Published in final edited form as:
Plant J. 2017 April ; 90(1): 133–146. doi:10.1111/tpj.13481.

Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants

Dong Wang[1,†], **Zhipeng Qu**[2,†], **Lan Yang**[1], **Qingzhu Zhang**[1], **Zhi-Hong Liu**[1], **Trung Do**[2], **David L. Adelson**[2], **Zhen-Yu Wang**[3], **Iain Searle**[2,*], and **Jian-Kang Zhu**[1,4,*]

[1]Shanghai Center for Plant Stress Biology, Shanghai Institute for Biological Science, Chinese Academy of Sciences, Shanghai 200032, China

[2]Department of Genetics and Evolution, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, 5005, Australia

[3]Hainan Key laboratory for Sustainable Utilization of Tropical Bioresources, College of Agriculture, Hainan University, Haikou, China

[4]Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA

SUMMARY

Noncoding RNAs have been extensively described in plant and animal transcriptomes by using highthroughput sequencing technology. Of these noncoding RNAs, a growing number of long intergenic noncoding RNAs (lincRNAs) have been described in multicellular organisms, however the origins and functions of many lincRNAs remain to be explored. In many eukaryotic genomes, transposable elements (TEs) are widely distributed and often account for large fractions of plant and animal genomes yet the contribution of TEs to lincRNAs is largely unknown. By using strand-specific RNA-sequencing, we profiled the expression patterns of lincRNAs in Arabidopsis, rice and maize, and identified 47 611 and 398 TE-associated lincRNAs (TE-lincRNAs), respectively. TE-lincRNAs were more often derived from retrotransposons than DNA transposons and as retrotransposon copy number in both rice and maize genomes so did TE-lincRNAs. We validated the expression of these TE-lincRNAs by strand-specific RT-PCR and also demonstrated tissue-specific transcription and stress-induced TE-lincRNAs either after salt, abscisic acid (ABA) or cold treatments. For Arabidopsis TE-lincRNA11195, mutants had reduced sensitivity to ABA as demonstrated by longer roots and higher shoot biomass when compared to wild-type. Finally, by altering the chromatin state in the Arabidopsis chromatin remodelling mutant *ddm1*, unique lincRNAs including TE-lincRNAs were generated from the preceding untranscribed regions and interestingly inherited in a wild-type background in subsequent generations. Our findings not only

[*]For correspondence: jkzhu@sibs.ac.cn or iain.searle@adelaide.edu.au.
[†]These authors contributed equally to this work.

AUTHORS' CONTRIBUTIONS

Experiments were designed by DW, ZQ, QZ, DA, IS and JK. DW, ZQ, LY, QZ, ZL, TD, and ZW conducted experiments and all authors analysed the data. DW, ZQ, IS wrote the manuscript and all authors edited the manuscript. All authors read and approved the final manuscript.

SUPPORTING INFORMATION

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

demonstrate that TE-associated lincRNAs play important roles in plant abiotic stress responses but lincRNAs and TE-lincRNAs might act as an adaptive reservoir in eukaryotes.

Keywords

transposable element; long intergenic noncoding RNAs; transposable elements-associated lincRNAs; abiotic stress; noncoding RNAs

INTRODUCTION

Noncoding RNAs (ncRNA) have been extensively described in plant and animal transcriptomes by using high-throughput sequencing technology. Besides canonical ncRNAs that include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear and small nucleolar RNAs, many regulatory ncRNAs have been characterized (Cech and Steitz, 2014). Small regulatory RNAs, including microRNAs and small interfering RNAs (siRNAs), have been demonstrated to play important roles in the regulation of eukaryotic gene expression through either transcriptional or post-transcriptional mechanisms (Bologna and Voinnet, 2014). These small RNAs are produced by cleavage of folded doublestranded RNA (dsRNA) derived from long noncoding RNA (lncRNA). A growing number of lncRNAs have been shown to function in gene regulation without being processed into small RNAs. In animals, to balance the copy number of X chromosomes between male and female cells, the lncRNA Xist recruits Polycomb group proteins to cause lysine 27 trimethylation in histone H3 (H3K27me3) to silence one X chromosome in females (Plath *et al.*, 2003). In plants, thousands of lncRNAs generated by the DNA-dependent RNA polymerase V are involved in RNA-directed DNA methylation and transcriptional gene silencing (Wierzbicki *et al.*, 2008).

Recently, one type of lncRNA, long intergenic noncoding RNAs (lincRNAs), have been identified by tiling arrays or RNA-sequencing in several plant species (Liu *et al.*, 2012; Li *et al.*, 2014; Zhang *et al.*, 2014). LincRNAs are defined as ncRNA longer than 200 nt that do not overlap with either protein-coding or other non-lincRNA types of genes (Ulitsky and Bartel, 2013). Some of them are known to play fundamental biological roles in plant development and physiology (Ariel *et al.*, 2014, 2015; Zhang *et al.*, 2014), such as *INDUCED BY PHOSHATE STARVATION1* (*IPS1*), that can inhibit the function of miR319 through target mimicry during inorganic phosphate starvation response (Franco-Zorrilla *et al.*, 2007). Although the functions of lincRNAs are beginning to be studied in plants, their origin still remains obscure.

Transposable elements (TEs) have been found to be widely distributed in many eukaryotic genomes, and constitute a large fraction of plant and animal genomes. In humans, more than two-thirds of mature lncRNAs contain an exon of at least partial TE origin (Kapusta *et al.*, 2013), and they are believed to contribute contemporary sequence elements to conserved lncRNAs in animals (Hezroni *et al.*, 2015). The contribution of TEs to lincRNA in plants is still unknown. In this report, we explored the contribution of TEs to lincRNAs in three plants species, with significantly different genomic TE diversity. Proportions of lincRNAs harbouring TEs are significantly higher in maize and rice than in Arabidopsis, which is consistent with the number of TEs in these genomes. We name these lincRNAs containing

TEs, TE-associated lincRNAs (TE-lincRNAs), and show that some of them are expressed in a tissue-specific pattern. Of particular interest was the observation that the expression pattern of some TE-lincRNAs varied in response to different stress conditions. Furthermore, *Arabidopsis thaliana* seedlings deficient in TE-lincRNA11195, were more resistant to abscisic acid (ABA) treatment when compared to wild-type (WT), indicating that this lincRNA was involved in the abiotic stress response. Importantly, unique lincRNAs, including TE-lincRNAs, were transcribed in seedlings with DDM1 (decrease in DNA methylation 1) loss of function, and these lincRNAs were inherited in subsequent generations in a WT background, suggesting that these unique lincRNAs produced by changing the chromatin status can be inherited.

RESULTS

Genome-wide identification of TE-lincRNAs in three plant species

To systematically identify TE-lincRNAs, we performed strand-specific RNA-sequencing from 2-week-old seedlings of three plant species. Because of the low expression levels of retrotransposon-derived lncRNAs reported in human and mouse (Fort *et al.*, 2014), we produced high-depth transcriptomes, of approximately 66 million, 173 million, and 256 million pair-end Illumina reads from three biological replicates of *Arabidopsis thaliana*, rice (*Oryza sativa* subsp. *japonica*) and maize (*Zea mays* subsp. *mays* var. B73), respectively (Table S1). We constructed a comprehensive pipeline to identify TE-lincRNAs, consisting of three key steps (Figure 1). First, transcripts from the three species were reconstructed from their RNA-seq datasets using Cufflinks (Trapnell *et al.*, 2010) after mapping reads to the corresponding reference genomes with TopHat2 (Kim *et al.*, 2013). Second, only transcripts greater than 200 nt and not overlapping annotated genes were kept, and we then removed potentially peptide/protein-coding transcripts by sequence similarity search against SWISSPROT and filtered out transcripts with open reading frames (ORFs) larger than 100 amino acids (aa) inside or 50 aa at end(s). After filtering, 205, 1229 and 773 transcripts remained, corresponding to lincRNAs in Arabidopsis, rice and maize, respectively (Table 1). Third, lincRNAs partially overlapping TE loci but not completely located inside TEs were classified as TE-lincRNAs. In the end, we identified 47, 611 and 398 TE-lincRNAs from Arabidopsis, rice and maize, respectively (Tables 1 and S2). The significantly larger proportion of TE-lincRNAs in rice and maize when compared to Arabidopsis is correlated with the increased number of TEs (Table 1). We then determined the genomic distribution of TE-lincRNAs in all three genomes and found that TE-lincRNAs were distributed on all nuclear chromosomes, but were not strongly correlated with the distributions of TEs along the chromosome (Figure S1).

We then compared some general characteristics of TE-lincRNAs and lincRNAs without TEs (designated as non-TE-lincRNAs). Both TE-lincRNAs and non-TE-lincRNAs share similar length distributions in all three species (Figure S2), while the average lengths of TE-lincRNAs are significantly longer than non-TE-lincRNAs in Arabidopsis (829 nt compared to 773 nt, *P*-value = 0.01347, Wilcoxon rank sum test), rice (1125 nt compared to 834 nt, *P*-value < 2.2e-16, Wilcoxon rank sum test) and maize (1343 nt compared to 753 nt, *P*-value < 2.2e-16, Wilcoxon rank sum test). The majority of lincRNAs, including TE-lincRNAs and non-TE-lincRNAs, are single-exon transcripts in all three species examined (Figure S3). There is no significant difference between the average exon numbers of TE-lincRNAs and non-TE-lincRNAs in Arabidopsis and rice, while only slightly significant lower average exon numbers for TE-lincRNAs in maize (1.6 compared to 1.5, *P*-value = 0.2507 in Arabidopsis; 1.6 compared to 1.7, *P*-value = 0.1432 in rice; 1.3 compared to 1.4, *P*-value = 0.007197 in maize; Wilcoxon rank sum test). These results indicated that TEs may have contributed to the extension of transcribed length of lincRNAs but not to splicing complexity in rice and maize. In addition, we scored the potential of RNA motifs embedded in TE-lincRNAs and non-TE lincRNAs by utilizing the Rfam database, and most lincRNAs, either TE-lincRNAs or non-TE-lincRNAs, have none or only one RNA motif (Figure S4 and Table S3). There was no significant difference with respect to the number of embedded RNA motifs between TE-lincRNAs and non-TE-lincRNAs (*P*-value = 0.8368 in Arabidopsis; *P*-value = 0.5387 in rice; *P*-value = 0.8285 in maize; Wilcoxon rank sum test). Next we determined if positional bias of lincRNAs with respect to corresponding neighboring protein- coding genes occurs in the three genomes. Both TE-lincRNAs and non-TE-lincRNAs showed biased distributions at 5′ or 3′ end 5 kilobase (kb) flanking regions of protein- coding genes (Figure S5). We also checked the correlation of expression profiles of TE-lincRNAs and non- TE-lincRNAs with their 10 closest genes at the 5′ end or 3′ end using public RNA-seq datasets (Figure S6a) (Filichkin *et al.*, 2010; Di *et al.*, 2014). We observed the significant high positive or negative expression correlation between some TE-lincRNAs or non-TE-lncRNAs with their neighbor genes, but not for all lincRNAs. Then we reconstructed the protein-coding and non-protein-coding RNA co-expression networks based on the expression profiles across these RNA-seq datasets, and 16 320 genes as well as 77 lincRNAs (including 12 TE-lincRNAs) were reconstructed into 21 co-expression sub-networks (Table S4). TE-lincRNAs were identified with high expression correlation with multiple protein-coding genes in co-expression sub-networks showing stress response (Figure S6b, c).

Examination of TE contributions to lincRNAs

Plant TEs are primarily of two types: class I (retroelement) transposing through an RNA intermediate (copy and paste mechanism) and class II (DNA element) using a DNA intermediate (cut and paste mechanism) to transpose (Bennetzen and Wang, 2014). These two types of TEs can be further classified into many families based on their sequence similarity (Wicker *et al.*, 2007), and each family of TEs has its own functional properties and evolutionary history. Therefore, we were interested in studying the contribution of different TE families to lincRNAs. In Arabidopsis, more than 40% of TE-lincRNAs (22 out of 47) contained 28 RC/Helitron TEs (Figure 2a and Table S5). In rice, the majority of TE-lincRNAs (228, 247 and 197 out of 611) harboured 424 miniature inverted-repeat transposable elements (MITEs), 438 unclassified transposons and 335 unclassified retrotransposons, respectively. While in maize, Gypsy (413 Gypsys contributed to 230 lincRNAs), LTR (172 LTRs contributed to 116 lincRNAs) and Copia (166 Copias contributed to 113 lincRNAs) retrotransposons were the three major contributors in inbred line B73 (Figure 2a and Table S5). Because the copy number of different TEs in these genomes differs, enrichment analysis was carried out to examine the significance of contributions of different TEs to lincRNAs. The short interspersed elements (SINEs) in

Arabidopsis made the most significant contribution to lincRNAs (Figure 2b). While unclassified transposons and unclassified retrotransposons, Ty3-gypsy and centromere-specific retrotransposons contributed remarkably to rice lincRNAs (Figure 2b). In maize, LTR and long interspersed elements (LINE) were most significant enriched TE families in lincRNAs (Figure 2b). Aside from TE families over-represented in their contribution to lincRNAs, there were some TE families under-represented in their contribution to lincRNAs. Nine TE families of Arabidopsis were excluded from lincRNA transcripts, including LINE/L1 with a copy number greater than 1000 (Table S5). In rice, no lincRNAs harboured segments of Mariner while DNA/hAT-Ac with a copy number of approximately 3200 was one of five TE families that did not contribute lincRNAs in maize (Table S5). These results suggest that different TE families have different contributions to lincRNAs in varied plant species. Compared to the two crops used in this study, more TE families tended to be depleted from lincRNAs in 2-week-old Arabidopsis seedlings.

With respect to the number of TEs contributing to individual TE-lincRNA, we found that the largest number of lincRNAs contain only a single TE, while some of TE-lincRNAs can contain up to 18 or 43 TEs, in rice and maize respectively (Figure S7(a)). Length of TE-lincRNAs contributed by more than one TEs are longer than those contributed by one TE (Figure S7b). When considering the coverage of lincRNAs contributed to by TEs, we found that many lincRNAs had a high percentage of TE content, especially in maize (Figure S7c). Conversely, most TEs that contributed to lincRNAs are fully inside TE-lincRNAs (Figure S7d). We also found that the percentage of TE-lincRNAs in identified lincRNAs was much higher than the percentage of genes contributed to by TEs (Figure S8a). Specifically, TE coverage in TE-lincRNAs was significantly higher than TE coverage in protein-coding genes in maize (mean coverage as 54.3% to 16.7%, *P*-value < $2.2e^{-16}$, Wilcoxon rank sum test), but not in Arabidopsis (mean coverage as 33.4–36.9%, *P*-value = 0.2894, Wilcoxon rank sum test) and rice (mean coverage as 35.6–38.5%, *P*-value = 0.1355, Wilcoxon rank sum test) (Figure S8b). In addition, we also checked the coordinates of TEs with respect to host lincRNAs. Most TEs were completely nested inside the lincRNAs, as we have shown in Figure S7 (d), while most of the remaining TEs were located within 500-bp flanking regions of lincRNAs (Figure S7e).

We also analysed the conservation of TE-lincRNA between Arabidopsis and rice according to the protocol described in the methods, but because the number of whole-genome pairwise alignments between maize and other species was small (only four), this conservation analysis was not performed for maize. The overall conservation levels of different genomic features were similar in both Arabidopsis and rice as measured by the phyloP score. As expected, the most conserved element was genes, the least conserved was TEs and TE-lincRNAs were more conserved than TEs (Figure 3). TE-lincRNAs and non-TE-lincRNAs had a similar level of conservation (Figure 3). This was broadly consistent with the idea that TEs embedded in lncRNAs were functionally or structurally constrained by evolution (Kapusta *et al.*, 2013).

Transcript profiling of TE-lincRNAs in Arabidopsis

Next we validated expression of TE-lincRNAs in seedlings of Arabidopsis, maize and rice by strand-specific reverse transcription (RT)-PCR. We selected 11 candidates for further expression analysis in Arabidopsis of which we validated expression for all of them (Figure 4a). All TE-lincRNAs tested were amplified from only one strand as expected from our directional RNA-seq data. Moreover, they were amplified from cDNA primed with oligodT indicating that the TE-lincRNAs were polyadenylated. Similarly in rice and maize, all three TE-lincRNAs from each species were confirmed to be expressed as all were amplified from strand-specific cDNA or oligodT primed cDNA (Figure 4b, c). To measure TE-lincRNAs transcript levels in different Arabidopsis tissues we performed digital PCR on a Fluidigm Biomark HD system. All TE-lincRNAs exhibited varied expression patterns in different tissues. For example, lincRNA18980 was found to be highly expressed in roots but almost not expressed in flowers, and TE-lincRNA3688, TE-lincRNA11344 and TE-lincRNA15772 showed very low levels of expression in root, flower and silique tissues (Figure 5(a)). In addition, transcript profiles of 205 Arabidopsis lincRNAs under different stress treatments were analysed using public RNA-seq data (Filichkin *et al.*, 2010). Compared with normal growth conditions, the expression patterns of many lincRNAs, including TE-lincRNAs, were altered in five stress conditions (Figure 5b). This observation was consistent with early studies that lncRNAs exhibit tissue-specific or spatiotemporal patterns (Cabili *et al.*, 2011; Derrien *et al.*, 2012; Goff *et al.*, 2015). Because many lncRNAs have been shown to be involved in gene regulation *in cis*, we further checked the correlation of expression between selected TE-lincRNAs and their neighbouring genes. For TE-lincRNA15772 and TE-lincRNA19433, there was no correlation between the expression of these TE-lincRNAs and their flanking genes; however, the transcript level of TE-lincRNA11344 was negatively correlated with expression of its neighbouring gene DPBF2 (Spearman's correlation, $r = -0.9036145$, P-value = 0.00208, Figure 5c), suggesting that TE-lincRNA11344 may function by downregulating the adjacent gene.

Mutations in Arabidopsis TE-lincRNA11195 cause resistance to abscisic acid

To investigate functional roles of TE-lincRNAs during stress conditions, we identified homozygous T-DNA insertion mutants in a number of TE-lincRNAs and screened the mutants under standard laboratory conditions and during ABA treatment (Alonso *et al.*, 2003). Strikingly, two independent T-DNA insertion alleles of TE-lincRNA11195 containing a LTR exhibited ABA resistant phenotypes (Figure 6). T-DNA insertions in both mutants caused TElincRNA11195 transcript to be undetectable (Figure 6a), and we designated these two lines as *11195-1* and *-2*. In the absence of exogenous ABA, *11195-1* and *11195-2* seedlings had similar growth when compared to WT (Figure 6b). However, after moving to media supplemented with 20 μM ABA, remarkably enhanced resistance was observed in the mutants compared with WT (Figure 6(b)). Both mutants had significantly increased primary root elongation when compared to WT under 20 μM ABA treatment (Figure 6(c) top panel, Two-sample independent *t*-test, $P < 0.05$), and a weak but non-significant enhancement in the fresh weight of aerial tissues (Figure 6c bottom panel, Two-sample independent *t*-test, $P > 0.05$). In addition, we tested whether TE-lincRNA11195 plays a role in seed germination. Mutants of *lincRNA11195* also showed insensitive to exogenous ABA at the stage of seed germination (Figure S9a), and were substantially insensitive to ABA in post-germination seedling development (Figure S9b, c, Two-sample independent *t*-test, $P < 0.01$).

To investigate the transcription regulation of TElincRNA11195, we measured TE-lincRNA11195 RNA abundance in WT and ABA insensitive mutants by RT-PCR. Abundance of lincRNA11195 increased more than two-fold under ABA treatment at 12 h in Col-0 (Figure S10). Mutant seedlings of the ABA receptors PYR1/PYL1/PYL4 inhibited the transcription of TE-lincRNA11195 (Figure S10). Together these findings clearly demonstrate that TE-lincRNA11195 is ABA responsive. To further explore the regulation and functional role of TE-lincRNA11195 during abiotic stress responses, TE-lincRNA11195 abundance was monitored in several stress conditions. Besides ABA treatment, salt and cold treatments changed the abundance of TE-lincRNA11195, but did not affect the adjacent gene expression (Figure S11a). Next we studied the role of TE-lincRNA11195 under salt treatment at both germination and post-germination seedling development. Seed germination and greening rates of seedlings were significantly higher in *lincRNA11195* mutants than WT (Figure S11b, c), suggesting that TE-lincRNA11195 is also involved in response to salt. Together these results indicated that TE-lincRNA11195 is involved in abiotic stress responses in plants.

In order to identify potential gene targets of TElincRNA11195, we performed RNA-seq on wild-type and *lincRNA11195* mutants under normal and ABA treated conditions. We used a Generalized Linear Model (GLM) to identify differential expressed genes in *lincRNA11195* amongst wild-type and ABA treatments and identified 8 and 10 genes that were significantly up- or down-regulated (Benjamini-Hochberg method adjusted *P*-value < 0.05), respectively (Figure 7a and Table S6). Gene Ontology (GO) enrichment analysis of the 100 most significantly differentially expressed genes indicated that genes involved in ‘response to salicylic acid stimulus’ are most significantly over-represented (Figure 7b). The genomic distribution of these 100 most significantly differentially expressed genes showed that they are distributed across all chromosomes (Figure 7c). Further molecular analysis, for example RIP to detect RNA–protein interactions, will be required to elucidate the function of TE-lincRNA11195.

TEs insertions are known to modify transcriptional responses in plants (Naito *et al.*, 2009; Ito *et al.*, 2011), and we evaluated the contribution of the LTR to TElincRNA11195 transcription under ABA treatment. Expression of TE-linc11195 in transgenic plants without the LTR was slightly higher than in plants with the LTR under control conditions; but the expression of TE-linc11195 harbouring the LTR had a greater increase than in plants without the LTR under ABA treatment (Figure S12), suggesting that TE enhances the extent of TE-linc11195 ABA response at the transcriptional level. We then investigated the expression of TE-linc11195 in the close relative *Arabidopsis lyrata* and *Capsella rubella*, as the DNA sequence of TE-linc11195 is present in both species. Transcript of TE-linc11195 could be detected in two-week-old seedlings of *A. thaliana* and *A. lyrata* (Figure S13), indicating that TE-linc11195 may function specifically at this stage in the Arabidopsis genus. Next we performed a pairwise sequence alignment of TE-linc11195 between *A. thaliana* and *A. lyrata* and this indicated the majority of the sequence is conserved between these two species

(Figure S14). We also performed a comparison of the secondary structures of TE-linc11195 in the two species and demonstrated they were largely conserved (Figure S14).

Characterization of unique TE-lincRNAs generated in loss of *ddm1* mutant plants

Chromatin changes can be triggered by fluctuations in the ambient environment (Talbert and Henikoff, 2014), and unique lncRNAs responsive to abiotic or biotic stress have also been characterized in plants (Di *et al.*, 2014; Zhu *et al.*, 2014). Therefore we were interested in the correlation between lincRNA expression and chromatin status in *ddm1.* Mutated chromatin-remodeling factor DDM1 alters the distribution of DNase I hypersensitive (DH) sites that are closely associated with RNA Polymerase II binding sites (Zhang *et al.*, 2012; Wang and Timmis, 2013). We generated a transcriptome dataset including approximately 70 million paired-end reads from 2-week-old *ddm1* seedlings (Table S1). As we expected, unique transcripts were detected from intergenic regions of plants defective for DDM1, and TE-lincRNAs as well as non-TE-lincRNAs were detected (Figure 8a and Table 1). There was a similar percentage of TE-lincRNAs found in the *ddm1* lincRNA repertoire (102 out of 446) compared to WT, nonetheless the total number of TE-lincRNAs and non-TE-lincRNAs was increased in *ddm1* Col (Table 1). 387 *ddm1* specific lincRNAs were found, and 192 of them were found to be covered by DH sites by checking their position and 1-kb flanking regions (Zhang *et al.*, 2012; Wang and Timmis, 2013), indicating that unique lincRNAs can be generated once nuclear chromatin state changes. Subsequently, the inheritance of these unique lincRNAs was studied in *ddm1* heterozygous seedlings produced by crossing *ddm1* homozygous plants with WT and by intercrossing the F1 to produce F2 plants (Figure 8b). Interestingly, transcripts of these lincRNAs could be detected in heterozygous F1 seedlings (Figure 8c) and strikingly in the subsequent F2 generation expression was independent of the *DDM1* genotype (Figure 8c and Table S7). Of interest, these *ddm1* specific lincRNAs were not expressed in some of *ddm1* homozygous seedlings, indicating that the inheritance of lincRNA is non-Mendelian (Table S7).

DISCUSSION

The importance of lncRNAs involved in biological processes has been extensively described in many plant species including crops thanks to advances in DNA sequencing technology (Li *et al.*, 2014; Zhang *et al.*, 2014; Wang *et al.*, 2015, 2016), but comprehensive understanding of their biological function and origin still remain elusive. Although TEs are proposed to be a major contributor to vertebrate lncRNAs (Kelley and Rinn, 2012; Kapusta *et al.*, 2013), their contribution to plant lncRNAs remains unclear. In this study, we mainly focused on the contribution of TEs to lincRNA in three plant species, one model dicotyledonous species (*Arabidopsis thaliana*) and two important monocotyledonous crops (rice and maize). In total, 47, 611 and 398 TE-lincRNAs were identified in Arabidopsis, rice and maize respectively by using high-quality RNA-seq data with high-depth stranded RNA-sequencing. In rice and maize, TEs occurred in approximately half of the lincRNAs identified from 2-week-old seedlings. Despite the small proportion of TEs in the *Arabidopsis thaliana* genome, more than 20% of identified lincRNAs included TEs. This demonstrates that TEs make a remarkable contribution to lincRNAs in plants particularly as TEs constitute the majority of DNA in many plant genomes. Furthermore, lincRNAs preferentially harbour TEs compared to protein- coding genes, which is an observation that is consistent with findings in mammals (Kapusta *et al.*, 2013).

While TEs are ubiquitous in lincRNAs from all three examined plants, some TE families are excluded from the lincRNA repertoire (Table S5). Moreover, the relative abundance of TE families within lincRNAs does not simply mirror that of the entire genome. For example, the copy number of SINEs is not high, but their contribution to lincRNA is significant in *Arabidopsis thaliana* (Figure 2b). These results show that contribution to lincRNA does not mean a close correlation with the number of TEs. Also the interspecific variations we observed in the coverage and type of TEs in lincRNAs reflect the abundance and intrinsic properties of certain TEs residing in the genome, and it further suggests that TEs play a role in the divergence of lincRNAs.

LincRNAs are known to exhibit organ-specific expression patterns in Arabidopsis (Liu *et al.*, 2012), and this pattern was also observed in TE-lincRNAs (Figure 5a). Furthermore, varied TE-lincRNAs expression was observed under different stress treatments (Figure 5b), indicating their transcription is responsive to abiotic stress. This hypothesis is supported by the ABA treatment result of TE-linc11195 There is an LTR in the 5′ terminal region of TE-linc11195, and knock out of TE-linc11195 caused an ABA insensitive phenotype for *Arabidopsis thaliana* seedlings comparing with WT (Figure 6 and Figure S9). Moreover, expression of this TE-lincRNA was completely blocked in seedlings mutated ABA receptors genes PYR1/PYL1/PYL4 (Figure S10). In addition, a salt insensitive phenotype was also observed in plants defected in TE-linc11195 at stages of seed germination and post-germination seedling development (Figure S11b, c). These results further indicate that TE-lincRNAs are involved in plants' responses to abiotic stress. Because it has been suggested that TEs provide unique sequence elements to conserved lncRNAs (Hezroni *et al.*, 2015), the contribution of the LTR to TE-linc11195 was also checked under ABA treatment. Interestingly, we found that this LTR could strengthen the extent of ABA response for TE-linc11195 (Figure S12), indicating that TEs play a biological role in the evolution of lincRNAs.

Changes in chromatin state caused the generation of unique lincRNAs in *ddm1* Col (Table 1 and Figure 8). These unique lincRNAs may also play a role in responses to stress, which may contribute to the biotic stress resistance found in *ddm1* Col (Dowen *et al.*, 2012). Our observation is also different from the previous suggestion that TE insertions give rise to functional lncRNAs (Ponting *et al.*, 2009), indicating that both TE-lincRNA and non-TE-lincRNA can simply arise by alteration of chromatin state. This finding provides an attractive hypothesis that chromatin altered by environmental factors can produce unique lincRNAs which may be functional when responding to the environment and can be inherited. Our hypothesis is also consistent with a previous suggestion that lncRNAs have a distinct advantage over proteins as gene regulators because they can be functional immediately upon transcription without needing to be translated into protein outside the nucleus (Johnson and Guigo, 2014). In the light of the many possible regulatory roles of lincRNAs, the environmentally triggered appearance of lincRNAs may diversify biological regulation of the organism and drive an increased rate of evolution. Our observation that TE-lincRNA11195 was transcribed in the genus Arabidopsis but not *Capsella* (Figure S13) might help explain lineage-specific changes in gene networks. As transposable elements are often clade specific, clade specific TE-lincRNAs would be expected to frequently arise. This idea could be tested by RNA-seq analysis to identify lineage-specific TE-lincRNAs from a number lineages combined with CRISPR/Cas genome editing to remove specific lineages of TE-lincRNAs.

CONCLUSION

We have identified 47, 611 and 398 TE-linRNAs in 2-weekold seedlings of *Arabidopsis thaliana*, rice and maize respectively. Different TE families have differing extents of contribution to lincRNAs. More importantly, we found that many TE-lincRNAs are potentially stress-responsive and may contribute to stress response. This was validated by the perturbation of one TE-lincRNA, lincRNA11195, which was found to be involved in the ABA response. Furthermore, unique TE-lincRNAs and non-TE-lincRNAs could be detected in mutants whose nuclear chromatin state had changed, and these unique lincRNAs were inherited. This research has evaluated the contribution of TEs to lincRNAs and demonstrated the important role played by TE-lincRNAs in response to stress.

EXPERIMENTAL PROCEDURES

RNA-seq library preparation and sequencing

Total RNAs were obtained from 2-week-old seedlings of Arabidopsis, rice and maize. The preparation of strand-specific RNA-seq libraries and deep sequencing were performed in the Shanghai Center for Plant Stress Biology (Shanghai, China). These libraries were constructed through applying TruSeq Stranded mRNA (Illumina, San Diego, CA, USA) in accordance with the manufacturer’s instruction. The quality of RNA-seq libraries were assessed by using a Fragment Analyzer (Advanced Analytical, IA, USA), and the resulting libraries were sequenced on an Illumina HiSeq 2500 instrument producing pair-end reads of 100 or 125 nucleotides. For *ddm1* Col, RNA was extracted from 2-week-old seedlings, and shipped to Beijing Genomics Institue (Shenzhen, China) for sequencing.

TE-lincRNA identification pipeline

Adaptors and low quality sequences were filtered with trimgalore (v0.3.3, –stringency 6). Then clean reads were aligned to reference genomes (TAIR10 for Arabidopsis, TIGR release 7 for rice and AGPv2 for maize) using Tophat2 with following parameters: -N 5 –read-edit-dist 5 (v2.0.14) (Kim *et al.*, 2013). Mapped reads from three biological replicates for Arabidopsis and rice were merged and then assembled with Cufflinks respectively (v2.2.1) (Trapnell *et al.*, 2010). For maize, mapped reads were assembled with Cufflinks firstly and then merged with Cuffmerge, due to the large number of mapped reads (Trapnell *et al.*, 2010). Annotated protein-coding genes or transcripts with protein encoding potential were filtered with following three steps: (i) remove short transcripts (shorter than 200 bp), intronic transcripts and transcripts overlapping with protein-coding genes (at least 1 bp overlapping); (ii) BLASTX against SWISSPROT protein sequence database (Camacho *et al.*, 2009); and (iii) remove transcripts with ORFs longer than 100 aa inside or 50 aa at end(s). The remaining transcripts were categorized as lincRNAs. Finally, genomic coordinates of lincRNAs were further checked with respect to TEs in Arabidopsis, rice and maize respectively. LincRNAs overlapping with but not fully inside TE (s) were characterised as TE-lincRNAs.

Sequence conservation analysis

Whole-genome level pairwise alignments of Arabidopsis with 23 other plants and rice with 27 other plants were downloaded from Ensemble Plants (Kersey *et al.*, 2012). Multiple alignments were obtained by merging pairwise alignments with multiz (Blanchette *et al.*, 2004). Phylogenetic models were estimated by applying phyloFit on four-fold degenerate (4d) sites according to the manual (Hubisz *et al.*, 2011). Based on the multiple alignments and estimated phylogenetic models, conservation scores for different genomic features, including protein-coding genes, TEs, TE-lincRNAs, non-TE-lincRNAs and intergenic intervals (the intergenic intervals were defined as the genomic intervals after removing all protein-coding genes and lincRNAs), were calculated by using phyloP with following parameters: –features –method SCORE – mode CONACC (Hubisz *et al.*, 2011).

RNA motif detection

Rfam 12.0 is a collection of noncoding RNA families by multiple sequence alignments, consensus secondary structures and covariance models (CMs) (Nawrocki *et al.*, 2015). The program ‘cmscan’ from the infernal package was used to search the lincRNA sequence against CM-format motifs in Rfam 12.0 with following parameter: −E $1e^{-1}$ (Nawrocki and Eddy, 2013). If multiple RNA motifs were identified from overlapped regions the one with the smallest E-value was selected.

Expression correlation analysis and co-expression network reconstruction

Variance-stabilizing transformation of raw counts for lincRNAs and protein-coding genes across multiple samples from public RNA-seq datasets (SRA00903 and GSE49325) were used to calculate pairwise correlation between transcripts. Pearson’s correlation was calculated between lincRNA and the 10 closest protein-coding genes. WGCNA was used to reconstruct Arabidopsis lincRNA and reference gene co-expression networks (Langfelder and Horvath, 2008).

Statistical analysis and data visualization

Statistical analysis and data visualization of characterises of TElinRNAs and non-TE-lincRNAs were performed with R and R packages (Lawrence *et al.*, 2009; R Development Core Team, 2010, Yin *et al.*, 2012).

Plant materials, stress treatment and PCR assay

Seeds of *C. rubella* and *A. thaliana* T-DNA insertion mutants including *11195-1* (CS843057), *11195-2* (CS834193) and *ddm1-10* (SALK_093009) were obtained from Arabidopsis Biological Resource Center (ABRC). ABA insensitive mutant used in this study is *pyr1*/*pyl1*/*pyl4* (Park *et al.*, 2009). For generating transgenic lincRNA11195 plants with or without the LTR, DNA fragments containing 1.5 kb upstream of lincRNA11195 and the fulllength or lacking LTR region lincRNA sequence plus a 200-bp downstream sequence with attB sites were amplified from Col- 0 genomic DNA, and were then cloned into Gateway vector pDONR207 (Invitrogen). Each insert was subsequently introduced into the Gateway pGWB1 vector by LR reaction (Invitrogen). All plasmids were transformed into *Agrobacterium tumefaciens* strain GV3101, and then transformed into *A. thaliana* plants of the mutant backgrounds via the floral dip method. Stress treatment was carried out as described previously (Zeller *et al.*, 2009). Preparation of cDNA and real-time quantitative PCR were performed according to the previous description (Wang *et al.*, 2014). RT-PCR and strand-specific RTPCR were carried out as described previously (Wierzbicki *et al.*, 2008). All experiments were carried out with at least three biological replicates. Details of the primers used in this study are listed in Table S8.

Gene differential expression analysis of *TE-lincRNA11195* mutant RNA-seq

Fourteen-day-old wild-type and *11195-2* seedlings were grown on half-strength MS medium then treated with either 0 or 100 μM ABA for 12 h, RNA extracted and then Illumina sequencing performed. Adaptor and low quality sequences were trimmed with trim_galore the same as above. Clean reads were aligned to reference genome using STAR_2.5.2a with following parameters: – outFilterMismatchNmax 10 –outFilterMismatchNoverLmax 0.05 –seedSearchStartLmax 30. Gene differential expression analysis was performed using edgeR with GLM method considering two factors: lincRNA11195 mutant and ABA treatment (Robinson *et al.*, 2010).

Sequence pairwise alignment and secondary structure prediction of TE-lincRNA11195 in *A. thaliana* and *A. lyrata*

Homolog of TE-lincRNA11195 in *A. lyrata* was determined using its sequence of *A. thaliana* blastn against *A. lyrata* genomic sequences (https://github.com/PacificBiosciences/DevNet/wiki/Arabidopsis-lyrata) and extended to the equivalent length of TE-linc11195 in *A. thaliana*. Sequence pairwise alignment of TE-lincRNA11195 between *A. thaliana* and *A. lyrata* was performed using ClustalX2 (Larkin *et al.*, 2007). The secondary structures of TE-lincRNA11195 in two species were predicted using RNA-fold with the default setting (Gruber *et al.*, 2008).

Availability of data and materials

The data sets supporting the results of this article are available in NCBI's GEO database repository, and are accessible through GEO accession number GSE76798.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded by the Chinese Academy of Sciences, and National Science Foundation of China (31401077) awarded to DW, by the Australian Research Council (ARC) through a Future Fellowship (FT130100525) awarded to IS and a MOET-VIED PhD scholarship awarded to TD. The authors declare no conflicts of interest.

References

Alonso JM, Stepanova AN, Leisse TJ, et al. Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science. 2003; 301:653–657. [PubMed: 12893945]

Ariel F, Jegu T, Latrasse D, Romero-Barrios N, Christ A, Benhamed M, Crespi M. Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. Mol Cell. 2014; 55:383–396. [PubMed: 25018019]

Ariel F, Romero-Barrios N, Jegu T, Benhamed M, Crespi M. Battles and hijacks: noncoding transcription in plants. Trends Plant Sci. 2015; 20:362–371. [PubMed: 25850611]

Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014; 65:505–530. [PubMed: 24579996]

Blanchette M, Kent WJ, Riemer C, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004; 14:708– 715. [PubMed: 15060014]

Bologna NG, Voinnet O. The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. Annu Rev Plant Biol. 2014; 65:473–503. [PubMed: 24579988]

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]

Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014; 157:77–94. [PubMed: 24679528]

Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]

Di C, Yuan J, Wu Y, et al. Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. Plant J. 2014; 80:848–861. [PubMed: 25256571]

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR. Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci USA. 2012; 109:E2183–E2191. [PubMed: 22733782]

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. 2010; 20:45–58. [PubMed: 19858364]

Fort A, Hashimoto K, Yamada D, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet. 2014; 46:558–566. [PubMed: 24777452]

Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, Garcia JA, Paz-Ares J. Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet. 2007; 39:1033–1037. [PubMed: 17643101]

Goff LA, Groff AF, Sauvageau M, et al. Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA. 2015; 112:6855–6862. [PubMed: 26034286]

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res. 2008; 36:W70–W74. [PubMed: 18424795]

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 2015; 11:1110– 1122. [PubMed: 25959816]

Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011; 12:41–51. [PubMed: 21278375]

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011; 472:115–119. [PubMed: 21399627]

Johnson R, Guigo R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA. 2014; 20:959– 976. [PubMed: 24850885]

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013; 9:e1003470. [PubMed: 23637635]

Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012; 13:R107. [PubMed: 23181609]

Kersey PJ, Staines DM, Lawson D, et al. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res. 2012; 40:D91–D97. [PubMed: 22067447]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14:R36. [PubMed: 23618408]

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. [PubMed: 19114008]

Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947–2948. [PubMed: 17846036]

Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009; 25:1841–1842. [PubMed: 19468054]

Li L, Eichten SR, Shimizu R, et al. Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. 2014; 15:R40. [PubMed: 24576388]

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell. 2012; 24:4333– 4345. [PubMed: 23136377]

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009; 461:1130–1134. [PubMed: 19847266]

Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013; 29:2933–2935. [PubMed: 24008419]

Nawrocki EP, Burge SW, Bateman A, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015; 43:D130–D137. [PubMed: 25392425]

Park SY, Fung P, Nishimura N, et al. Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. Science. 2009; 324:1068–1071. [PubMed: 19407142]

Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. Role of histone H3 lysine 27 methylation in X inactivation. Science. 2003; 300:131–135. [PubMed: 12649488]

Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136:629–641. [PubMed: 19239885]

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140. [PubMed: 19910308]

Talbert PB, Henikoff S. Environmental responses mediated by histone variants. Trends Cell Biol. 2014; 24:642–650. [PubMed: 25150594]

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511– 515. [PubMed: 20436464]

Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013; 154:26–46. [PubMed: 23827673]

Wang D, Timmis JN. Cytoplasmic organelle DNA preferentially inserts into open chromatin. Genome Biol Evol. 2013; 5:1060–1064. [PubMed: 23661564]

Wang D, Qu Z, Adelson DL, Zhu JK, Timmis JN. Transcription of nuclear organellar DNA in a model plant system. Genome Biol Evol. 2014; 6:1327–1334. [PubMed: 24868015]

Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, Wang P, Liu N, Lindsey K, Zhang X. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). New Phytol. 2015; 207:1181–1197. [PubMed: 25919642]

Wang X, Ai G, Zhang C, Cui L, Wang J, Li H, Zhang J, Ye Z. Expression and diversification analysis reveals transposable elements play important roles in the origin of Lycopersicon-specific lncRNAs in tomato. New Phytol. 2016; 209:1442–1455. [PubMed: 26494192]

Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007; 8:973–982. [PubMed: 17984973]

Wierzbicki AT, Haag JR, Pikaard CS. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell. 2008; 135:635–648. [PubMed: 19013275]

Yin T, Cook D, Lawrence M. ggbio: an R package for extending the grammar of graphics for genomic data. Genome Biol. 2012; 13:R77. [PubMed: 22937822]

Zeller G, Henz SR, Widmer CK, Sachsenberg T, Ratsch G, Weigel D, Laubinger S. Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. Plant J. 2009; 58:1068–1082. [PubMed: 19222804]

Zhang W, Zhang T, Wu Y, Jiang J. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. Plant Cell. 2012; 24:2719–2731. [PubMed: 22773751]

Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, Qu LH, Shu WS, Chen YQ. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol. 2014; 15:512. [PubMed: 25517485]

Zhu QH, Stephen S, Taylor J, Helliwell CA, Wang MB. Long noncoding RNAs responsive to Fusarium oxysporum infection in Arabidopsis thaliana. New Phytol. 2014; 201:574–584. [PubMed: 24117540]

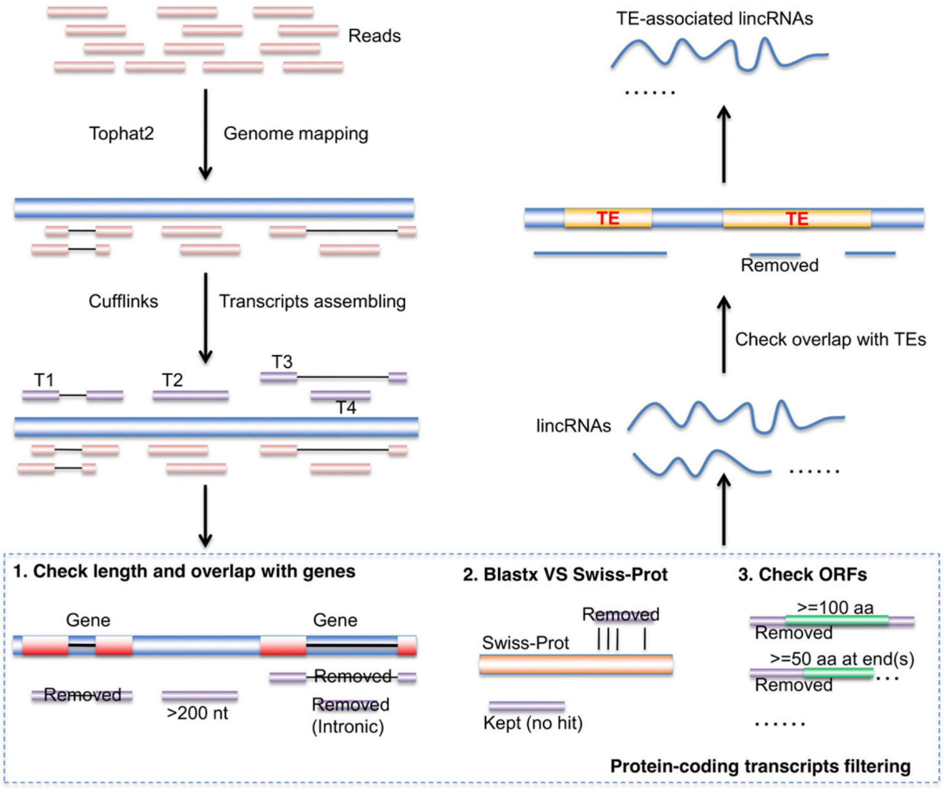

Figure 1. Identification of TE-associated lincRNAs from RNA-seq data

Quality-checked short reads were mapped to the reference genome using TopHat2 and Cufflinks was then used to assemble the mapped reads into longer transcripts. To filter out protein-coding transcripts and canonical noncoding RNA the following three steps were undertaken. First, transcripts shorter than 200 nt were removed and the remaining were tested for overlap with annotated genes. Those transcripts that either overlapped with annotated genes by at least one base pair or that were located in the intronic regions of genes were removed. Second, transcripts with high similarity to known protein motifs were identified by BLASTX searches against the SWISS_PROT database and then removed. The last step involved inspecting the transcript ORFs and removing transcripts with ORFs longer than 100 amino acids (aa) inside the transcript or longer that 50 aa at transcript end(s). These remaining transcripts were classified as candidate lincRNAs. TE-associated lincRNAs were identified as those that overlapped with transposable element (TE) loci but did not fully reside within a TE. [Colour figure can be viewed at wileyonlinelibrary.com].

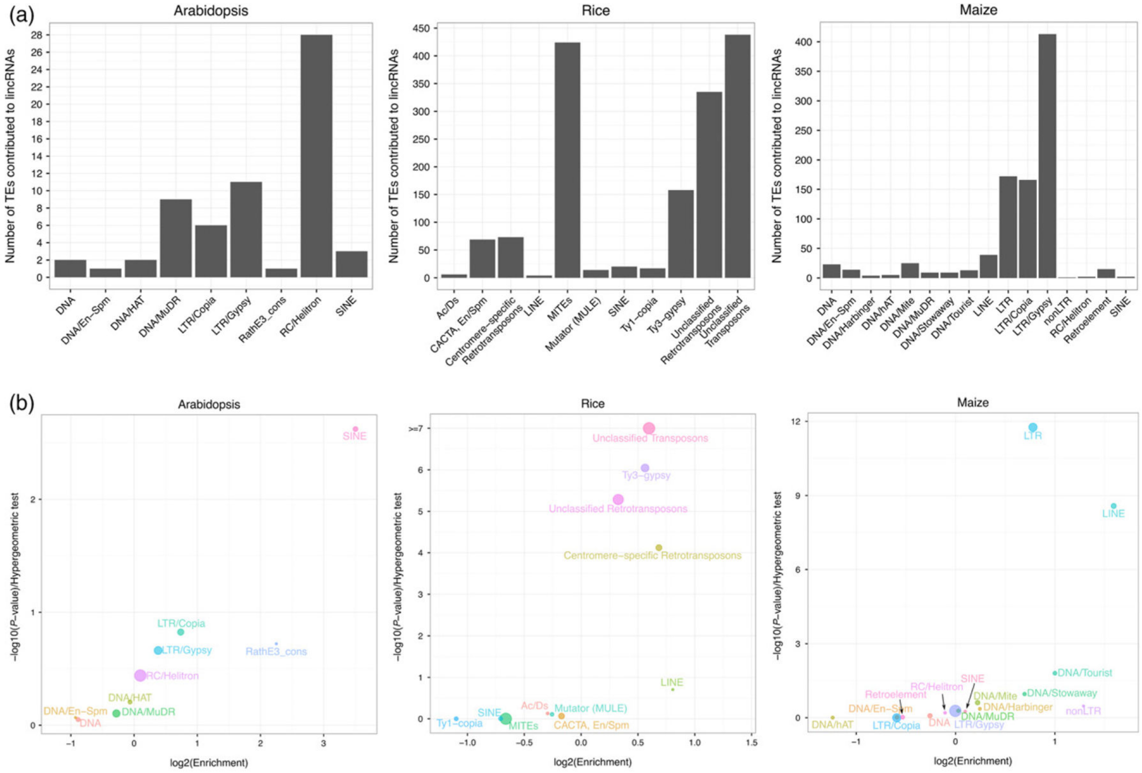

Figure 2. Occurrence and enrichment of different TE families in lincRNAs from Arabidopsis, rice and maize

(a) Bar charts showing the number of TEs from different families contributing to lincRNAs. (b) Bubble charts describing the over-representation of different TE families contributing to TE-associated lincRNAs. X axis represents the fold of enrichment of different TE families contributing to lincRNAs. Y axis represents statistical significance of the over-representation of different TE families contributing to lincRNAs (*P*-value, hypergeometric test). Sizes of bubbles indicate proportions of TEs in each TE family with respect to total number of TEs contributing to lincRNAs. [Colour figure can be viewed at wileyonlinelibrary.com].

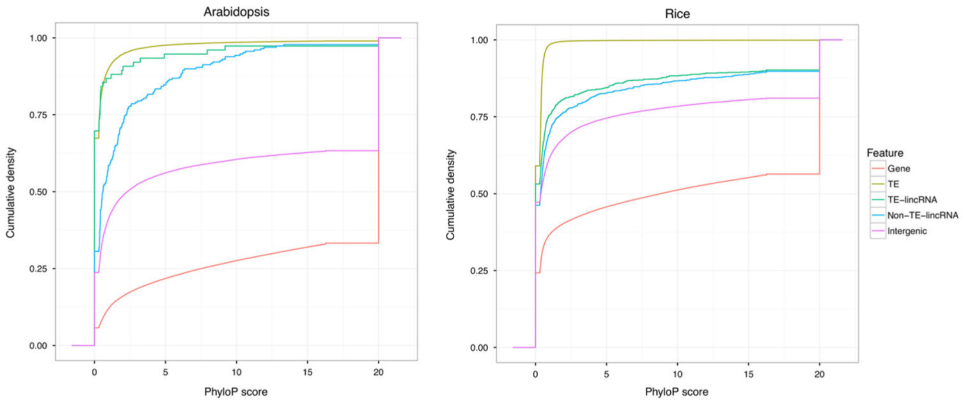

Figure 3. Level of conservation of TE-lincRNAs, non-TE-lincRNAs, genes, TEs and intergenic regions in Arabidopsis and rice

The cumulative distributions of phyloP scores derived from 24-way (Arabidopsis) and 28-way (rice) whole-genome alignments are presented. [Colour figure can be viewed at wileyonlinelibrary.com].

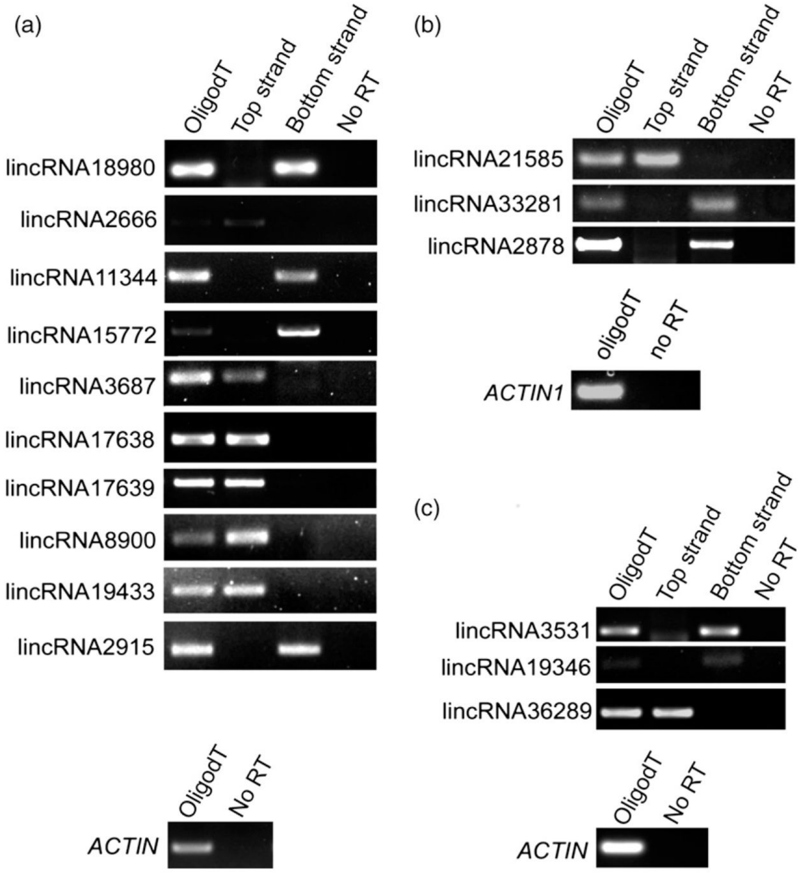

Figure 4. Detection of TE-lincRNAs in three species

(a–c) Strand-specific RT-PCR analysis was carried out on selected TE-lincRNA transcripts in either (a) *A. thaliana*; (b) *Oryza sativa* subs. *japonica*; or (c) *Zea mays* B73. Either oligodT, top strand or bottom strand-specific primers were used in the reverse transcription cDNA synthesis. Control RT-PCRs using either *ACTIN* or *ACTIN1* primes are shown below each panel.

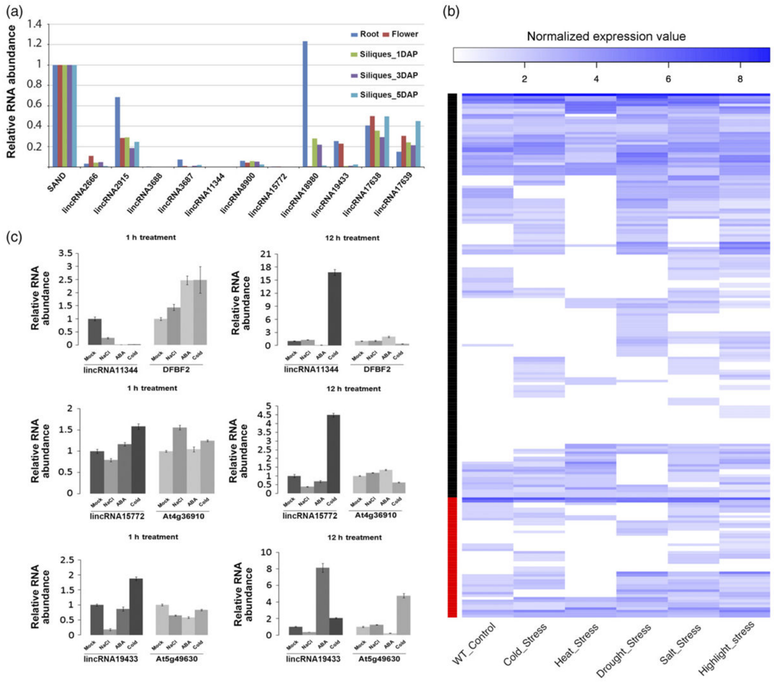

Figure 5. Expression pattern of TE-lincRNAs

(a) Expression of TE-lincRNAs in different Arabidopsis tissues. cDNA abundance was normalized using the SAND transcript.

(b) Heatmap showing expression profiles of Arabidopsis lincRNAs under different stress conditions. Expression value was normalized by variance-stabilizing transformation of raw counts. Black sidebar: 154 non-TE-lincRNAs; red bar: 47 TE-lincRNAs.

(c) Expression of selected TE-associated lincRNAs and neighbouring genes under different conditions. *ACTIN7* was used as a control in the qRT-PCR experiments of this study. [Colour figure can be viewed at wileyonlinelibrary.com].

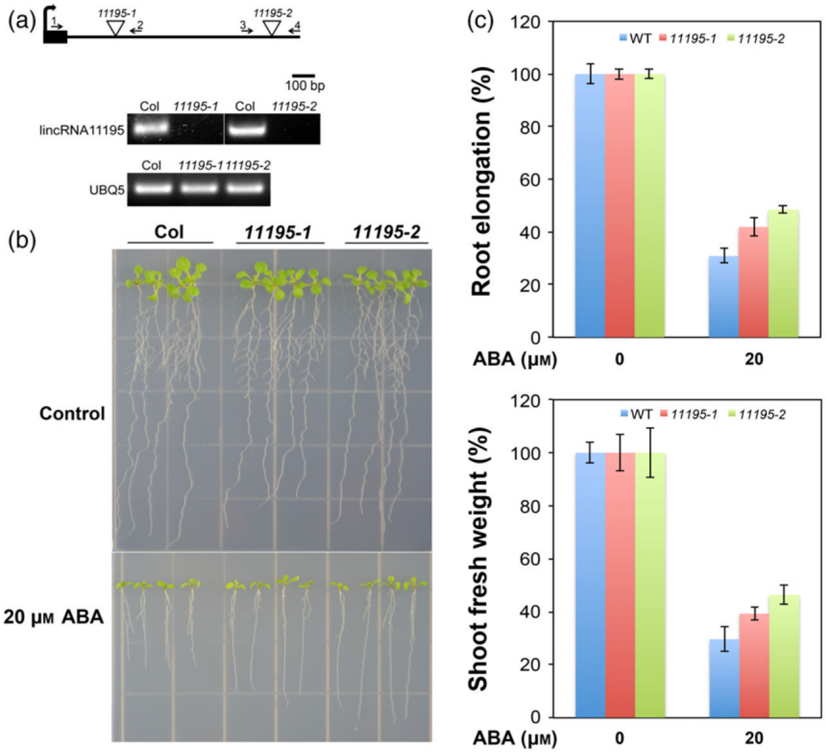

Figure 6. Arabidopsis TE-associated lincRNA11195 mediates ABA responses

(a) Expression analysis of the TE-associated lincRNA11195 in wild-type (Col-0) and two T-DNA insertion mutant alleles (lincRNA *11195-1* and *11195-2*). Bold right curved arrow shows the direction of transcription of lincRNA11195. The two primer pairs shown (1 and 2 for *11195-1* and 3 and 4 for *11195-2*) were used to amplify the TElincRNA11195 transcript. UBQ5 was used as a positive control.

(b) *TE-linc11195* mutants are insensitive to ABA.

(c) Root length and fresh shoot weight of seedlings shown in (b). Both graphs are presented as the percentage relative to growth on control half-strength MS medium. ABA assays were performed by stratifying seeds at 4°C for 3 days, followed by growth of seedlings for 5 days on half MS media, then seedlings were transferred onto half-strength MS medium supplemented with or without 20 μM ABA, and grown for an additional 8 days. Error bars stand for standard deviation ($n = 20$). [Colour figure can be viewed at wileyonlinelibrary.com].

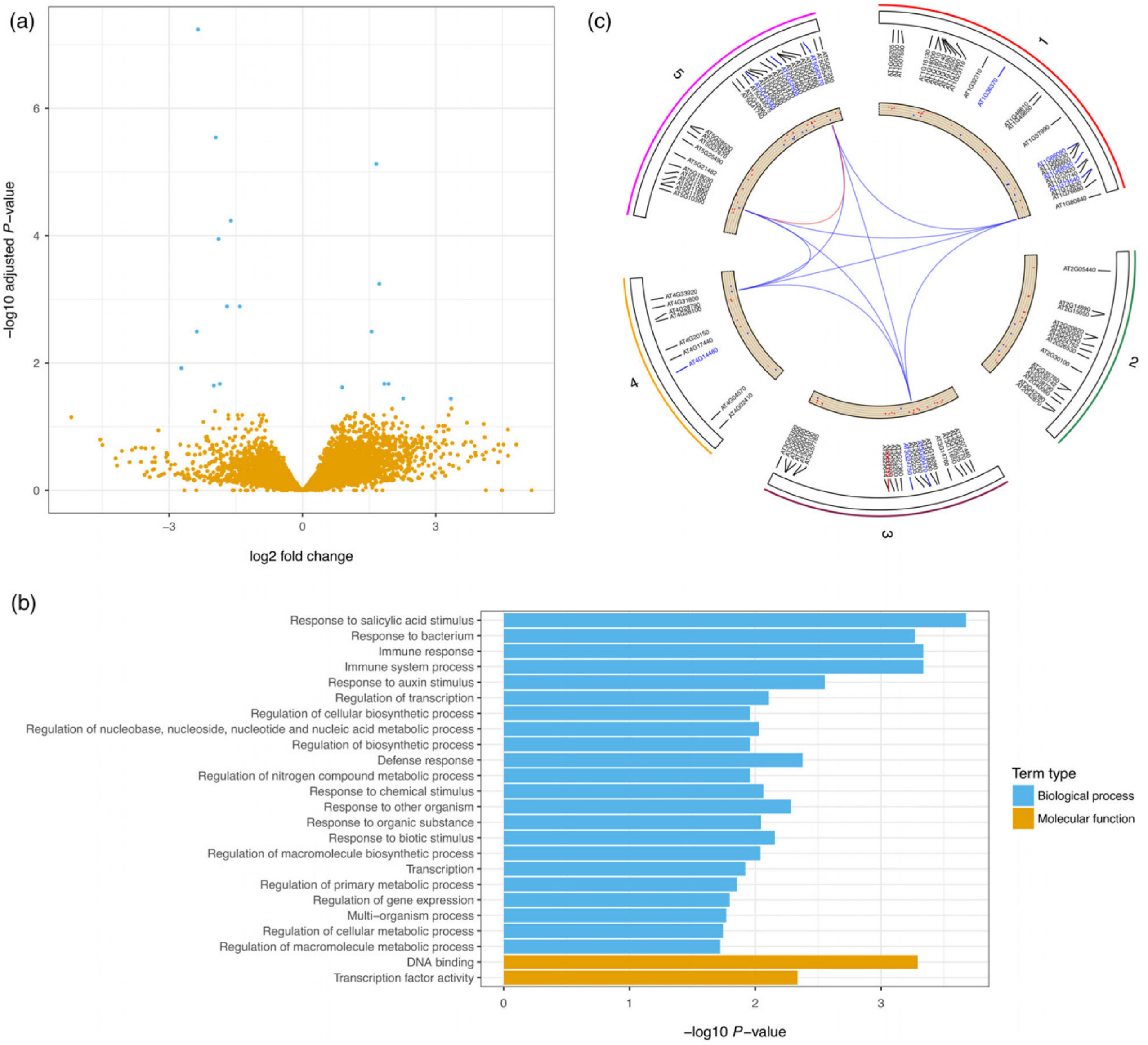

Figure 7. Gene differential expression analysis of *TE-linc11195* mutant using RNA-seq

(a) Volcano plot showing log2 fold changes versus statistical significances of genes. Blue dots represent statistically significant differentially expressed genes (Benjamini–Hochberg method adjusted *P*-value < 0.05).

(b) GO enrichment analysis of 100 most significantly differentially expressed genes.

(c) Genomic distribution of 100 most significantly differentially expressed genes. Gene labels with blue colour are top 10 most significantly expressed genes. Scatter plot inside inner track represents log2-fold changes of genes, therefore, red and blue dots represent up- and down- regulated genes respectively. Links inside circle plot represent five genes associated with most significant over-represented GO term 'response to salicylic acid stimulus', blue and red lines represent between- and in- chromosome connections respectively. [Colour figure can be viewed at wileyonlinelibrary.com].

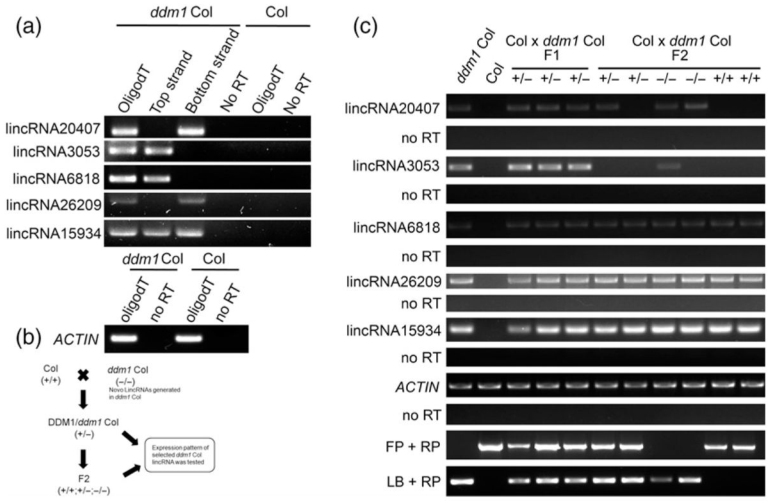

Figure 8. Characterization of unique lincRNAs generated by loss of DDM1

(a) Strand-specific RT-PCR analysis was performed on selected lincRNAs only present in the *ddm1* mutant, three TE-lincRNAs: lincRNA20407, lincRNA3053 and lincRNA6818; two non-TE-lincRNAs: lincRNA26209 and lincRNA159.

(b, c) Expression pattern of *ddm1* dependent lincRNAs in subsequent generations. The − or + symbol indicates the presence or absence of the mutant or wildtype DDM1 allele, respectively. Actin was used as a positive control. FP, RP and LB are primers used to genotype the plants. Primers LB and RP indicate the presence of the *ddm1* T-DNA and primers FP and RP indicate the presence of wild-type allele.

Table 1

Summary of lincRNAs identified in this study

Species	Number of total lincRNAs	Number of TE-associated lincRNAs	Proportion of transposable elements in genome (%)	Proportion of TE-associated lincRNAs in total lincRNAs (%)
A. thaliana	205	47	14	22.9
O. sativa subsp. *japonica*	1229	611	35	49.7
Z. mays B73	773	398	76	51.5
A. thaliana (*ddm1* mutant)	446	102	14	22.9