources (purple). Single exon and multiple exons are shown in dark and light colors, respectively. **(D)** The distribution of transcript length for lincRNAs (purple) and protein-coding genes (green).

Figure 2. Sequence conservation analysis for lincRNAs of different evolutionary age categories. (A) Simplified species tree reporting the number of lincRNAs found for the different age categories, which are also indicated by the grey circle sizes. Numbers in parenthesis report the number of genomes included per clade to assess sequence similarity and define a lincRNA's age category. Boxplot showing the average phastCons score for **(B)** gene bodies and **(C)** promoter regions (2kb upstream of transcription start site) of different lincRNA age categories and gene types (lincRNAs, pri-miRNAs, protein-coding genes). The numbers in parentheses report the number of gene bodies and promoters with at least 50% of informative nucleotides over the total number of gene bodies and promoters in that category, respectively. PhastCons score ranges from 0 (not under selection) to 1 (strong negative selection). P-values for pairwise Mann–Whitney *U* test are shown using the horizontal lines connecting the series.

Figure 3. Expression analysis of lincRNAs. (A) Line chart showing the distribution of expression breadth for lincRNAs, pri-miRNAs and protein-coding genes across all samples **(B)** Distribution of the maximum TPM expression levels for different lincRNA age categories and gene types (lincRNAs, pri-miRNAs and protein-coding genes). **(C)** Distribution of tissue specificity tau scores for different

lincRNA age categories and gene types. Tau scores range from zero to one, where zero means widely expressed, and one means very specifically expressed (detectable in only one cluster). The black dotted line represents a tau score of 0.97. **(D)** Bar chart reporting the number of expressed and highly-specific expressed lincRNAs for the different expression clusters. Cluster numbers and descriptions are shown below the chart, with numbers in parenthesis indicating the number of samples present per cluster. **(E)** Heatmap showing the proportion of highly-specific expressed lincRNAs in each cluster for each lincRNA age category. Cluster descriptions are the same as in panel D. Numbers in parenthesis report the number of genes per age category together with the fraction of lincRNAs showing highly-specific expression.

Figure 4. Chromatin state and TF ChIP-Seq peak annotation for different gene types (A) Dendrogram showing the enrichment for different lincRNA gene sets (x-axis) towards different chromatin states (CS) (y-axis). The values report the product of -log10(q-value) and the enrichment fold. Only significant enrichment values are reported (q-value < 0.05). **(B)** The proportion of peak-gene pairs present in single replicate (blue) and two or more ChIP-Seq replicates (grey). **(C)** The percentage of three gene types assigned to peaks in two or more ChIP-Seq replicates. **(D)** Distributions of the number of TF binding events for lincRNA evolutionary age categories and gene types (lincRNAs, pri-miRNAs and protein-coding genes).

Figure 5. Overview of TF-lincRNA regulatory interactions in different expression clusters and age categories. The dot sizes represent the number of the lincRNAs while the color represents the statistical significance. Cluster descriptions are the same as in Figure 3D. **(A)** Bubble chart showing the enrichment of TF binding for expressed lincRNAs in different expression clusters. TFs lacking significant enrichment in any of the 22 clusters are not shown. **(B)** Bubble chart showing the enrichment of TF binding for highly-expressed lincRNAs in different expression clusters. TFs lacking significant enrichment in

any of the 22 clusters are not shown. **(C)** Bubble chart showing the enrichment of TF binding for expressed lincRNAs in different age categories.

Figure 6. Experimental validation and characterization of TF-lincRNA regulatory interactions. (A) Heatmap showing log2 fold change (FC) of lincRNA relative expression levels in transcription factor (TF) overexpressing lines (35S::*MYB44*, 35S::*PIF4* and *KAN1-GR*) or TF knockout lines (*rga28* and *pifq* (*pif1/pif2/pif3/pif4* quadruple mutants)) vs. control wild-type lines in 14-day old roots. Expression values were determined by RT-qPCR. Asterisks indicate statistically significant differences (*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$) in an unpaired two-tailed Student's t-Test (n =3). Solid boxes indicate TF ChIP peaks <2kb from the lincRNA gene. Dashed boxes indicate TF ChIP peaks were identified only in one ChIP-Seq replicate or the lincRNA is not the closest gene to the TF peak. **(B)** Heatmap showing KAN1 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. The first line shows the expression profile of the TF under investigation while the other lines show groups of target genes with similar expression. Next to the group number, the number of target genes is reported and in parenthesis the fraction of lincRNAs per group is indicated. The numbers in squares and circles report the number of GWAS root-related and RT-qPCR experimentally validated lincRNA genes, respectively. **(C)** Heatmap showing MYB44 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. Group information is the same as in panel B. **(D)** Heatmap showing PIF4 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. Group information is the same as in panel B, but here triangles and circles report the RT-qPCR experimentally validated lincRNA genes for the *pifq* and *35S::PIF4* lines, respectively.

TABLES

Table 1. Summary of qPCR validation for TF-lincRNA gene pairs.

Line	Total number of validated lincRNAs (1)	Has a peak and is DE (2)	No peak and not DE	Has a peak and is not DE	No peak and is DE	Has a putative peak and is DE (3)	Accuracy (4)	Precision	Recall
35S:*MYB44*	24	11	5	6	0	2	0.73	0.65	1.00
35S:*PIF4*	26	8	4	9	4	1	0.48	0.47	0.67
KAN1-GR	24	6	6	7	3	2	0.55	0.46	0.67
Pifq	25	8	6	8	1	2	0.61	0.50	0.89
rga-28	25	3	11	8	2	1	0.58	0.27	0.60

(1) In total 27 lincRNAs were tested in one or more lines.

(2) DE refers to differential expression, indicating deregulation in a TF perturbation line.

(3) These peak-lincRNA pairs do not satisfy our stringent peak definition (present in two or more replicates and present within 2kb of the gene body). Here a peak is only identified in one ChIP-Seq replicate or the lincRNA is not the closest target gene to this peak.

(4) The genes present in the category “Has a putative peak and is DE” were excluded to compute Accuracy, Precision and Recall.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215,** 403-410.

Ariel, F., Lucero, L., Christ, A., Mammarella, M.F., Jegu, T., Veluchamy, A., Mariappan, K., Latrasse, D., Blein, T., Liu, C., Benhamed, M., and Crespi, M. (2020). R-Loop Mediated trans Action of the APOLO Long Noncoding RNA. Mol Cell **77,** 1055-1065 e1054.

Bardou, F., Ariel, F., Simpson, C.G., Romero-Barrios, N., Laporte, P., Balzergue, S., Brown, J.W., and Crespi, M. (2014). Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. Dev Cell **30,** 166-176.

Bazin, J., Romero, N., Rigo, R., Charon, C., Blein, T., Ariel, F., and Crespi, M. (2018). Nuclear Speckle RNA Binding Proteins Remodel Alternative Splicing and the Non-coding Arabidopsis Transcriptome to Regulate a Cross-Talk Between Auxin and Immune Responses. Frontiers in Plant Science **9**.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc Natl Acad Sci U S A **107,** 18724-18728.

Bhogireddy, S., Mangrauthia, S.K., Kumar, R., Pandey, A.K., Singh, S., Jain, A., Budak, H., Varshney, R.K., and Kudapa, H. (2021). Regulatory non-coding RNAs: a new frontier in regulation of plant biology. Functional & Integrative Genomics **21,** 313-330.

Blein, T., Balzergue, C., Roule, T., Gabriel, M., Scalisi, L., Francois, T., Sorin, C., Christ, A., Godon, C., Delannoy, E., Martin-Magniette, M.L., Nussaume, L., Hartmann, C., Gautheret, D., Desnos, T., and Crespi, M. (2020). Landscape of the Noncoding Transcriptome Response of Two Arabidopsis Ecotypes to Phosphate Starvation. Plant Physiol **183,** 1058-1072.

Bu, D.C., Luo, H.T., Jiao, F., Fang, S.S., Tan, C.F., Liu, Z.Y., and Zhao, Y. (2015). Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. Science China-Life Sciences **58,** 787-798.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev **25,** 1915-1927.

Chen, L., Zhu, Q.H., and Kaufmann, K. (2020). Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. Planta **252,** 92.

Chen, Q.S., Liu, K., Yu, R., Zhou, B.L., Huang, P.P., Cao, Z.X., Zhou, Y.Q., and Wang, J.H. (2021). From "Dark Matter" to "Star": Insight Into the Regulation Mechanisms of Plant Functional Long Non-Coding RNAs. Frontiers in Plant Science **12**.

Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J **89,** 789-804.

Chow, C.N., Lee, T.Y., Hung, Y.C., Li, G.Z., Tseng, K.C., Liu, Y.H., Kuo, P.L., Zheng, H.Q., and Chang, W.C. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. Nucleic Acids Research **47,** D1155-D1163.

Corona-Gomez, J.A., Coss-Navarrete, E.L., Garcia-Lopez, I.J., Klapproth, C., Perez-Patino, J.A., and Fernandez-Valverde, S.L. (2022). Transcriptome-guided annotation and functional classification of long non-coding RNAs in Arabidopsis thaliana. Sci Rep **12,** 14063.

Deng, F., Zhang, X., Wang, W., Yuan, R., and Shen, F. (2018). Identification of Gossypium hirsutum long non-coding RNAs (lncRNAs) under salt stress. BMC Plant Biol **18,** 23.

Devaiah, B.N., Nagarajan, V.K., and Raghothama, K.G. (2007). Phosphate homeostasis and root development in Arabidopsis are synchronized by the zinc finger transcription factor ZAT6. Plant Physiol **145,** 147-159.

Dorrity, M.W., Alexandre, C.M., Hamm, M.O., Vigil, A.L., Fields, S., Queitsch, C., and Cuperus, J.T. (2021). The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution. Nat Commun **12,** 3334.

Fang, S.S., Zhang, L.L., Guo, J.C., Niu, Y.W., Wu, Y., Li, H., Zhao, L.H., Li, X.Y., Teng, X.Y., Sun, X.H., Sun, L., Zhang, M.Q., Chen, R.S., and Zhao, Y. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Research **46,** D308-D314.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44,** D279-285.

Fukuda, M., Nishida, S., Kakei, Y., Shimada, Y., and Fujiwara, T. (2019). Genome-Wide Analysis of Long Intergenic Noncoding RNAs Responding to Low-Nutrient Conditions in Arabidopsis thaliana: Possible Involvement of Trans-Acting siRNA3 in Response to Low Nitrogen. Plant Cell Physiol **60,** 1961-1973.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., Dewar, K., Stinchcombe, J.R., Schoen, D.J., Wang, X., Schmutz, J., Town, C.D., Edger, P.P., Pires, J.C., Schumaker, K.S., Jarvis, D.E., Mandakova, T., Lysak, M.A., van den Bergh, E., Schranz, M.E., Harrison, P.M., Moses, A.M., Bureau, T.E., Wright, S.I., and Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet **45,** 891-898.

Hawker, N.P., and Bowman, J.L. (2004). Roles for Class III HD-Zip and KANADI genes in Arabidopsis root development. Plant Physiol **135,** 2261-2270.

Hazarika, R.R., Serra, M., Zhang, Z., Zhang, Y., Schmitz, R.J., and Johannes, F. (2022). Molecular properties of epimutation hotspots. Nat Plants **8,** 146-156.

He, H., Zhou, Y.F., Yang, Y.W., Zhang, Z., Lei, M.Q., Feng, Y.Z., Zhang, Y.C., Chen, Y.Q., Lian, J.P., and Yu, Y. (2021). Genome-Wide Analysis Identified a Set of Conserved lncRNAs Associated with Domestication-Related Traits in Rice. International Journal of Molecular Sciences **22**.

Heo, J.B., and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science **331,** 76-79.

Heyndrickx, K.S., Van de Velde, J., Wang, C., Weigel, D., and Vandepoele, K. (2014). A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. Plant Cell **26,** 3894-3910.

Hu, L., Xu, Z., Hu, B., and Lu, Z.J. (2017). COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. Nucleic Acids Res **45,** e2.

Huang, Y., Sicar, S., Ramirez-Prado, J.S., Manza-Mianza, D., Antunez-Sanchez, J., Brik-Chaouche, R., Rodriguez-Granados, N.Y., An, J., Bergounioux, C., Mahfouz, M.M., Hirt, H., Crespi, M., Concia, L., Barneche, F., Amiard, S., Probst, A.V., Gutierrez-Marcos, J., Ariel, F., Raynaud, C., Latrasse, D., and Benhamed, M. (2021). Polycomb-dependent differential chromatin compartmentalization determines gene coregulation in Arabidopsis. Genome Research **31,** 1230-+.

Hupalo, D., and Kern, A.D. (2013). Conservation and functional element discovery in 20 angiosperm plant genomes. Mol Biol Evol **30,** 1729-1744.

Jha, U.C., Nayyar, H., Jha, R., Khurshid, M., Zhou, M., Mantri, N., and Siddique, K.H.M. (2020). Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation. BMC Plant Biol **20,** 466.

Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. Annual Review of Plant Biology **57,** 19-53.

Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S.R., Finn, R.D., Bateman, A., and Petrov, A.I. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res **49,** D192-D200.

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

Ke, L.L., Zhou, Z.W., Xu, X.W., Wang, X., Liu, Y.L., Xu, Y.T., Huang, Y., Wang, S.T., Deng, X.X., Chen, L.L., and Xu, Q. (2019). Evolutionary dynamics of lincRNA transcription in nine citrus species. Plant Journal **98,** 912-927.

Kim, D.H., and Sung, S. (2017). Vernalization-Triggered Intragenic Chromatin Loop Formation by Long Noncoding RNAs. Dev Cell **40,** 302-312 e304.

Kindgren, P., Ard, R., Ivanov, M., and Marquardt, S. (2018). Transcriptional read-through of the long non-coding RNA SVALKA governs plant cold acclimation. Nat Commun **9,** 4561.

Kramer, M.C., Kim, H.J., Palos, K.R., Garcia, B.A., Lyons, E., Beilstein, M.A., Nelson, A.D.L., and Gregory, B.D. (2022). A Conserved Long Intergenic Non-coding RNA Containing snoRNA Sequences, lncCOBRA1, Affects Arabidopsis Germination and Development. Front Plant Sci **13,** 906603.

Li, Q.Q., Zhang, Z., Zhang, C.X., Wang, Y.L., Liu, C.B., Wu, J.C., Han, M.L., Wang, Q.X., and Chao, D.Y. (2022). Phytochrome-interacting factors orchestrate hypocotyl adventitious root initiation in Arabidopsis. Development **149**.

Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P.N. (2016). High-Resolution Expression Map of the Arabidopsis Root Reveals Alternative Splicing and lincRNA Regulation. Dev Cell **39,** 508-522.

Li, S.X., Yu, X., Lei, N., Cheng, Z.H., Zhao, P.J., He, Y.K., Wang, W.Q., and Peng, M. (2017). Genome-wide identification and functional prediction of cold and/or drought-responsive lncRNAs in cassava. Scientific Reports **7**.

Liu, F., Marquardt, S., Lister, C., Swiezewski, S., and Dean, C. (2010). Targeted 3' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing. Science **327,** 94-97.

Liu, J., Wang, H., and Chua, N.H. (2015). Long noncoding RNA transcriptome of plants. Plant Biotechnol J **13,** 319-328.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S.L., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012). Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis. Plant Cell **24,** 4333-4345.

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., Yi, X., Xu, W., and Su, Z. (2018). PCSD: a plant chromatin state database. Nucleic Acids Res **46,** D1157-D1167.

Lucero, L., Ferrero, L., Fonouni-Farde, C., and Ariel, F. (2021). Functional classification of plant long noncoding RNAs: a transcript is known by the company it keeps. New Phytologist **229,** 1251-1260.

Ma, J.C., Bai, X.T., Luo, W.C., Feng, Y.N., Shao, X.M., Bai, Q.X., Sun, S.J., Long, Q.M., and Wan, D.S. (2019). Genome-Wide Identification of Long Noncoding RNAs and Their Responses to Salt Stress in Two Closely Related Poplars. Frontiers in Genetics **10**.

Maher, K.A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D.A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M.W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S.M., and Deal, R.B. (2018). Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. Plant Cell **30,** 15-36.

Mattioli, K., Volders, P.J., Gerhardinger, C., Lee, J.C., Maass, P.G., Mele, M., and Rinn, J.L. (2019). High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. Genome Research **29,** 344-355.

Merelo, P., Xie, Y.K., Brand, L., Ott, F., Weigel, D., Bowman, J.L., Heisler, M.G., and Wenkel, S. (2013). Genome-Wide Identification of KANADI1 Target Genes. Plos One **8**.

Mohammadin, S., Edger, P.P., Pires, J.C., and Schranz, M.E. (2015). Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. BMC Plant Biol **15,** 217.

Moison, M., Pacheco, J.M., Lucero, L., Fonouni-Farde, C., Rodriguez-Melo, J., Mansilla, N., Christ, A., Bazin, J., Benhamed, M., Ibanez, F., Crespi, M., Estevez, J.M., and Ariel, F. (2021). The lncRNA APOLO interacts with the transcription factor WRKY42 to trigger root hair cell expansion in response to cold. Mol Plant **14,** 937-948.

Moubayidin, L., Salvi, E., Giustini, L., Terpstra, I., Heidstra, R., Costantino, P., and Sabatini, S. (2016). A SCARECROW-based regulatory circuit controls Arabidopsis thaliana meristem size from the root endodermis. Planta **243,** 1159-1168.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29,** 2933-2935.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature **505,** 635-+.

Nelson, A.D.L., Devisetty, U.K., Palos, K., Haug-Baltzell, A.K., Lyons, E., and Beilstein, M.A. (2017). Evolinc:A Tool for the Identification and Evolutionary Comparison of Long Intergenic Non-coding RNAs. Frontiers in Genetics **8**.

Nelson, A.D.L., Forsythe, E.S., Devisetty, U.K., Clausen, D.S., Haug-Batzell, A.K., Meldrum, A.M.R., Frank, M.R., Lyons, E., and Beilstein, M.A. (2016). A Genomic Analysis of Factors Driving lincRNA Diversification: Lessons from Plants. G3-Genes Genomes Genetics **6,** 2881-2891.

O'Maoileidigh, D.S., Graciet, E., and Wellmer, F. (2014). Gene networks controlling Arabidopsis thaliana flower development. New Phytol **201,** 16-30.

Pacheco, J.M., Mansilla, N., Moison, M., Lucero, L., Gabarain, V.B., Ariel, F., and Estevez, J.M. (2021). The lncRNA APOLO and the transcription factor WRKY42 target common cell wall EXTENSIN encoding genes to trigger root hair cell elongation. Plant Signal Behav **16,** 1920191.

Palos, K., Dittrich, A.C.N., Yu, L.A., Brock, J.R., Railey, C.E., Wu, H.Y.L., Sokolowska, E., Skirycz, A., Hsu, P.Y., Gregory, B.D., Lyons, E., Beilstein, M.A., and Nelson, A.D.L. (2022). Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. Plant Cell.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12,** 2825-2830.

Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Res **9**.

Pfeiffer, A., Shi, H., Tepperman, J.M., Zhang, Y., and Quail, P.H. (2014). Combinatorial Complexity in a Transcriptionally Centered Signaling Hub in Arabidopsis. Molecular Plant **7,** 1598-1618.

Ponting, C.P., and Haerty, W. (2022). Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. Annu Rev Genomics Hum Genet.

Potter, K.C., Wang, J., Schaller, G.E., and Kieber, J.J. (2018). Cytokinin modulates context-dependent chromatin accessibility through the type-B response regulators. Nat Plants **4,** 1102-1111.

Qi, X., Xie, S., Liu, Y., Yi, F., and Yu, J. (2013). Genome-wide annotation of genes and noncoding RNAs of foxtail millet in response to simulated drought stress by deep sequencing. Plant Mol Biol **83,** 459-473.

Qin, T., Zhao, H., Cui, P., Albesher, N., and Xiong, L. (2017). A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. Plant Physiol **175,** 1321-1336.

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

Rai, M.I., Alam, M., Lightfoot, D.A., Gurha, P., and Afzal, A.J. (2019). Classification and experimental identification of plant long non-coding RNAs. Genomics **111,** 997-1005.

Ramskold, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol **5,** e1000598.

Ransbotyn, V., Yeger-Lotem, E., Basha, O., Acuna, T., Verduyn, C., Gordon, M., Chalifa-Caspi, V., Hannah, M.A., and Barak, S. (2015). A combination of gene expression ranking and co-expression network analysis increases discovery rate in large-scale mutant screens for novel Arabidopsis thaliana abiotic stress genes. Plant Biotechnol J **13,** 501-513.

Ransohoff, J.D., Wei, Y., and Khavari, P.A. (2018). The functions and unique features of long intergenic non-coding RNA. Nat Rev Mol Cell Biol **19,** 143-157.

Rigo, R., Bazin, J., Romero-Barrios, N., Moison, M., Lucero, L., Christ, A., Benhamed, M., Blein, T., Huguet, S., Charon, C., Crespi, M., and Ariel, F. (2020). The Arabidopsis lncRNA ASCO modulates the transcriptome through interaction with splicing factors. EMBO Rep **21,** e48977.

Roule, T., Crespi, M., and Blein, T. (2022a). Regulatory long non-coding RNAs in root growth and development. Biochem Soc Trans **50,** 403-412.

Roule, T., Christ, A., Hussain, N., Huang, Y., Hartmann, C., Benhamed, M., Gutierrez-Marcos, J., Ariel, F., Crespi, M., and Blein, T. (2022b). The lncRNA MARS modulates the epigenetic reprogramming of the marneral cluster in response to ABA. Mol Plant **15,** 840-856.

Sanchita, Trivedi, P.K., and Asif, M.H. (2020). Updates on plant long non-coding RNAs (lncRNAs): the regulatory components. Plant Cell Tissue and Organ Culture **140,** 259-269.

Sarropoulos, I., Marin, R., Cardoso-Moreira, M., and Kaessmann, H. (2019). Developmental dynamics of lncRNAs across mammalian organs and species. Nature **571,** 510-+.

Schmitz, R.J., Grotewold, E., and Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell **34,** 718-741.

Severing, E., Faino, L., Jamge, S., Busscher, M., Kuijer-Zhang, Y., Bellinazzo, F., Busscher-Lange, J., Fernandez, V., Angenent, G.C., Immink, R.G.H., and Pajoro, A. (2018). Arabidopsis thaliana ambient temperature responsive lncRNAs. BMC Plant Biol **18,** 145.

Shafiq, S., Li, J.R., and Sun, Q.W. (2016). Functions of plants long non-coding RNAs. Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms **1859,** 155-162.

Shea, D.J., Nishida, N., Takada, S., Itabashi, E., Takahashi, S., Akter, A., Miyaji, N., Osabe, K., Mehraj, H., Shimizu, M., Seki, M., Kakizaki, T., Okazaki, K., Dennis, E.S., and Fujimoto, R. (2019). Long noncoding RNAs in Brassica rapa L. following vernalization. Sci Rep **9,** 9302.

Shuai, P., Liang, D., Tang, S., Zhang, Z.J., Ye, C.Y., Su, Y.Y., Xia, X.L., and Yin, W.L. (2014). Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in Populus trichocarpa. Journal of Experimental Botany **65,** 4975-4983.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15,** 1034-1050.

Song, L., Huang, S.S.C., Wise, A., Castanon, R., Nery, J.R., Chen, H.M., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science **354**.

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., and Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research **41**.

Sun, Z., Nair, A., Chen, X., Prodduturi, N., Wang, J., and Kocher, J.P. (2017). UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. Sci Rep **7,** 14196.

Szczesniak, M.W., Rosikiewicz, W., and Makalowska, I. (2016). CANTATAdb: A Collection of Plant Long Non-Coding RNAs. Plant Cell Physiol **57,** e8.

Szczesniak, M.W., Kubiak, M.R., Wanowska, E., and Makalowska, I. (2021). Comparative genomics in the search for conserved long noncoding RNAs. Essays Biochem **65,** 741-749.

Tannenbaum, M., Sarusi-Portuguez, A., Krispil, R., Schwartz, M., Loza, O., Benichou, J.I.C., Mosquna, A., and Hakim, O. (2018). Regulatory chromatin landscape in Arabidopsis thaliana roots uncovered by coupling INTACT and ATAC-seq. Plant Methods **14,** 113.

Togninalli, M., Seren, U., Freudenthal, J.A., Monroe, J.G., Meng, D., Nordborg, M., Weigel, D., Borgwardt, K., Korte, A., and Grimm, D.G. (2020). AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana. Nucleic Acids Res **48,** D1063-D1068.

Tominaga-Wada, R., and Wada, T. (2016). The ectopic localization of CAPRICE LIKE MYB3 protein in Arabidopsis root epidermis. J Plant Physiol **199,** 111-115.

Tran Vdu, T., Souiai, O., Romero-Barrios, N., Crespi, M., and Gautheret, D. (2016). Detection of generic differential RNA processing events from RNA-seq data. RNA Biol **13,** 59-67.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics **17,** 601-614.

Van Bel, M., Silvestri, F., Weitz, E.M., Kreft, L., Botzki, A., Coppens, F., and Vandepoele, K. (2021). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. Nucleic Acids Res.

Van de Velde, J., Van Bel, M., Vaneechoutte, D., and Vandepoele, K. (2016). A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. Plant Physiol **171,** 2586-2598.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. Plant Physiology **150,** 535-546.

Vaneechoutte, D., and Vandepoele, K. (2019). Curse: building expression atlases and co-expression networks from public RNA-Seq data. Bioinformatics **35,** 2880-2881.

Wang, H., Niu, Q.W., Wu, H.W., Liu, J., Ye, J., Yu, N., and Chua, N.H. (2015). Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits. Plant Journal **84,** 404-416.

Wu, H., Yang, L., and Chen, L.L. (2017). The Diversity of Long Noncoding RNAs and Their Generation. Trends in Genetics **33,** 540-552.

Xuan, H.D., Zhang, L.Z., Liu, X.S., Han, G.M., Li, J., Li, X., Liu, A.G., Liao, M.Z., and Zhang, S.H. (2015). PLNlncRbase: A resource for experimentally identified lncRNAs in plants. Gene **573,** 328-332.

Yamada, M. (2017). Functions of long intergenic non-coding (linc) RNAs in plants. J Plant Res **130,** 67-73.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21,** 650-659.

Yu, C.P., Lin, J.J., and Li, W.H. (2016). Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci Rep **6,** 25164.

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., Chang, X., Dong, W., Ho, S.Y.W., Liu, X., Song, A., Chen, J., Guo, W., Wang, Z., Zhuang,

Y., Wang, H., Chen, X., Hu, J., Liu, Y., Qin, Y., Wang, K., Dong, S., Liu, Y., Zhang, S., Yu, X., Wu, Q., Wang, L., Yan, X., Jiao, Y., Kong, H., Zhou, X., Yu, C., Chen, Y., Li, F., Wang, J., Chen, W., Chen, X., Jia, Q., Zhang, C., Jiang, Y., Zhang, W., Liu, G., Fu, J., Chen, F., Ma, H., Van de Peer, Y., and Tang, H. (2020). The water lily genome and the early evolution of flowering plants. Nature **577,** 79-84.

Zhao, Q., Li, M., Jia, Z., Liu, F., Ma, H., Huang, Y., and Song, S. (2016). AtMYB44 Positively Regulates the Enhanced Elongation of Primary Roots Induced by N-3-Oxo-Hexanoyl-Homoserine Lactone in Arabidopsis thaliana. Mol Plant Microbe Interact **29,** 774-785.

Zhao, X.Y., Li, J.R., Lian, B., Gu, H.Q., Li, Y., and Qi, Y.J. (2018). Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. Nature Communications **9**.

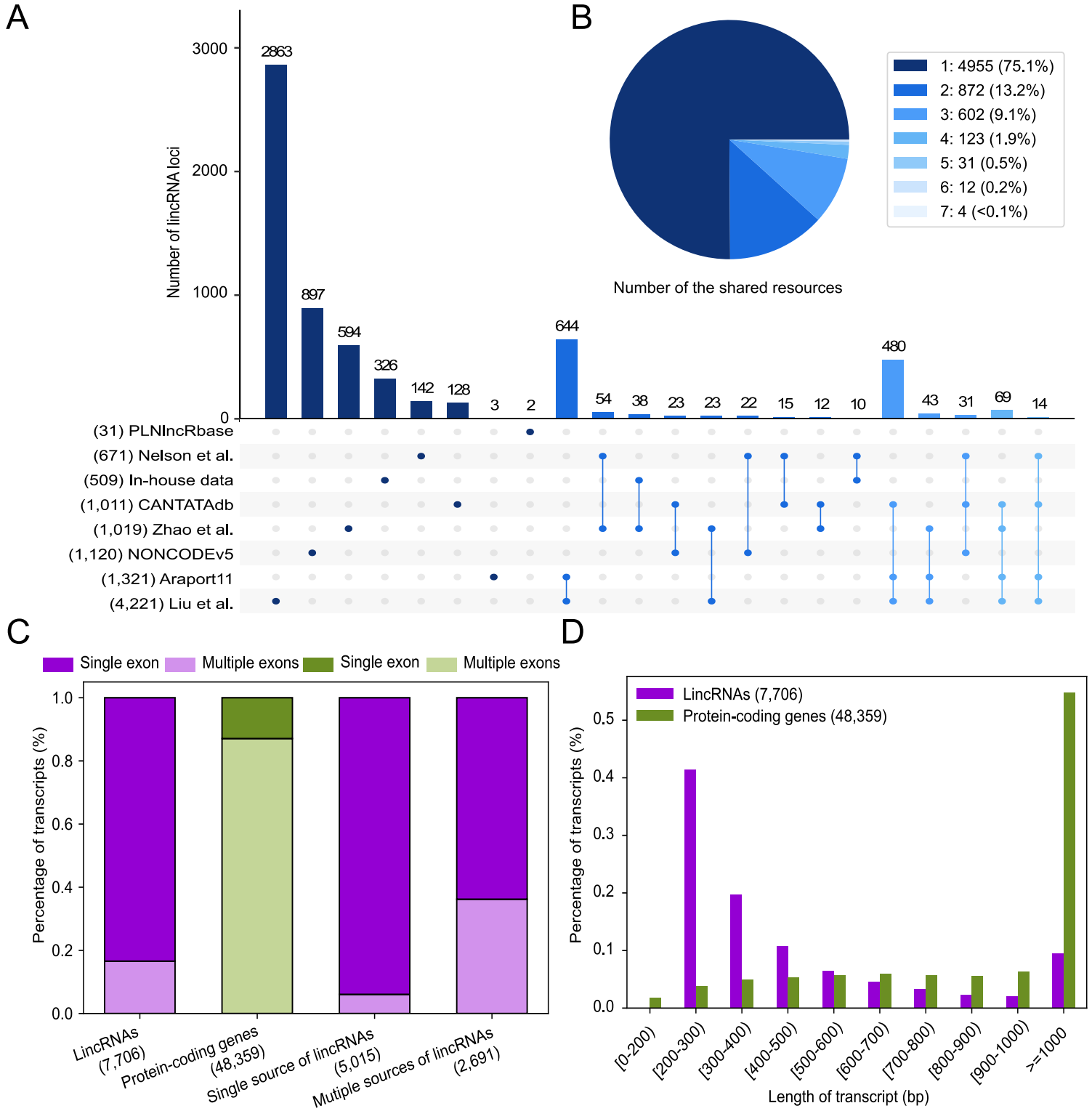

Figure 1. Overlap and gene features of Arabidopsis lincRNA annotations. (A) Upset plot showing the intersection of lincRNA annotation in the eight resources. Each row represents a resource, reporting in parenthesis the total number of lincRNA transcripts before merging. LincRNA annotations unique to a single resource are represented as a single circle while circles connected by lines represent the intersection of lincRNA loci shared between various resources. The bar chart indicates the number of unique lincRNA loci and intersectional lincRNA loci, displaying only intersections that contain at least ten lincRNA loci. More complex overlap patterns are not shown to maintain figure clarity. **(B)** The pie chart shows the proportion of lincRNA loci supported by one or more resources. **(C)** The distribution of exon number for all lincRNA transcripts (purple), protein-coding transcripts (green), transcripts of lincRNAs supported by single resource (purple) and multiple resources (purple). Single exon and multiple exons are shown in dark and light colors, respectively. **(D)** The distribution of transcript length for lincRNAs (purple) and protein-coding genes (green).

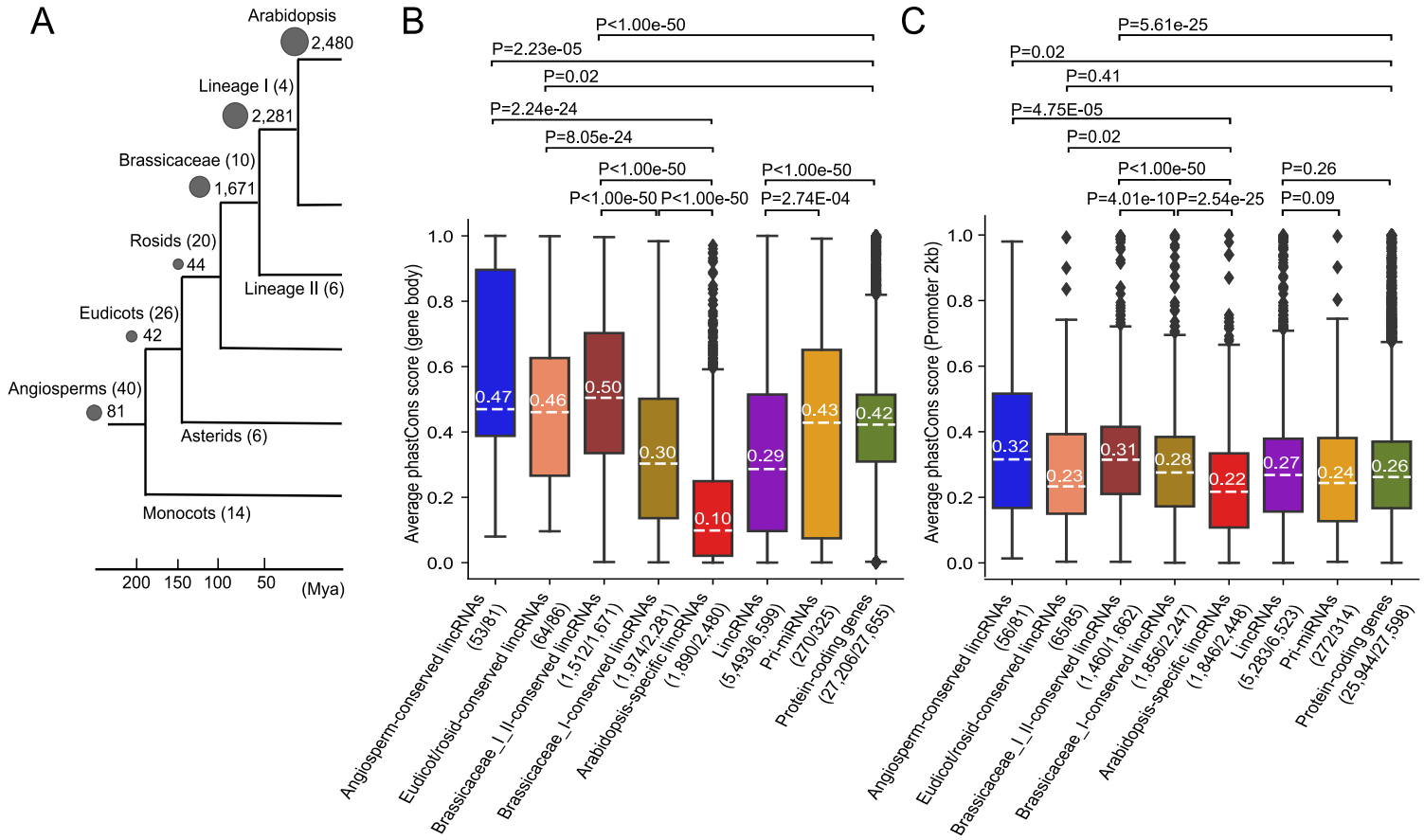

Figure 2. Sequence conservation analysis for lincRNAs of different evolutionary age categories. (A) Simplified species tree reporting the number of lincRNAs found for the different age categories, which are also indicated by the grey circle sizes. Numbers in parenthesis report the number of genomes included per clade to assess sequence similarity and define a lincRNA's age category. Boxplot showing the average phastCons score for **(B)** gene bodies and **(C)** promoter regions (2kb upstream of transcription start site) of different lincRNA age categories and gene types (lincRNAs, pri-miRNAs, protein-coding genes). The numbers in parentheses report the number of gene bodies and promoters with at least 50% of informative nucleotides over the total number of gene bodies and promoters in that category, respectively. PhastCons score ranges from 0 (not under selection) to 1 (strong negative selection). P-values for pairwise Mann–Whitney *U* test are shown using the horizontal lines connecting the series.

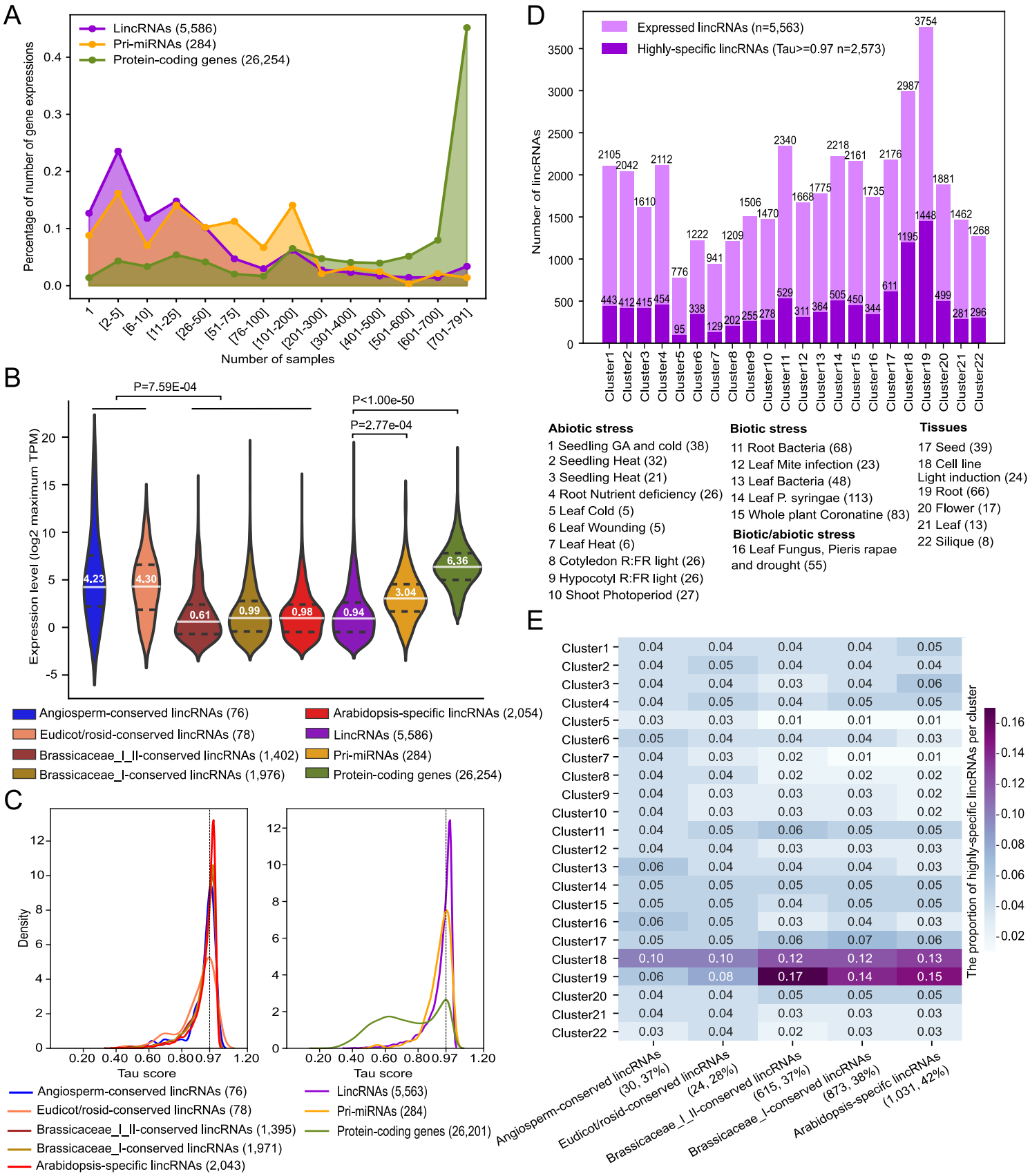

Figure 3. Expression analysis of lincRNAs. (A) Line chart showing the distribution of expression breadth for lincRNAs, pri-miRNAs and protein-coding genes across all samples **(B)** Distribution of the maximum TPM expression levels for different lincRNA age categories and gene types (lincRNAs, pri-miRNAs and protein-coding genes). **(C)** Distribution of tissue specificity tau scores for different lincRNA age categories and gene types. Tau scores range from zero to one, where zero means widely expressed, and one means very specifically expressed (detectable in only one cluster). The black dotted line represents a tau score of 0.97. **(D)** Bar chart reporting the number of expressed and highly-specific expressed lincRNAs for the different expression clusters. Cluster numbers and descriptions are shown below the chart, with numbers in parenthesis indicating the number of samples present per cluster. **(E)** Heatmap showing the proportion of highly-specific expressed lincRNAs in each cluster for each lincRNA age category. Cluster descriptions are the same as in panel D. Numbers in parenthesis report the number of genes per age category together with the fraction of lincRNAs showing highly-specific expression.

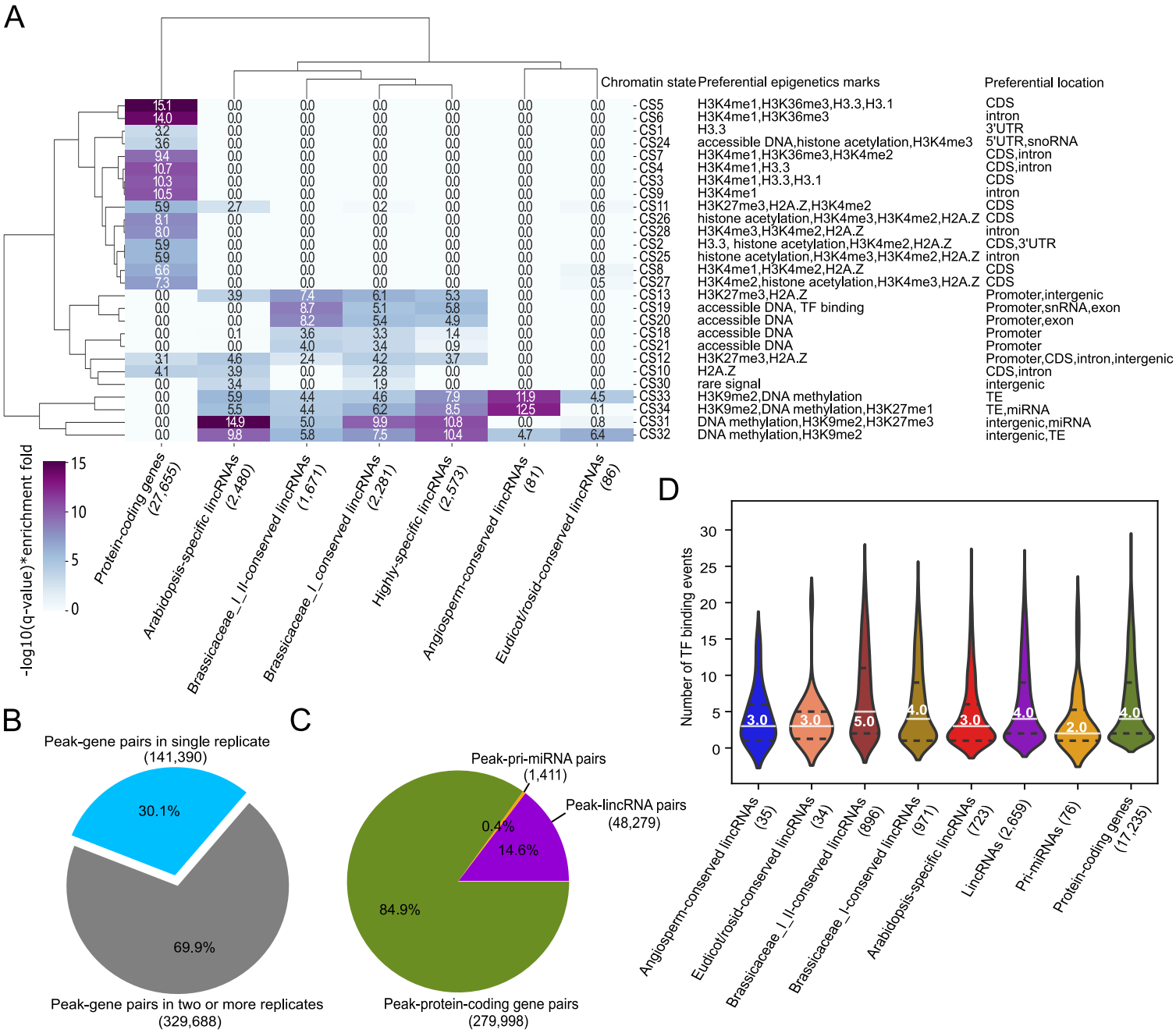

Figure 4. Chromatin state and TF ChIP-Seq peak annotation for different gene types. (A) Dendrogram showing the enrichment for different lincRNA gene sets (x-axis) towards different chromatin states (CS) (y-axis). The values report the product of -log10(q-value) and the enrichment fold. Only significant enrichment values are reported (q-value < 0.05). **(B)** The proportion of peak-gene pairs present in single replicate (blue) and two or more ChIP-Seq replicates (grey). **(C)** The percentage of three gene types assigned to peaks in two or more ChIP-Seq replicates. **(D)** Distributions of the number of TF binding events for lincRNA evolutionary age categories and gene types (lincRNAs, pri-miRNAs and protein-coding genes).

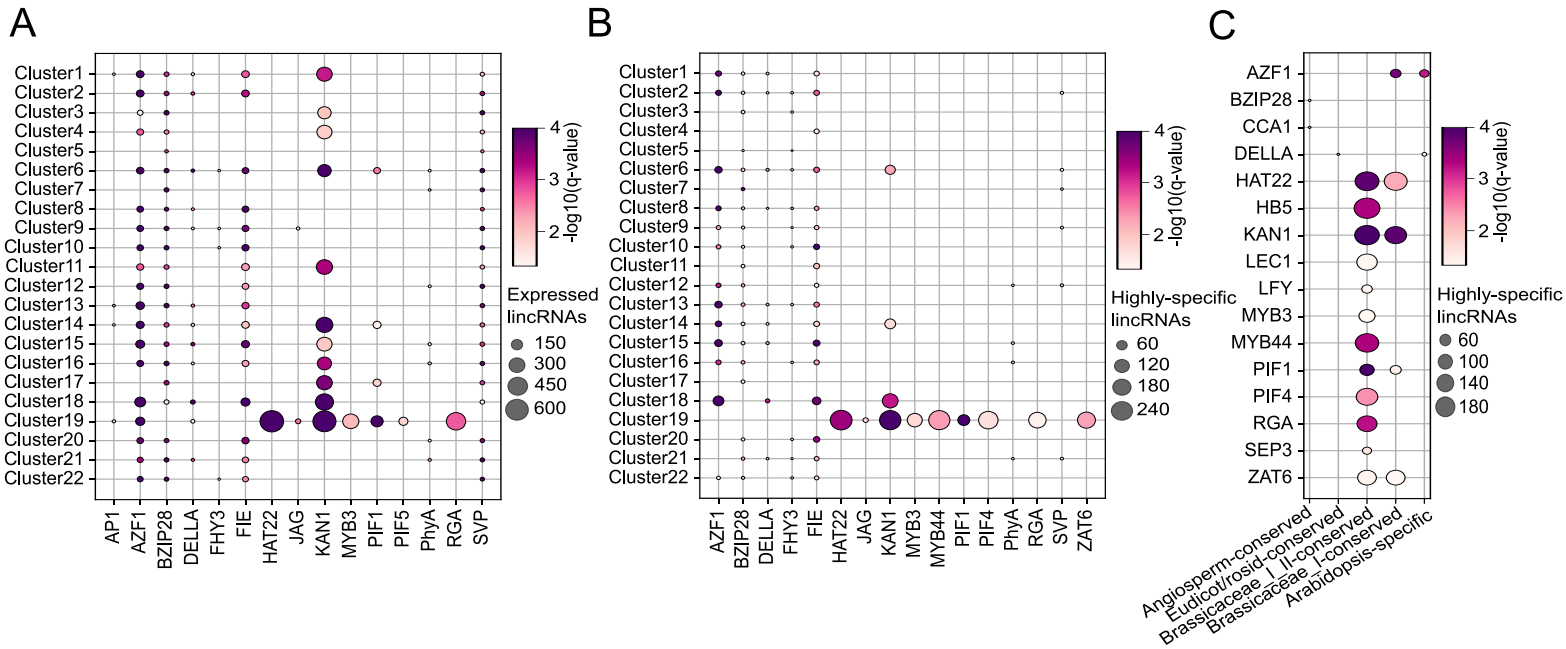

Figure 5. Overview of TF-lincRNA regulatory interactions in different expression clusters and age categories. The dot sizes represent the number of the lincRNAs while the color represents the statistical significance. Cluster descriptions are the same as in Figure 3D. **(A)** Bubble chart showing the enrichment of TF binding for expressed lincRNAs in different expression clusters. TFs lacking significant enrichment in any of the 22 clusters are not shown. **(B)** Bubble chart showing the enrichment of TF binding for highly-expressed lincRNAs in different expression clusters. TFs lacking significant enrichment in any of the 22 clusters are not shown. **(C)** Bubble chart showing the enrichment of TF binding for expressed lincRNAs in different age categories.

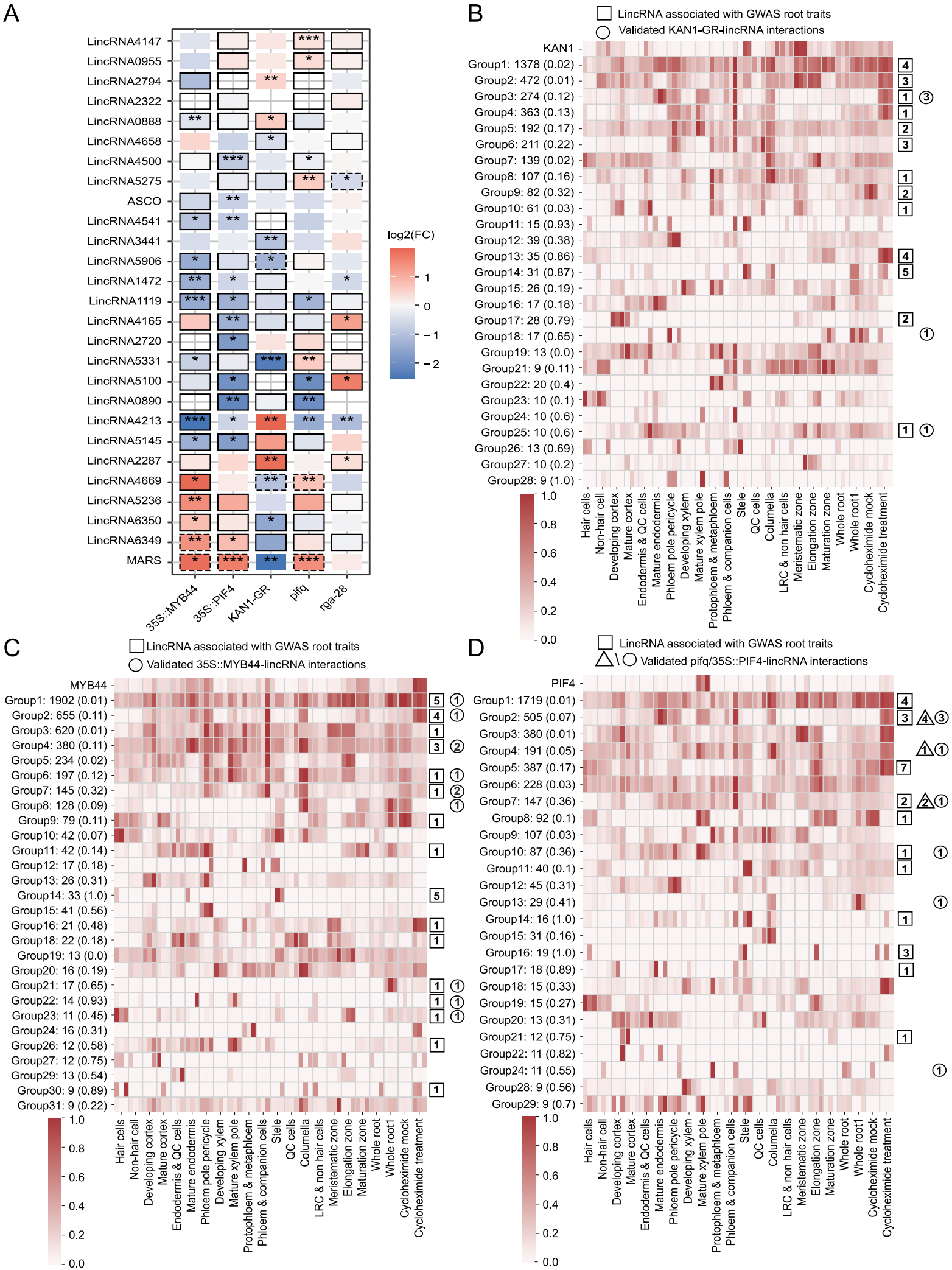
A
LincRNA4147
LincRNA0955
LincRNA2794
LincRNA2322
LincRNA0888
LincRNA4658
LincRNA4500
LincRNA5275
ASCO
LincRNA4541
LincRNA3441
LincRNA5906
LincRNA1472
LincRNA1119
LincRNA4165
LincRNA2720
LincRNA5331
LincRNA5100
LincRNA0890
LincRNA4213
LincRNA5145
LincRNA2287
LincRNA4669
LincRNA5236
LincRNA6350
LincRNA6349
MARS
35S::MYB44
35S::PIF4
KAN1-GR
pifq
rga-28
log2(FC)
1
0
−1
−2
B
LincRNA associated with GWAS root traits
Validated KAN1-GR-lincRNA interactions
KAN1
Group1: 1378 (0.02)
Group2: 472 (0.01)
Group3: 274 (0.12)
Group4: 363 (0.13)
Group5: 192 (0.17)
Group6: 211 (0.22)
Group7: 139 (0.02)
Group8: 107 (0.16)
Group9: 82 (0.32)
Group10: 61 (0.03)
Group11: 15 (0.93)
Group12: 39 (0.38)
Group13: 35 (0.86)
Group14: 31 (0.87)
Group15: 26 (0.19)
Group16: 17 (0.18)
Group17: 28 (0.79)
Group18: 17 (0.65)
Group19: 13 (0.0)
Group21: 9 (0.11)
Group22: 20 (0.4)
Group23: 10 (0.1)
Group24: 10 (0.6)
Group25: 10 (0.6)
Group26: 13 (0.69)
Group27: 10 (0.2)
Group28: 9 (1.0)
Hair cells
Non-hair cell
Developing cortex
Mature cortex
Endodermis & QC cells
Mature endodermis
Phloem pole pericycle
Developing xylem
Mature xylem pole
Protophloem & metaphloem
Phloem & companion cells
Stele
QC cells
Columella
LRC & non hair cells
Meristematic zone
Elongation zone
Maturation zone
Whole root
Whole root1
Cycloheximide mock
Cycloheximide treatment
C
LincRNA associated with GWAS root traits
Validated 35S::MYB44-lincRNA interactions
MYB44
Group1: 1902 (0.01)
Group2: 655 (0.11)
Group3: 620 (0.01)
Group4: 380 (0.11)
Group5: 234 (0.02)
Group6: 197 (0.12)
Group7: 145 (0.32)
Group8: 128 (0.09)
Group9: 79 (0.11)
Group10: 42 (0.07)
Group11: 42 (0.14)
Group12: 17 (0.18)
Group13: 26 (0.31)
Group14: 33 (1.0)
Group15: 41 (0.56)
Group16: 21 (0.48)
Group18: 22 (0.18)
Group19: 13 (0.0)
Group20: 16 (0.19)
Group21: 17 (0.65)
Group22: 14 (0.93)
Group23: 11 (0.45)
Group24: 16 (0.31)
Group26: 12 (0.58)
Group27: 12 (0.75)
Group29: 13 (0.54)
Group30: 9 (0.89)
Group31: 9 (0.22)
D
LincRNA associated with GWAS root traits
Validated pifq/35S::PIF4-lincRNA interactions
PIF4
Group1: 1719 (0.01)
Group2: 505 (0.07)
Group3: 380 (0.01)
Group4: 191 (0.05)
Group5: 387 (0.17)
Group6: 228 (0.03)
Group7: 147 (0.36)
Group8: 92 (0.1)
Group9: 107 (0.03)
Group10: 87 (0.36)
Group11: 40 (0.1)
Group12: 45 (0.31)
Group13: 29 (0.41)
Group14: 16 (1.0)
Group15: 31 (0.16)
Group16: 19 (1.0)
Group17: 18 (0.89)
Group18: 15 (0.33)
Group19: 15 (0.27)
Group20: 13 (0.31)
Group21: 12 (0.75)
Group22: 11 (0.82)
Group24: 11 (0.55)
Group28: 9 (0.56)
Group29: 9 (0.7)

Figure 6. Experimental validation and characterization of TF-lincRNA regulatory interactions. (A) Heatmap showing log2 fold change (FC) of lincRNA relative expression levels in transcription factor (TF) overexpressing lines (35S::MYB44, 35S::PIF4 and KAN1-GR) or TF knockout lines (rga28 and pifq (pif1/pif2/pif3/pif4 quadruple mutants)) vs. control wild-type lines in 14-day old roots. Expression values were determined by RT-qPCR. Asterisks indicate statistically significant differences (*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$) in an unpaired two-tailed Student's t-Test (n =3). Solid boxes indicate TF ChIP peaks <2kb from the lincRNA gene. Dashed boxes indicate TF ChIP peaks were identified only in one ChIP-Seq replicate or the lincRNA is not the closest gene to the TF peak. **(B)** Heatmap showing KAN1 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. The first line shows the expression profile of the TF under investigation while the other lines show groups of target genes with similar expression. Next to the group number, the number of target genes is reported and in parenthesis the fraction of lincRNAs per group is indicated. The numbers in squares and circles report the number of GWAS root-related and RT-qPCR experimentally validated lincRNA genes, respectively. **(C)** Heatmap showing MYB44 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. Group information is the same as in panel B. **(D)** Heatmap showing PIF4 ChIP-Seq peak-based target genes in the cluster 19 root expression atlas. Group information is the same as in panel B, but here triangles and circles report the RT-qPCR experimentally validated lincRNA genes for the pifq and 35S::PIF4 lines, respectively.