Short title: **lncRNA variation and silencing in *A. thaliana***

Population-level annotation of lncRNA transcription in *Arabidopsis* reveals extensive variation associated with TE-like silencing

Aleksandra E. Kornienko[1,*], Viktoria Nizhynska[1], Almudena Molla Morales[1], Rahul Pisupati[1], Magnus Nordborg[1,*]

[1] Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter, Dr. Bohr-Gasse 3, 1030, Vienna, Austria

*corresponding authors

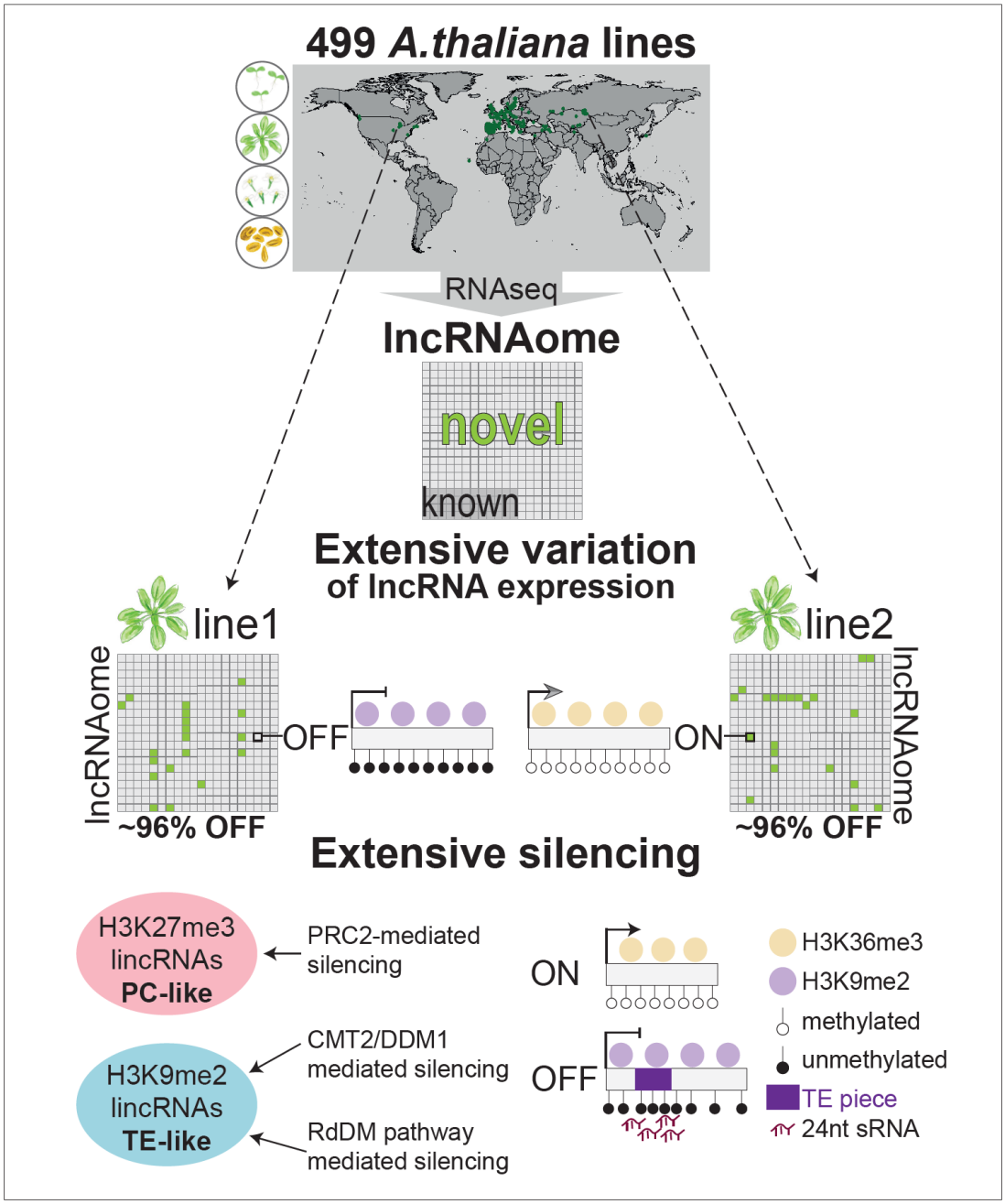
- **Fig. 1. Annotation.** Mapping lncRNA transcription in hundreds of accessions and several tissues reveals thousands of novel lncRNAs.
- **Fig. 2. Expression variation.** lncRNAs display extensive natural expression variability and appear to be largely silent.
- **Fig. 3. Epigenetic patterns.** Epigenetic patterns of lncRNAs in *A. thaliana* indicate ubiquitous silencing.
- **Fig. 4. Epigenetic variation**. LncRNAs display increased epigenetic variation that explains expression variation of many lncRNAs.
- **Fig. 5. TE content in TAIR10.** Many lincRNAs contain pieces of TEs that affect their silencing and variation.
- **Fig. 6. Copy number in TAIR10.** Copy number of lincRNAs affects their epigenetic patterns and variability.
- **Fig. 7. Silencing.** lincRNAs are silenced by PC-like and TE-like mechanisms.
- **Fig. 8. Silencing.** TE pieces appear to attract silencing to lincRNA loci.

Abstract

Long non-coding RNAs (lncRNAs) are under-studied and under-annotated in plants. In mammals, lncRNA loci are nearly as ubiquitous as protein-coding genes, and their expression has been shown to be highly variable between individuals of the same species. Using *A. thaliana* as a model, we aimed to understand the true scope of lncRNA transcription across plants from different regions and study its natural variation. Using RNA-seq data spanning hundreds of natural lines and several developmental stages to create a more comprehensive annotation of lncRNAs, we found over 10,000 new loci — three times as many as in the current public annotation. While lncRNA transcription is ubiquitous in the genome, most loci appear to be actively silenced and their expression is extremely variable between natural lines. This high expression variability is largely caused by the high variability of repressive chromatin levels at lncRNA loci. This was particularly common for intergenic lncRNAs, where pieces of transposable elements (TEs) present in 50% of the loci are associated with increased silencing and variation, and such lincRNAs tend to be targeted by TE silencing machinery. We create the most comprehensive *A. thaliana* lncRNA annotation to date and improve our understanding of plant lncRNA genome biology, raising fundamental questions about what causes transcription and what causes silencing across the genome.

Keywords: *Arabidopsis thaliana*, long non-coding RNAs, gene expression, lncRNA annotation, epigenetics, natural variation, transposon silencing

Graphical abstract



Introduction

Long non-coding RNAs (lncRNAs) are a relatively new and still enigmatic class of genes that are increasingly recognized as important gene regulators participating in nearly every biological process (Statello et al., 2021). There are more lncRNA than protein-coding genes in the human genome (Volders et al., 2019) and they are apparently abundant in the genomes of all eukaryotes (Mattick and Rinn, 2015; Kapusta and Feschotte, 2014). In human and mouse, lncRNAs have been shown to be involved in various diseases (Wapinski and Chang, 2011; Batista and Chang, 2013), and medical applications have been proposed (Wahlestedt, 2013). Although many lncRNAs have proven functions, the vast majority have not been studied or proven functional (Leone and Santoro, 2016), and many knock-outs of seemingly functional candidates showed no phenotype (Sauvageau et al., 2013), leading to continuous debate about the functionality and importance of lncRNAs as a class (Mattick et al., 2023). Evolutionary studies of lncRNAs revealed low sequence conservation and highly divergent expression when compared to protein-coding genes (Necsulea and Kaessmann, 2014; Nelson et al., 2017), yet some signs of conservation and selection have also been found (Johnsson et al., 2014; Mattick et al., 2023). Several studies have looked at how lncRNAs differ between closely related species such as rat and mouse (Kutter et al., 2012), human and chimp (Necsulea and Kaessmann, 2014) or different plant species (Nelson et al., 2017; Zhu et al., 2022), but few have looked at differences within one species (Melé et al., 2015). Recently it was shown that lncRNAs display salient interindividual expression variation in human (Kornienko et al., 2016) and mouse (Andergassen et al., 2017), much higher than that of protein-coding genes, but the meaning, causes and consequences of this high variability are unknown.

Arabidopsis thaliana has higher natural genetic variability than humans (1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium, 2016) and represents an interesting and convenient model for studying lncRNAs in plants. This matters because most research on lncRNAs has been performed in human and mouse (Rinn and Chang, 2020), and relatively little is known about lncRNAs in plants (Liu et al., 2015; Budak et al., 2020). Several studies have identified and annotated lncRNAs in plant species such as *A. thaliana* (Liu et al., 2012), wheat (Xin et al., 2011), maize (Li et al., 2014) and strawberry (Kang and Liu, 2015), but, although several databases have been created, the number and comprehensiveness of plant lncRNA annotations are poorer than that of human and mouse (Xie et

al., 2014; Xuan et al., 2015; Paytuví Gallart et al., 2016; Zhu et al., 2022). Nonetheless, it is clear that lncRNAs do regulate genes in plants (Liu et al., 2015; Whittaker and Dean, 2017; Chen et al., 2023) and that lncRNA expression is particularly responsive to stress and environmental factors (Wang et al., 2017; Budak et al., 2020). Understanding the real scope of lncRNA transcription in plants could identify novel candidates for further functional studies and shed light on the genome biology of lncRNAs in plants and beyond.

While many lncRNAs have been shown to participate in epigenetic silencing or activation of protein-coding genes (Statello et al., 2021), there is much less research on the epigenetic regulation of lncRNAs themselves. In *A. thaliana,* the epigenetic patterns of some functional lncRNAs have been thoroughly studied (Whittaker and Dean, 2017), but little is known about the epigenetics of lncRNAs on a genome-wide level. While high epigenetic variation was reported between natural accessions (Kawakatsu et al., 2016), it is not clear how it affects lncRNAs.

LncRNAs are known to sometimes originate from transposable elements (TEs) (Zhu et al., 2022; Kapusta et al., 2013), yet what implications this has for their expression, epigenetics and variation is not well known. Similarly, while aberrant lncRNA copy number has been connected to disease and other phenotypes (Xu et al., 2020; Athie et al., 2020), general information about lncRNA copy number and its consequences is missing, in particular in plants.

In this study, we aimed to study the extent and natural variability of lncRNA transcription in the model plant *A. thaliana*. We annotated lncRNAs using data from 499 natural accessions, finding thousands of new loci and generating an extended lncRNA annotation. We find that lncRNAs show high expression and epigenetic variability between natural accessions and are generally silenced in any given accession. Epigenetic variability explains expression variation of many lncRNAs. Intergenic lincRNAs show particularly high variability and can be divided into PC-like and TE-like loci that show differences in epigenetic patterns, copy number and—most importantly—the presence of pieces of TE sequences. We found that short pieces of TEs are prevalent in intergenic lncRNAs, likely attracting TE-like silencing to their loci. We provide new insights into the biology of lncRNAs in plants, identify a major role of TE-likeness in lncRNA silencing, and provide an extensive annotation and data resource for the *A. thaliana* community.

Results

Transcriptome annotation from hundreds of natural lines reveals thousands of novel lncRNAs

To investigate the extent of lncRNA transcription in *A. thaliana*, we used newly generated and publicly available (Kawakatsu et al., 2016; Cortijo et al., 2019) PolyA+ stranded RNA-seq data spanning 5 different tissues/developmental stages (seedling, 9-leaf rosette, leaves from 14-leaf rosette, flower and pollen) and 499 natural accessions (Fig. 1A, see Suppl. Table S1 for accession overview, Table S2 for RNA-seq samples overview, and Table S3 for RNA-seq mapping statistics). To create a cumulative transcriptome annotation, we mapped the RNA-seq data from all samples onto the TAIR10 genome, assembled transcriptomes from each accession/tissue separately (Suppl. Table S4) and then used a series of merging and filtering steps creating one cumulative annotation, which we then classified into several gene classes (Fig. 1B, Methods, Suppl. Fig. S1). We used Araport11 and TAIR10 gene annotations (Cheng et al., 2017; TAIR10 annotation) to guide the classification of transcripts corresponding to protein-coding genes (PC genes), pseudogenes, TE genes and TE fragments, rRNAs and tRNA, and used an additional protein-coding potential filtering step to identify a set of lncRNAs (Suppl. Fig. S1 and S2A). Our transcriptome annotation performed well in assembling known PC genes and known lncRNAs (Suppl. Fig. S2B).

In total, we identified 23,676 protein-coding and 11,295 lncRNA loci, i.e., almost one third (29%) of the cumulative transcriptome annotation consisted of lncRNA loci (Fig. 1C). The resulting annotation is highly enriched in lncRNAs (Suppl. Fig. S2C), containing 10,315 novel loci when compared to the public Araport11 annotation (Fig. 1C) and 7,774 novel loci when compared to the recent large-scale lncRNA identification in *A. thaliana* Col-0 accession (Kindgren et al., 2020). Our annotation extends the lncRNA portion of the reference genome from 2.2% to 10.7 %, or ~13 Mbp in total (Suppl. Fig. S2D). We were also able to detect and annotate many TE genes and TE fragments across accessions/tissues (Fig. 1C), finding spliced isoforms for 579 TE genes previously annotated as single-exon (Suppl. Fig. S2E-H).

Figure 1

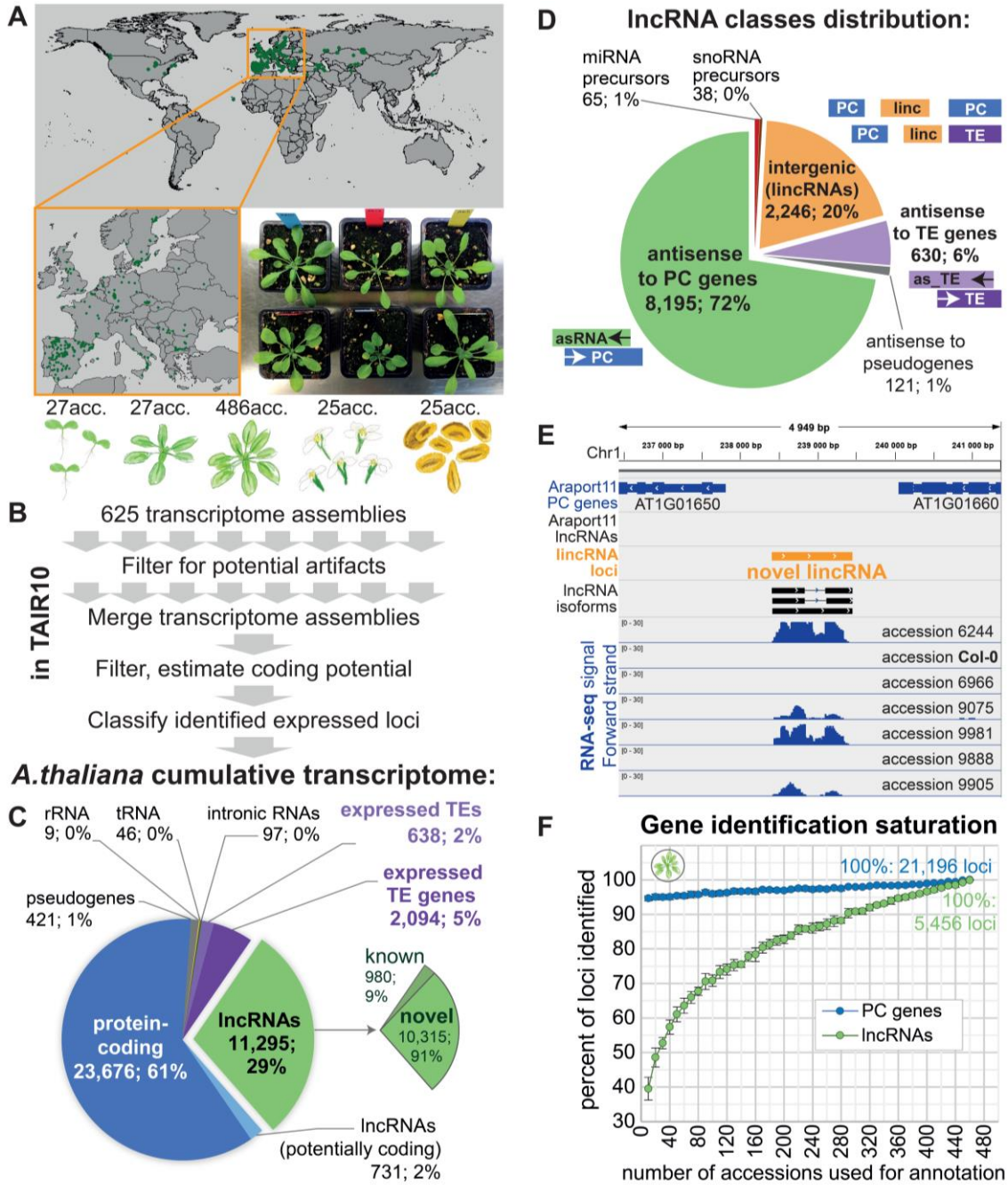


Fig. 1. Mapping lncRNA transcription in hundreds of accessions and several tissues reveals thousands of novel lncRNAs.

A. Origins of the *A. thaliana* accessions used for transcriptome annotation and an example photograph of 6 different accessions in the growth chamber. **B.** Overview of the pipeline used for cumulative transcriptome annotation. Tissues from left to right: seedlings, rosette, flowers, pollen. **C.** The distribution of types of loci in the cumulative annotation. **D.** The distribution of lncRNA positional classes. **E.** An example of a novel intergenic lncRNA on chromosome 1. Expression in 7 different *A. thaliana* accessions is shown. **F.** The number of lncRNA and PC loci identified as a function of the number of accessions used, relative to the number identified using 460 accessions. Random subsampling of accessions was performed in 8 replicates and the error bars indicate the standard deviation across replicates.

We classified lncRNAs based on their genomic position (Fig. 1D). The largest group (8,195, or 72%) were antisense (AS) lncRNAs that overlapped annotated PC genes in the antisense direction (Suppl. Fig. S3). We found that 8,083 Araport11 PC genes have an antisense RNA partner, over five times more than in the Araport11 reference annotation. Previous studies have reported ubiquitous, unstable antisense transcription (Yuan et al., 2015; Li et al., 2013) as well as the activation of antisense transcription upon stress (Xu et al., 2021), however, our data contained exclusively PolyA+ RNA-seq in normal conditions, so we can conclude that relatively stable polyadenylated antisense transcripts can be produced over almost a third of PC genes in *A. thaliana* across different accessions and tissues.

The second largest class with 2,246 loci (20%) were intergenic lncRNAs (lincRNAs) (Fig.1D-E). The third largest class (630, or 6%) consisted of lncRNAs that were antisense to TE genes (AS-to-TE lncRNAs). The other 3 classes constituted <3% of all lncRNA loci and we will ignore them in what follows.

As expected, the genomic distribution of AS lncRNAs and AS-to-TE lncRNAs mirrored the annotation of PC and TE genes respectively, with the former enriched in chromosome arms, and the latter near centromeres. LincRNAs were also enriched near centromeres but could be found genome-wide (Suppl. Fig. S4).

Analyzing more accessions and tissues reveals more lncRNA loci

We hypothesized that a major reason we discovered so many new lncRNA loci was that our annotation was based on hundreds of accessions, while most previous studies used only the reference accession Col-0 (Cheng et al., 2017; Yuan et al., 2015, 2016). If lncRNAs are very variably expressed between individuals, as has previously been found in humans (Kornienko et al., 2016), data from a single accession would uncover only the subset expressed in this particular accession. To test this, we subsampled the unified rosette RNA-seq dataset from the 1001 Genome project (Kawakatsu et al., 2016) and ran the annotation pipeline many times (Methods). This saturation analysis showed that we could increase the number of annotated lncRNA loci 2.5-fold by increasing the number of accessions from 10 to 460 (Fig. 1F). Unlike PC genes, the number of lncRNAs strongly depended on the sample size and showed no sign of saturating at 460 accessions (Suppl. Fig. S5A, B).

To confirm that the observed increase was not simply due to increased sequencing coverage, we compared these results with very high-coverage RNA-seq data from Col-0 only (Cortijo et al., 2019). Applying subsampling to these data, we found a much slower increase that saturated early and could not possibly explain the results in Fig. 1F (Suppl. Fig. S5C).

It is well-known that lncRNAs can be specific to tissue and developmental stage (Cabili et al., 2011), so it is likely that our use of multiple tissues helped identify more loci. Our final annotation, which was based on seedlings, rosettes, flowers, and pollen from multiple accessions (Fig. 1B) revealed 11,265 lncRNA loci, while the 460-accession rosette analysis above gave us only 5,456 lncRNA loci (Fig. 1F). To understand the effect of adding different tissues better, we performed another saturation analysis where we varied both the number of tissues and accessions (Suppl. Fig. S6). While the number of loci always increased with the number of accessions, the number of tissues used mattered even more. In particular, adding flowers or pollen to the analysis produced a big jump in the number of genes identified. For example, when using 20 accessions and 4 tissues allowed the identification of ~3 times more lncRNAs than when using just seedling data (Suppl. Fig. S6). Adding flowers alone nearly doubled the number of lncRNAs identified.

To summarize, by combining RNA-seq data from hundreds of accessions and four developmental stages, we provide a massively expanded lncRNA annotation for *A. thaliana*, identifying over ten thousand novel loci. Our extended lncRNA annotation is available in the supplement.

High lncRNA expression variability between accessions

We have seen that including more accessions allowed the identification of more lncRNA loci (Fig. 1G) and hypothesized that the reason was that not every lncRNA was expressed in every accession. Indeed, this appears to be the case: our analysis showed that while most PC genes were expressed in nearly all accessions, most lncRNAs, as well as most TE genes and fragments, were expressed (TPM>0.5) in less than 5% of accessions (Fig. 2A). Our analysis of the expression frequency (ON/OFF state) of the four main types of loci showed that while about 50% of PC loci were expressed in every accession, the same was true for no more than 1% of AS lncRNAs, lincRNAs and TE genes (Suppl. Fig. S7).

Figure 2

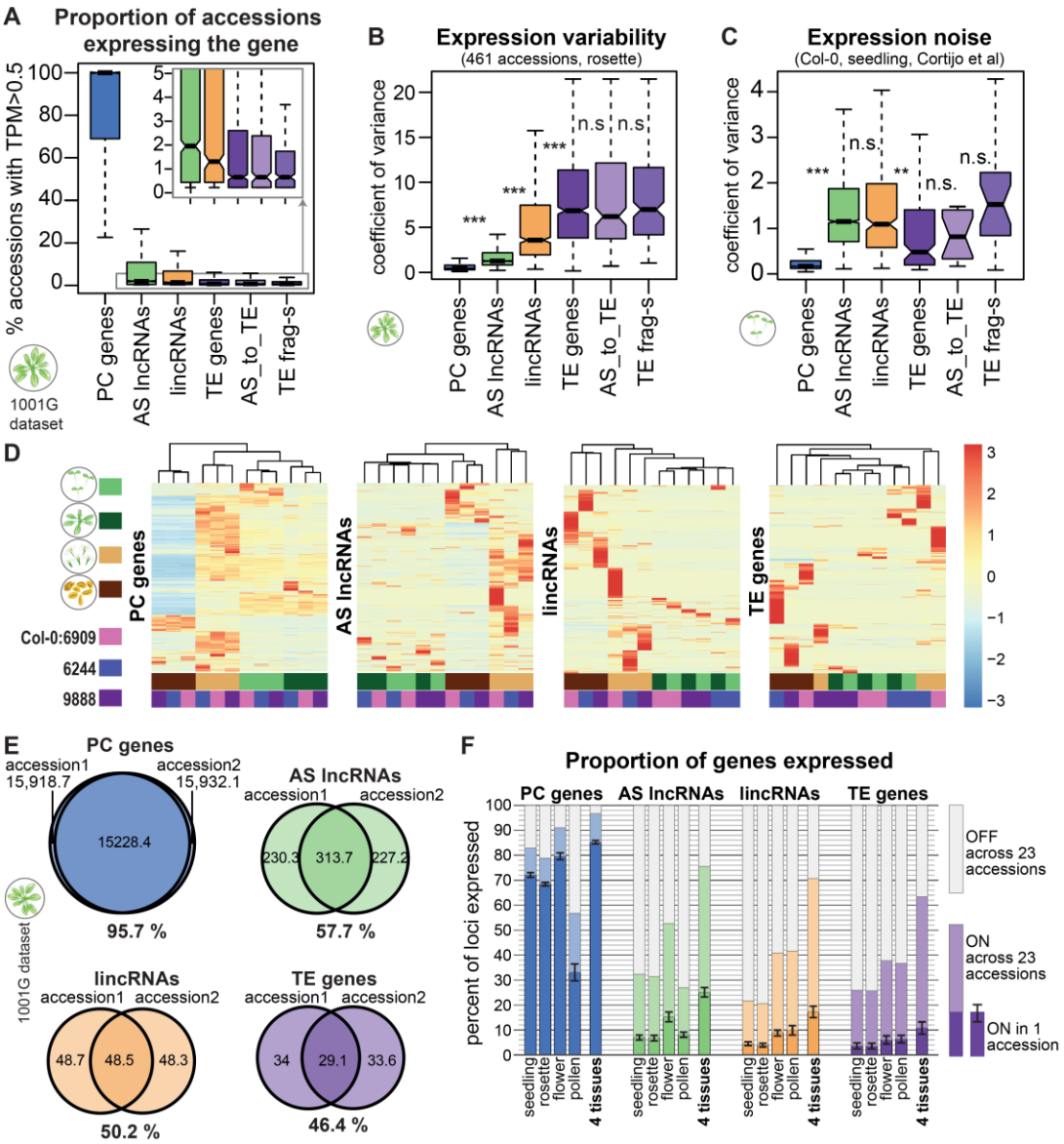


Fig. 2. lncRNAs display extensive natural expression variability and appear to be largely silent.

A. The fraction of accessions in the 1001 Genomes dataset (Kawakatsu et al., 2016) where the gene is expressed (TPM > 0.5). Only genes that are expressed in at least one accession are plotted. **B.** Coefficient of variance of expression in 461 accessions from 1001 Genomes dataset (Kawakatsu et al., 2016). Only genes with TPM > 1 in at least one accession are plotted. **C.** Expression noise calculated from 14 technical replicates of Col-0 seedlings; expression noise value averaged across 12 samples is displayed (Cortijo et al., 2019). Only genes with TPM > 1 in at least one sample are plotted. Boxplots: Outliers are not shown, and p-values were calculated using Mann-Whitney test on equalized sample sizes: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. **D.** Gene expression levels for different types of genes in 4 tissues for the reference accession Col-0 (6909) and 2 randomly picked accessions. Heatmaps built using "pheatmap" in R with scaling by row. Only genes expressed in at least one sample are plotted. Clustering trees for rows not shown. **E.** Average number of genes expressed in an accession and its randomly selected partner accession from the 1001G dataset, and the number of genes expressed (TPM > 0.5) in both accessions. Percentages indicate the overlap between accessions. **F.** The proportion of genes expressed in one accession in seedlings, 9-leaf rosettes, flowers, pollen, or all 4 tissues combined (dark bars). The error bars show standard deviation between 23 accessions. The light part of the bars displays the additional proportion of genes that can be detected as expressed when all 23 accessions are considered.

To quantify the natural expression variability of lncRNAs and other gene types, we calculated the coefficient of variance using rosette data across 461 accessions (Kawakatsu et al., 2016). Similarly to human (Kornienko et al., 2016) and mouse (Andergassen et al., 2017), both AS lncRNAs and lincRNAs are significantly more variable than protein-coding genes (Fig. 2B). In particular, lincRNAs showed expression variability almost at the level of TE genes and fragments. The variability of lncRNAs that were antisense to TE genes was similar to that of TE genes and fragments (Fig. 2B). Analyzing expression variability in other rosette datasets (Suppl. Fig. S8A-B) and other tissues (Suppl. Fig. S8C-E) confirmed these results.

Two factors crucially affect expression variability values and must be controlled for when comparing lncRNAs to PC genes: gene length and absolute expression level. lncRNAs are known to be shorter and have lower expression than PC genes (Cabili et al., 2011) and we confirmed this in our data (Suppl. Fig. S9A,B). Both gene length and absolute expression level are negatively correlated with the coefficient of variance and while this holds true for every gene type, the anticorrelation slopes are different (Suppl. Fig. S9C,D). When we control for expression level, the trend shown in Fig2B is preserved (Suppl. Fig. S8A). When we control for both expression and gene length the trend is preserved for lincRNAs but AS lncRNAs are similar to PC genes (Suppl. Fig. S9E,F), which might be explained by particularly high variability of short PC genes (Cortijo et al., 2019).

As the 14-leaf rosette dataset produced for this paper contained 2-4 repeats for each accession, we could assess the level of intra-accession expression variation. For all classes of genes except AS lncRNAs, the intra-accession expression variation was significantly lower than the inter-accession variation and the difference between the classes of genes mirrored inter-accession variation (Suppl. Fig. S10A,B), which suggests that compared to PC genes, the expression of lincRNAs and TEs is more unstable and prone to be affected by the precise conditions or noise, while much of the AS lncRNA expression variation between accessions might be defined by generally unstable expression. To estimate the actual noise lncRNA expression we analyzed the RNA-seq data from a gene expression study that sampled *A. thaliana* seedlings 12 times over 24 hours with 14 technical replicates per sample (Cortijo et al., 2019). We found that lncRNAs have significantly noisier expression (Fig. 2C) as well as higher circadian expression variability (Suppl. Fig. S10C).

Interestingly, while TE genes showed higher variability between accessions (Fig. 2B), both lincRNAs and AS lncRNAs were noisier than TE genes (Fig. 2C).

To illustrate the extent of lncRNA expression variation we plotted expression across 4 tissues in 3 sample accessions as a heatmap for different types of genes (Fig. 2D). While PC gene expression clusters the samples according to tissue, lincRNA and TE expression clusters according to accession (AS lncRNA is similar to PC gene but is noisier). While pollen samples always cluster separately due to pollen's particular transcriptome (Slotkin et al., 2009), the expression of lincRNAs and TEs was strikingly different between accessions. It was also notable that pollen expressed particularly many lincRNAs and TE genes, while flowers seem to have higher expression for all 4 gene types. In general, Fig. 2D illustrates how few lncRNAs are expressed in each accession and how striking the inter-accession variation is. Two randomly chosen accessions effectively express the same PC genes, whereas they share only about a half of lncRNAs expressed (Fig. 2E). Furthermore, while ~70% of PC genes were expressed in seedlings and rosettes of any given accession, only 7% of AS lncRNAs and 4% of lincRNAs were expressed in these tissues (Fig. 2F). Strikingly, almost twice as many AS lncRNA loci were expressed in flowers, but not in pollen, while twice as many lincRNAs were expressed in both flowers and pollen (Fig. 2F). Like lincRNAs, TE genes show increased expression in flowers and pollen, which has been shown before (Slotkin et al., 2009). Across all 4 tissues in 23 accessions, 96% of PC genes, 76% of AS lncRNAs, 71% of lincRNAs, and 63% of TE genes are expressed (Fig 2.F) thus covering many more lncRNAs, in line with the identification saturation analysis (Fig. 1H) and the increased individual- and tissue-specificity of lncRNAs (Fig. 2B,D).

In summary, lncRNA expression showed high variability between accessions, between tissues, and also between replicates. lincRNAs differ from AS lncRNAs in that they show higher expression variability and increased expression in pollen, but both classes are predominantly silent in any given sample.

The epigenetic landscape of lncRNA loci suggests ubiquitous silencing

To characterize the epigenetic patterns of lncRNAs in *A. thaliana* and investigate their apparently ubiquitous silencing, we performed ChIP-seq and bisulfite-seq in multiple accessions using leaves from 14-leaf rosettes (Methods, Suppl. Fig. S11A, Suppl. Tables S5, S6). For the ChIP

experiments, we chose 2 active marks—H3K4me3 and H3K36me3—and 3 repressive marks associated with different types of silencing: H1, H3K9me2 and H3K27me3 (Suppl. Fig. S11B). H1 has been shown to be involved in silencing TEs, but also antisense transcription (Choi et al., 2020), H3K9me2 is a common heterochromatin mark and is known to silence TEs (Zemach et al., 2013), while H3K27me3 is commonly associated with PRC2 silencing and is mostly found on PC genes (Feng and Jacobsen, 2011). H3K27me3 can also be found on some TEs when the normal silencing machinery is deactivated (Déléris et al., 2021; Zhao et al., 2022).

First, we analyzed the ChIP-seq data focusing on the reference accession Col-0. The gene-body profiles of the different histone modifications are distinct for different types of genes (Fig. 3A). For example, while AS lncRNAs and PC genes showed similar levels of chromatin modifications—which is expected given their overlapping positions—their profiles differed, with PC genes showing a characteristic drop in H1 and H3K9me2 at their transcription start site (TSS) and increase towards the transcription end site (TES), whereas AS lncRNAs showed an even distribution across the gene body (Fig. 3A). LincRNAs and TE genes showed increased heterochromatic marks H3K9me2 and H1 and reduced active marks H3K36me3 and H3K4me3 (Fig. 3A). Calculating normalized and replicate-averaged ChIP-seq coverage over the whole locus (Fig. 3B, Suppl. Fig. S11D) and the promoter region (Suppl. Fig. S11E) confirmed the above observations. Overall, AS lncRNAs were similar to PC genes, while lincRNAs were intermediate between PC genes and TE genes in the heterochromatic marks and the lowest in the active marks (Fig. 3B, Suppl. Fig. S11D, E).

Next, we analyzed the bisulfite sequencing data, quantifying DNA methylation in three different contexts: CG, CHG, and CHH, where H stands for A, C, or T (Methods). CHG and CHH methylation are common in plants and are involved in TE-silencing (Fultz et al., 2015). While PC genes and AS lncRNAs as expected displayed low levels of CG methylation and no CHG or CHH methylation, lincRNAs exhibited a very significant methylation increase in all three contexts (Fig. 3C, Suppl. Fig. S12A, B). Interestingly, the distribution of lincRNA CG-methylation was bimodal (Fig. 3C: right, Suppl. Fig. S12C, D), with some loci looking like PC genes, while others looked like TE genes.

Figure 3

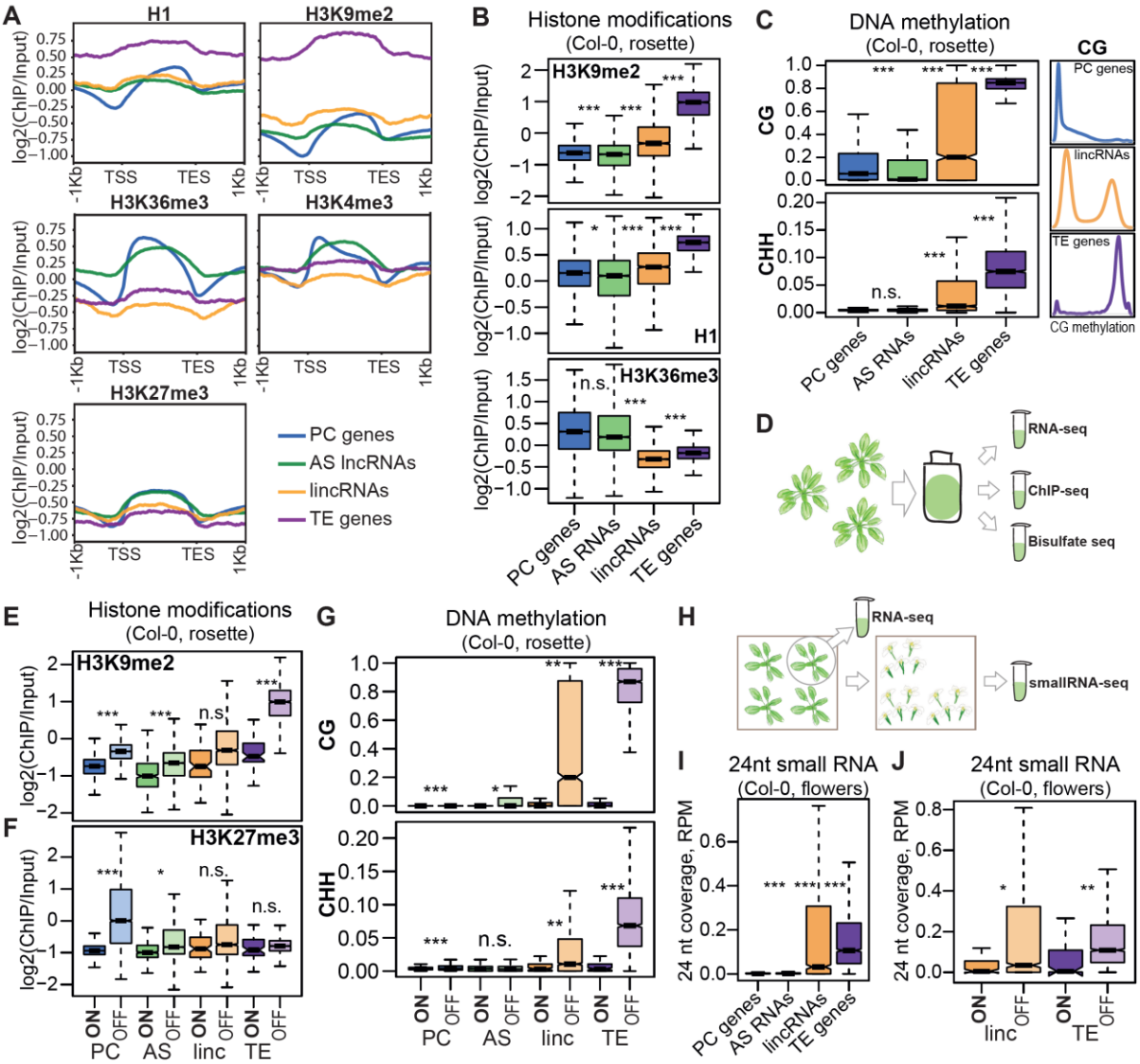


Fig. 3. Epigenetic patterns of lncRNAs in *A. thaliana* indicate ubiquitous silencing.

A. Averaged profiles of the input-normalized ChIP-seq signal for H1, H3K9me2, H3K36me3, H3K4me3 and H3K27me3 over 4 gene types from our cumulative transcriptome annotation. The plots show data from Col-0 rosettes, replicate 2. All genes, expressed and silent in Col-0, are used for the analysis. Profiles were built using plotProfile from deeptools (Ramírez et al., 2016). **B.** H3K9me2, H1 and H3K36me3 histone modifications in Col-0 rosette. The log2 of the gene-body coverage normalized by input and averaged between 2 replicates is plotted. **C.** Left: CG and CHH DNA methylation levels in Col-0 rosette. Right: density of CG methylation level for PC gene, lincRNA and TE gene loci. Methylation level is calculated as the ratio between the number of methylated and unmethylated reads over all Cs in the respective context (CG/CHH) in the gene body and averaged over 4 replicates. **D.** The scheme of the experiment: same tissue was used for RNA-seq, ChIP-seq and Bisulfite-seq in this study. **E.** H3K9me2 input-normalized coverage separately for expressed (ON, TPM>0.5) and silent genes (OFF, TPM<0.5). **F.** H3K27me3 normalized coverage separately for expressed (ON, TPM>0.5) and silent (OFF, TPM<0.5) genes. **G.** Methylation levels for expressed (ON, TPM>0.5) and silent (OFF, TPM<0.5) genes. H. **I.** Coverage of 24nt small RNAs in the gene body, calculated as the number of 24nt reads mapping to the locus divided by the total number of reads and the locus length. **J.** Coverage of 24nt small RNA separately for expressed (ON, TPM>0.5) and silent (OFF, TPM<0.5) genes. P-values were calculated using Mann-Whitney test on equalized sample sizes: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not shown.

As TE genes and lincRNAs are enriched next to the centromeres while PC genes and AS RNAs are not (Suppl. Fig. S4), we checked if the observed epigenetic differences held true when controlling for chromosomal position. While pericentromeric genes within 2 Mb of the centromeres all showed more heterochromatic patterns, the observed trends for histone marks and DNA methylation held true, especially for the genes further than 2 Mb from the centromeres (Suppl. Fig. S13, S14).

To confirm that the repressive marks at lncRNA loci are associated with silencing, we checked for epigenetic differences between the expressed and silent genes (the same samples were used for ChIP-seq/bisulfite-seq and RNA-seq, Fig. 3D). The repressive H3K9me2 (Fig. 3E, Suppl. Fig. S15A) and H1 (Suppl. Fig. S15B, C) were significantly higher on silent genes. While that was true for all gene categories, the H3K9me2 difference for TE genes was particularly high, underlining the fact that TEs are normally silenced by H3K9me2 (Feng and Jacobsen, 2011). Another repressive mark, H3K27me3, also showed significantly higher levels on silent genes of all categories, but here PC genes showed a striking increase while TE genes were minimally different (Fig. 3F, Suppl. Fig. S15D). We also found that silent AS lncRNAs showed increased CG methylation and silent lincRNAs showed strikingly increased CG and CHH methylation, although less so than TE genes (Fig3.G, Suppl. Fig. S16). Expressed PC genes had higher CG gene body methylation than silent ones which is a known phenomenon of however yet unclear function (Bewick and Schmitz, 2017).

Since lincRNAs showed both CHH methylation and H3K9me2, characteristic for TEs and absent from PC genes (Fultz et al., 2015), we performed small RNA sequencing of flowers (Fig. 3H, Suppl. Table S7) to look for evidence of targeting by the 24nt small RNAs that are normally involved in TE silencing by RNA-directed DNA methylation (RdDM) (Matzke and Mosher, 2014). This analysis demonstrated that small RNAs were indeed targeting many lincRNAs (Fig. 3I) and were associated with silencing for both lincRNAs and TE genes (Fig. 3J). Analyzing published sRNA-seq data (Papareddy et al., 2020) from leaves and a very early embryonic stage known as “early heart” where RdDM-mediated silencing is particularly active (Papareddy et al., 2020) confirmed 24nt sRNA targeting of lincRNAs, with early heart showing very high levels of sRNAs (Suppl. Fig. S17).

Figure 4

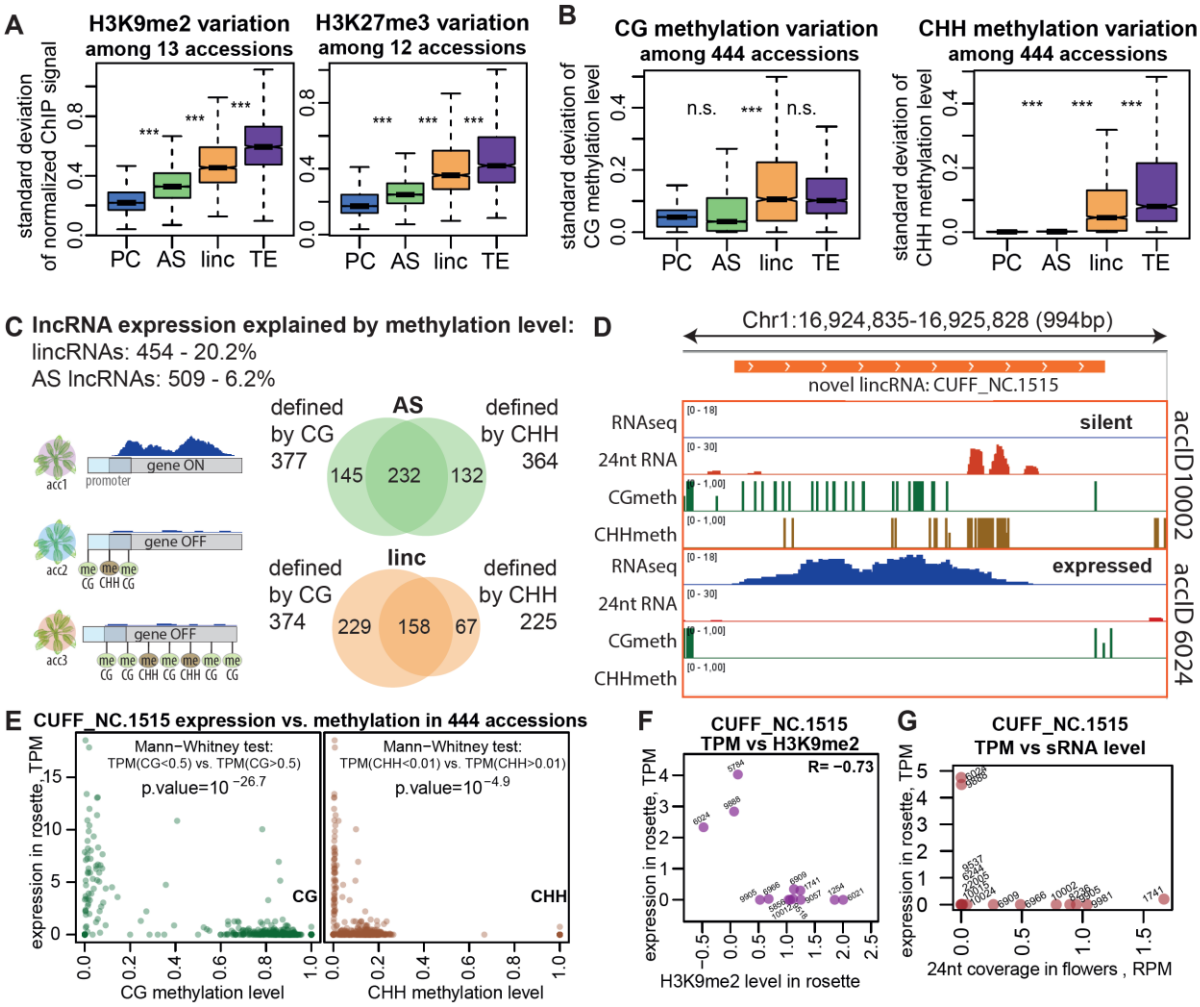


Fig. 4. LncRNAs display increased epigenetic variation that explains expression variation of many lncRNAs.

A. Standard deviation of input and quantile normalized coverage (see Methods) of H3K9me2 (left) and H3K27me3 (right) in rosettes across 13 and 12 accessions respectively. **B.** Standard deviation of CG (left) and CHH (right) methylation levels across 444 accessions (rosettes, 1001G dataset, (Kawakatsu et al., 2016)). P-values were calculated using Mann-Whitney test on equalized sample sizes: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not shown. **C.** The summary of lncRNAs for which expression can be explained by methylation (Suppl. Fig. S20B). The colored circles show the overlap between loci for AS lncRNAs (green) and lincRNAs (orange) that were found to be defined by CG or CHH methylation level. **D.** An example of a lincRNA defined by CG and CHH methylation. The figure shows RNA-seq signal (forward strand), CG and CHH methylation levels in rosette and the 24nt sRNA signal in flowers in 2 accessions. **E.** The plots show the expression level as a function of CG/CHH methylation for the example lincRNA across 444 accessions (Kawakatsu et al., 2016). The results of the Mann-Whitney tests used for defining the explanatory power of CG/CHH methylation are shown. **F.** Expression in rosettes vs. H3K9me2 level in rosettes of the example lincRNA in 13 accessions. **G.** Expression in rosettes vs. 24nt sRNA coverage in flowers of the example lincRNA in 14 accessions.

Epigenetic variation explains expression variation of many lncRNAs

In the previous section we described the epigenetic patterns within the reference accession, Col-0, only. As we produced ChIP-, bisulfite- and small-RNA-sequencing data for several accessions (Suppl. Tables S5, S6, S7) we were able to confirm that the epigenetic patterns we observed in Col-0 were similar in other accessions (Suppl. Fig. S18). However, while the overall patterns were similar, the variability between accessions at particular loci was very high for both lncRNA (especially lincRNA) and TE genes (Fig. 4A, B, Suppl. Fig. S19).

To test whether epigenetic variation can explain expression variation, we analyzed rosette methylation and expression data from 444 accessions (Kawakatsu et al., 2016) and found that for 454 lincRNAs and 509 AS lncRNAs, expression across accessions was indeed defined by the level of CG/CHH methylation at their gene body or promoter (Fig. 4C, Suppl. Fig. S20A). While these numbers correspond to only 20.2% and 6.2% of all lincRNAs and AS lncRNAs respectively, the analysis could only be performed on a limited number of informative loci with sufficiently high methylation variation and expression frequency (Suppl. Fig. S20B, Methods). Among these informative loci, we could explain expression variation by DNA methylation variation for 50.7% of lincRNAs and 21.5% of AS lncRNAs.

An example of such a lncRNA is displayed in Fig. 4D: the accession that expresses the lincRNA lacks CG and CHH methylation in the locus as well as 24nt sRNAs while the accession where the lincRNA is silent has both CG and CHH methylation, and 24nt sRNAs in flowers. The epigenetic variation in this locus is extensive and quite dichotomous with very strong association between the presence of methylation and the lack of expression and vice versa (Fig. 4E). For the samples where data were available, the repressive histone modifications H1 (Suppl. Fig. S20C) and H3K9me2 (Fig. 4F), as well as 24nt sRNA coverage (Fig. 4G), were also anticorrelated with expression across accessions.

In summary, we found that lncRNAs display distinctive epigenetic patterns consistent with the above observation of the lack of expression and suggesting ubiquitous silencing. Compared to PC genes, lncRNAs displayed increased epigenetic variation between natural accessions that explained expression variation of ~20% of lincRNAs and ~6% of AS lncRNAs. Many lincRNAs showed TE-like epigenetic status that was associated with silencing and were even targeted by

24nt siRNAs characteristic of the RdDM pathway of TE silencing. Because of the interesting patterns and the outstanding variation we observed for lincRNAs, we focus on them in what follows.

lincRNAs are enriched for TE pieces

Several similarities between lincRNAs and TE genes are apparent. First, both showed increased expression in flowers and pollen (Fig. 2F). Second, lincRNA expression was dramatically more variable than that of PC genes and antisense lncRNAs — almost at the level of TE genes (Fig. 2A). Third, our survey of the epigenetic landscape showed that lincRNAs display TE-like characteristics, though to a lesser extent (Fig. 3A,B,F). Fourth, similarly to expression variation, lincRNA repressive chromatin levels were more variable than that of PC genes and AS lncRNAs, reaching towards TE genes (Fig. 4A, B). Furthermore, it is known that lncRNAs can originate from TEs and contain parts of their sequences (Kapusta et al., 2013), and that TE domains within lncRNAs can play significant roles in lncRNA biology (Johnson and Guigó, 2014), such as their nuclear export or retention (Lubelsky and Ulitsky, 2018), or even have a crucial role for their function (Colognori et al., 2020). Based on this, we decided to investigate if TE sequences contribute to lincRNA loci in *A. thaliana* and affect their expression, epigenetics, and variability.

We used a BLAST-based analysis to identify sequences similar to TAIR10-annotated TEs inside loci and their borders (Fig. 5A, Methods). We called each match a “TE piece”, merged overlapping same-direction TE pieces into “TE patches”, and further refer to each TE-like region of a locus as a “TE patch” (Fig. 5A). LincRNA loci were clearly enriched in TE patches compared to AS lncRNAs and PC genes, as well as randomly picked intergenic regions of corresponding length (Fig. 5B, Methods). 52% of lincRNAs but only 27% of matching intergenic controls contained a TE patch. We also observed an enrichment of TE patches in upstream and downstream lincRNA border regions compared to matching controls (Fig. 5B). On a per-kb basis, lincRNA borders showed the highest density of TE patches, even higher than that of lincRNA loci, and the difference between lincRNAs and other gene types became even more prominent (Fig. 5C). TE genes had fewer TE patches per 1kb than lincRNAs, presumably because TE genes usually contain one big TE patch – the full TE — while lincRNAs contained several smaller patches (Suppl. Fig. S21A).

Figure 5

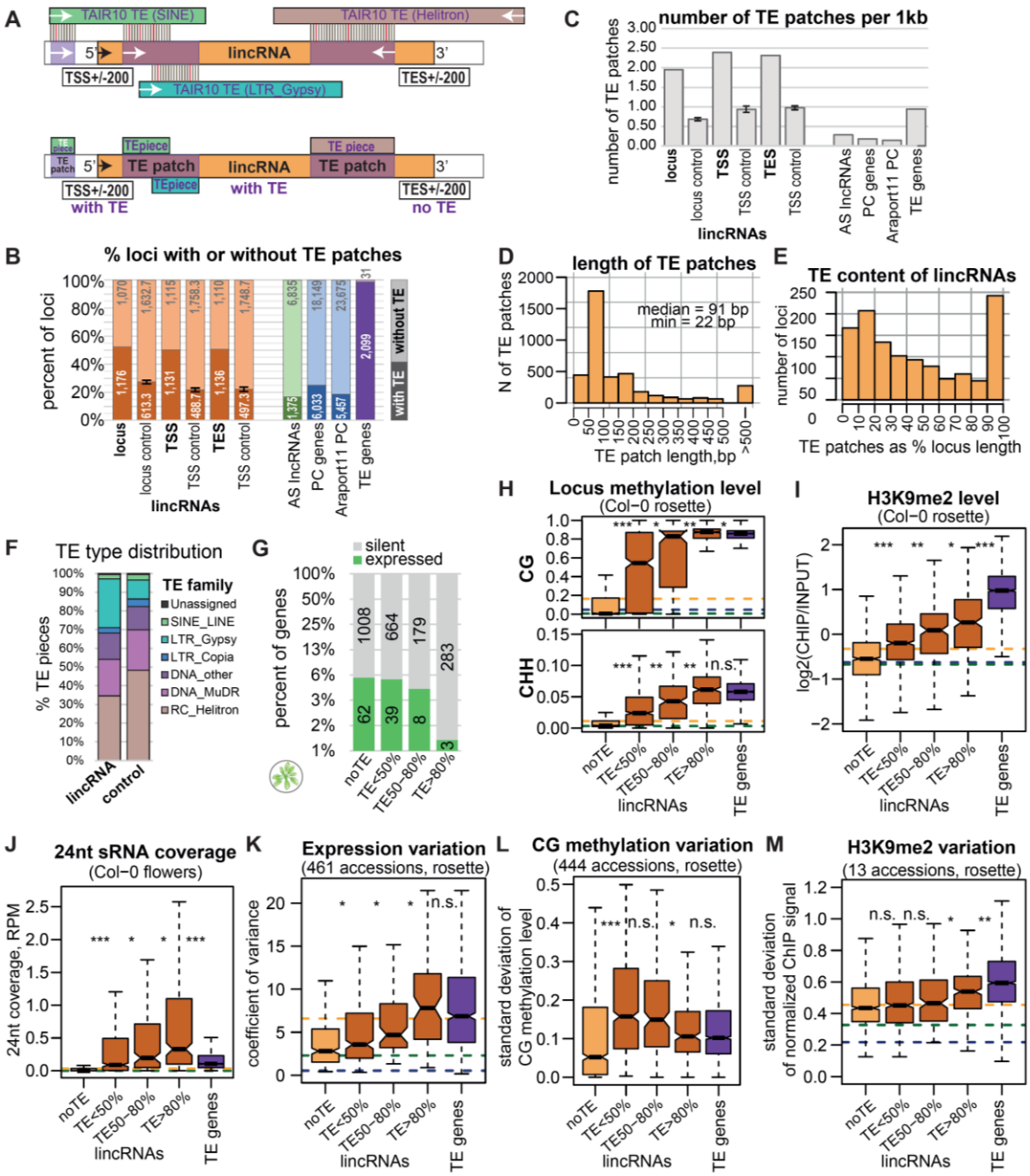


Fig. 5. Many lincRNAs contain pieces of TEs that affect their silencing and variation.

A. Outline of TE-content analysis. Top: TAIR10-annotated TEs were blasted to the sequences of lincRNAs (and other loci). Bottom: The mapped pieces of different TEs overlapping in the same direction were merged into "TE patches". The upstream and downstream "borders" of genes were analyzed in the same way. **B.** The fraction of loci containing a TE piece. The intergenic controls for lincRNAs, lincRNA TSS+/-200bp and TES +/-200bp were obtained by shuffling the corresponding loci within intergenic regions (lincRNAs excluded) 3 times and averaging the results. The error bars on controls represent the standard deviation between the 3 shuffling replicates. **C.** The number of TE patches per 1 kb. **D.** Distribution of the length of TE patches (any relative direction) within lincRNA loci. **E.** TE content distribution among lincRNAs. TE patches in any relative direction. The loci with large TE content are those where the TE patches are mapping antisense to the lincRNA locus. **F.** The proportion of TE pieces of different TE families inside different types of loci. **G.** The proportion of expressed lincRNA as a function of their TE-content. The y axis is

displayed in log scale. **H, I, J.** Levels of methylation (**H**), H3K9me2 (**I**), and 24nt sRNAs (**J**) for lincRNA loci as a function of TE-content, with TE genes for comparison. **K,L,M.** Expression variability between 461 accessions (Kawakatsu et al., 2016) (**K**), standard deviation of CG methylation levels across 444 accessions (Kawakatsu et al., 2016) (**L**) and standard deviation of quantile and input normalized H3K9me2 levels in rosettes across 13 accessions (**M**) of lincRNA loci as a function of TE-content, with TE genes for comparison. P-values were calculated using Mann-Whitney tests: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not shown.

It is important to note that lincRNA are not simply expressed TEs. Our lncRNA annotation pipeline required that a lncRNA did not overlap any TE genes and allowed for a maximum of 60% same-strand exonic overlap with annotated TE fragments (Suppl. Fig. S1). While only 22% of lincRNAs had a same-strand exonic overlap with a TE fragment, 52% contained a TE patch (Fig. 5B), both sense and antisense to the lincRNA direction (Fig. 5A, Suppl. Fig. S21B). The protein-coding potentials of lincRNAs and TE genes were also very different (Suppl. Fig. S2A). TE patches within lincRNAs (and other genes) were generally short with a median length of 91bp and a minimal length of 22bp (Fig. 5D) —much shorter than TAIR10-annotated TE fragments or the TE patches our analysis identifies in TE genes (Suppl. Fig. S20C, D). The relative TE content of TE-containing lincRNA loci differed greatly, with a few fully covered by TE patches, corresponding to lincRNAs that are antisense to a TE fragment (Fig. 5E, Suppl. Fig. S22A). On average, lincRNAs contained 309bp of same-strand TE-like sequences and 436bp of antisense TE-like sequences per kb. LincRNA loci were particularly enriched in LTR_Gypsy TE sequences compared to matching intergenic controls and other gene types (Fig. 5F, Suppl. Fig. S22B), and this enrichment was particularly pronounced in lincRNAs with antisense TE patches (Suppl. Fig. S22C-D). We did not however observe any particular localization for TE pieces from different families within the lincRNA loci (Suppl. Fig. S22E-F).

TE-content of lincRNAs affects their expression and epigenetics

We asked if the TE content of a lincRNA affects its expression, epigenetic characteristics, and their variability. Indeed, when binned based on relative TE-sequence content, lincRNAs with higher TE content were less often expressed (Fig. 5G) and showed higher levels of CG and CHH methylation (Fig. 5H), H3K9me2 (Fig. 5I) and 24nt siRNA (Fig. 5J). The expression variation (Fig. 5K, Suppl. Fig. S23) and epigenetic variation (Fig. 5L, M, Suppl. Fig. S24) of lincRNAs also depended on the TE content, but not as strongly.

LincRNAs are enriched in the pericentromeric regions (Suppl. Fig. S4) that are naturally enriched in TEs and heterochromatin, which might confound our TE piece (Fig. 5B-C) and epigenetic analyses (Fig. 5H-J). Controlling for the proximity to centromeres, we first found that while all gene types have higher TE-content closer to centromeres, the trend observed in Fig. 5B was preserved (Suppl. Fig. S25A). Second, while all pericentromeric lincRNAs, even those without TE patches, showed high repressive chromatin, the level of heterochromatic marks at lincRNA loci further from centromeres strongly depended on their TE content (Suppl. Fig. S25B-D). Furthermore, while 24nt sRNA coverage was generally low near centromeres (consistent with previous findings, see (Sigman and Slotkin, 2016)), it strongly depended on the TE-content in chromosome arms (Suppl. Fig. S25E). Thus, the presence and the relative size of TE-sequences inside lincRNA loci are indeed associated with a more repressive chromatin state irrespective of chromosomal location.

In summary, we found that intergenic lncRNAs are highly enriched for short pieces of TEs. About half of lincRNAs have a TE sequence inside, and higher TE-content is associated with more repressive epigenetic marks when comparing different lincRNAs in the genome.

Copy number of lincRNAs affects their expression variability and epigenetic patterns

Apart from expression variability, epigenetic patterns and TE-sequence content, another classical TE feature was evident for lincRNAs: lincRNAs are often present in multiple copies (Suppl. Fig. S26A). We decided to investigate this pattern further and see whether it affects their epigenetic patterns and expression.

We used a blast-based approach to look for multiple gene copies in TAIR10 (Methods) and found that lincRNAs were much more commonly multiplicated than PC genes and AS lncRNAs, with 28% being present in more than one copy and 8% in more than 10 copies (Fig. 6A). Again, lincRNAs were intermediate between PC genes and TE genes. We split lincRNAs into four categories: single- and multi-copy lincRNAs with and without TE patches (Fig. 6B, top). Similarly to the overall lincRNA distribution (Fig. 5B), about a half of single-copy lincRNAs contained a TE patch, while for multi-copy lincRNAs it was the majority (Fig. 6B). LincRNAs with higher copy numbers also showed higher TE-sequence content (Suppl. Fig. S26B). Although Helitrons

have the highest copy number in the *A. thaliana* genome (Quesneville, 2020), lincRNA with pieces of LTR_gypsy elements showed the highest copy number (Suppl. Fig. S26C), even when the TE-sequence content was no more than 20% of the locus (Suppl. Fig. S26D).

Figure 6

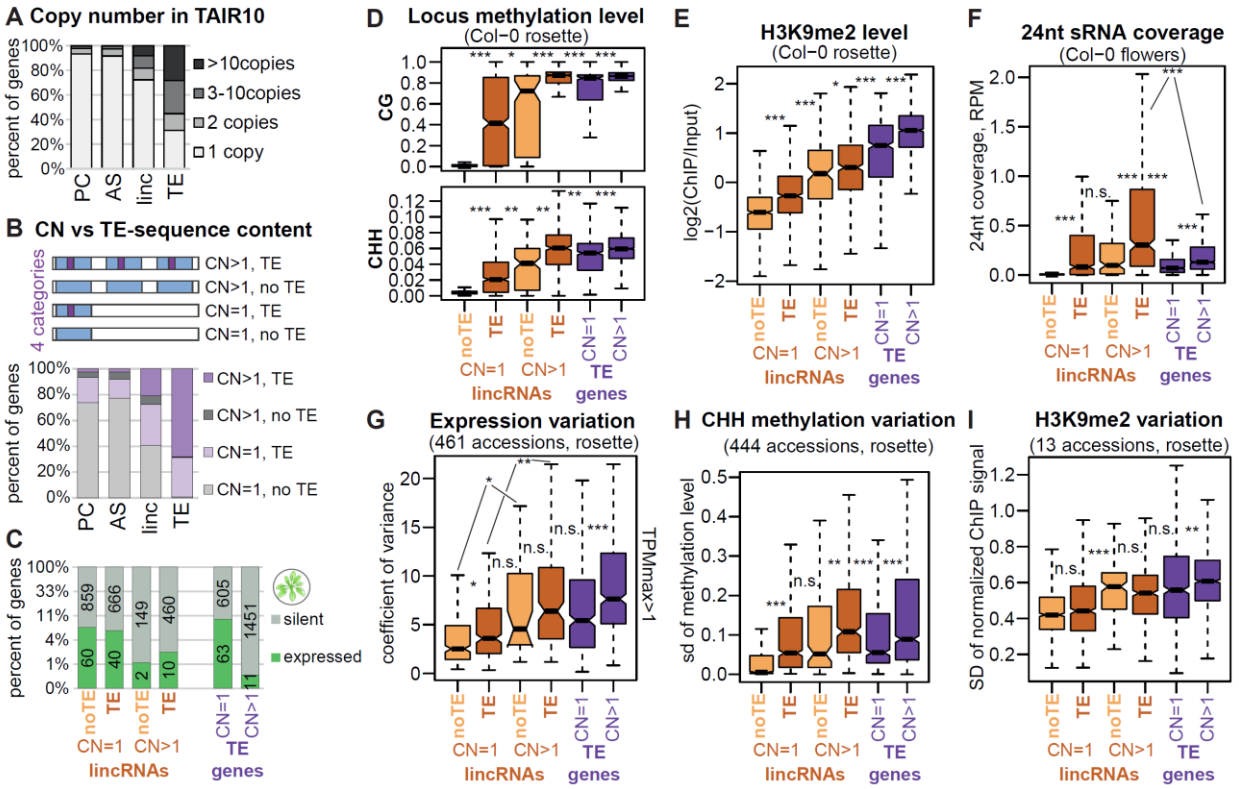


Fig. 6. Copy number of lincRNAs affects their epigenetic patterns and variability.

A. The copy number distribution for PC genes, AS lncRNAs, lincRNAs and TE genes from the cumulative transcriptome annotation in the TAIR10 genome. **B.** Top: the scheme of the 4 types of loci, bottom: the distribution of copy number of the 4 types of loci: 1 copy with no TE patch, 1 copy with a TE patch, multiple copies with no TE patch in the original locus and multiple copies with a TE patch in the original locus. **C.** The bar plot shows the proportion of the 4 types of lincRNAs, and 2 types of TE genes expressed (TPM>0.5, green) or silent (TPM<0.5, gray) in Col-0 rosettes. **D-F.** The boxplots show the CG and CHH methylation (**D**), H3K9me2 level (**E**) in Col-0 rosettes and 24nt sRNA coverage in Col-0 flowers (**F**) for the 4 types of lincRNAs and 2 types of TE genes. **G-I.** The boxplots show expression (**G**), CG methylation (**H**) and H3K9me2 (**I**) variability for the 4 types of lincRNAs and 2 types of TE genes. P values in the boxplots are calculated using Mann-Whitney test: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not plotted.

We analyzed the features of the four categories of lincRNAs and found that increased copy number was associated with reduced expression (Fig. 6C) and increased repressive chromatin (Fig. 6D-F, Suppl. Fig. S27A-F), as well as expression and epigenetic variability (Fig. 6G-I, Suppl. Fig. S27G-I). The presence of a TE patch within multi-copy lincRNA loci was associated with strikingly

increased CG and CHH methylation (Fig. 6D) and 24nt siRNA targeting (Fig. 6F) but did not seem to affect the level H3K9me2 and H1 levels and variability (Fig. 6E, I, Suppl. Fig. S27D, I) .

In summary, we showed that many lincRNAs are present in multiple copies and that increased copy number is associated with increased silencing and variability. The effect is in addition to the effect of TE patches, in that lincRNA with both multiple copies and the TE patches (i.e. most TE-like), showed the highest level of silencing.

lincRNAs are silenced by TE-like and PC-like mechanisms

We have seen that lincRNAs are ubiquitously silenced, with ~96% of loci OFF in rosettes of any particular accession (Fig. 2F) and very few accessions expressing any particular lincRNA (Fig. 2A). We have also seen that TE pieces inside lincRNAs are associated with repressive chromatin and siRNA targeting (Fig. 4F-H), at least when comparing lincRNA loci within a single genome. We investigated these patterns in greater detail, connecting them to known silencing pathways.

First, we observed that silent lincRNAs show a dichotomy when it comes to which silencing mark – H3K9me2 or H3K27me3 – is covering the locus (Fig. 7A). The same dichotomy was found across all gene types, but whereas almost all TE genes showed H3K9me2 silencing and most PC genes and AS lncRNAs were covered with H3K27me3, lincRNAs were split into two large categories (Fig. 7B, Suppl. Fig. S28). We thus defined two non-overlapping classes of lincRNAs: H3K9me2 and H3K27me3 lincRNAs, or K9 and K27 lincRNAs for short (Fig. 7A). K27 lincRNAs were almost free of TE patches, were present in one copy, and showed low DNA methylation and 24nt siRNA targeting, while K9 lincRNAs tended to have higher TE content, multiple copies, and show strikingly more DNA methylation and sRNA targeting (Fig. 7C-G). We thus could also label K27 and K9 lincRNAs as PC-like and TE-like respectively. The bimodality in the epigenetic features we observed before (Fig. 3) could be explained by lincRNAs being a heterogeneous group of PC-like and TE-like lincRNAs with different features.

PC-like lincRNAs are likely silenced by PRC2 that is known to establish the H3K27me3 repressive mark (Hansen et al., 2008), however, we did not try to confirm this hypothesis. Instead, we focused on the TE-like lincRNAs and hypothesized that the TE-like epigenetic patterns we observed were due to TE-silencing pathways.

Figure 7

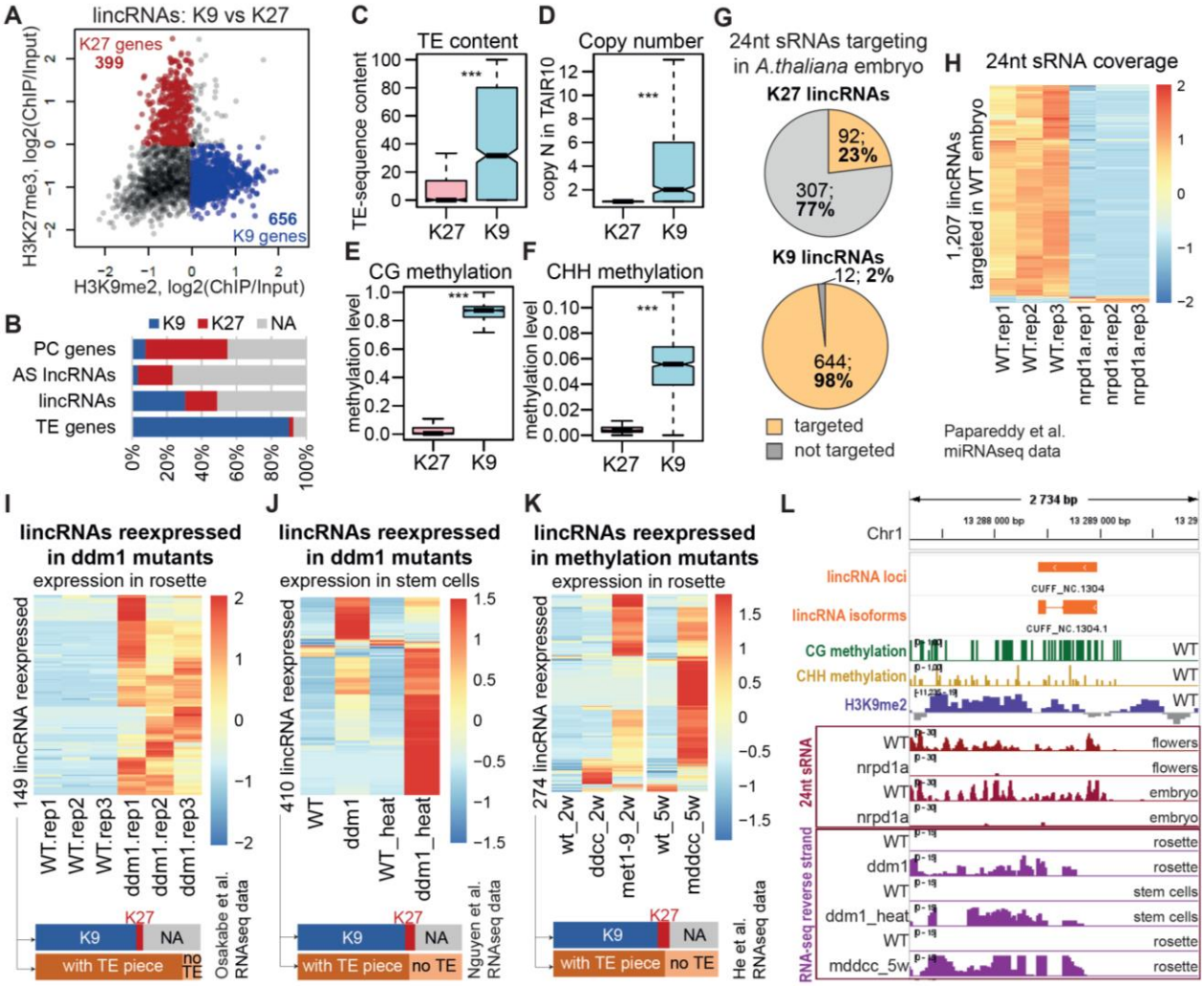


Fig. 7. lincRNAs are silenced by PC-like and TE-like mechanisms.

A. H3K27me3 vs. H3K9me2 level on lincRNA loci in Col-0 14-leaf rosettes (average of 2 replicates). K27 genes: red, K27 signal>0, K9 signal<0. K9 genes: blue, K27 signal<0, K9 signal>0. **B.** The distribution of K9 (blue) and K27 (red) genes among the 4 gene types. NA: genes with neither mark (gray, K27 signal<0, K9 signal<0). **C-F.** The boxplots show the relative TE-sequence content (**C**), copy number (**D**), CG (**E**) and CHH methylation level (**F**) of lincRNA loci classified as K27 and K9 genes. Outliers not plotted. P-values calculated using Mann-Whitney tests: ***$p<10^{-10}$. **G.** Relative number of K27 and K9 lincRNAs targeted by 24nt sRNAs (RPM>0.03) in *A. thaliana* embryos ("early heart" stage) (Papareddy et al., 2020). The small RNA coverage is averaged across 3 replicates. **H.** 24nt sRNA coverage in *A. thaliana* embryos ("early heart" stage) in the wild type (WT, Col-0) and in PolIV-deficient mutants (*nrpd1a*, Col-0 background) (Papareddy et al., 2020). 1,207 lincRNAs that are targeted (RPM>0.03, average of 3 replicates) by 24nt sRNAs in the WT are plotted. **I.** Expression level of the 149 lincRNAs re-expressed upon *ddm1* knockout in Col-0 rosettes (Osakabe et al., 2021). The bars at the bottom show the distribution of K9 (blue) vs. K27 (red) and TE-containing (dark orange) vs. TE-free (light orange) loci among the re-expressed lincRNAs (same for **J** and **K**). **J.** Expression level of the 410 lincRNAs re-expressed in stem cells upon *ddm1* knockout in Col-0 with or without heat stress treatment (Nguyen et al., 2023). **K.** Expression level of lincRNAs re-expressed upon DNA methylases knockouts in Col-0 rosettes (He et al., 2022) (see Methods). Heatmaps built using "pheatmap" in R with scaling by row. No column clustering, row clustering trees not displayed. **L.** An example of a lincRNA epigenetically silenced in Col-0 WT but expressed in the silencing mutants.

TEs in plants are thought to be silenced by two main mechanisms. First, in the RNA-directed DNA methylation mechanism known as RdDM (Onodera et al., 2005), PolIV-transcribed RNA from TE loci is turned into 24-nt small RNAs that guide DNA methylation machinery to the locus of transcription as well as to all homologous loci, allowing this mechanism to recognize and silence newly inserted TEs as well (Fultz et al., 2015). The second mechanism known to maintain TE silencing involves DDM1, MET1, CMT2 and CMT3 working together to establish the repressive H3K9me2 histone mark and DNA methylation at TE loci (Osakabe et al., 2021; Sigman and Slotkin, 2016). To test whether lincRNAs are also actively silenced by these mechanisms, we made use of publicly available RNA-seq data from knockouts of the TE silencing machinery in *A. thaliana*.

First, we analyzed the effect of inactivating the RdDM pathway. We have seen above that 24nt siRNA targeted ~50% of lincRNA loci in flowers. Analysis of the small RNA data from Papareddy et al (Papareddy et al., 2020) showed that 54% of lincRNA loci are targeted in early embryos and this targeting is highly specific to K9 lincRNAs (Fig. 7G). Knocking out NRPD1, the largest subunit of PolIV, causes a dramatic loss of 24nt small RNA coverage over 98% of those lincRNAs in embryos (Fig. 7H) as well as flowers (Suppl. Fig. S29) (Papareddy et al., 2020).

Next, we checked for the effect of removing DDM1, a key factor in TE silencing (Osakabe et al., 2021). Upon *ddm1* knock-out in Col-0, 149 of our lincRNAs were re-expressed in rosettes (Fig. 7I) and 410 lincRNAs were re-expressed in stem cells (Fig. 7J). Heat stress combined with knocking out *ddm1* was particularly beneficial for the reactivation of lincRNA, which is similar to TE behavior (Nguyen et al., 2023) (Suppl. Fig. S30). The removal of CG and non-CG DNA methylation in Arabidopsis also allowed re-expression of many lincRNAs in rosette (Fig. 7K). The re-expressed lincRNAs were again predominantly K9 lincRNAs and thus mostly TE-containing (Fig. 7I-K, bottom, Suppl. Table S8).

The re-expression of lincRNAs in the knockouts indicates two important points. First, that lincRNAs, predominantly the TE-like lincRNAs, are indeed silenced by the TE-silencing machinery. Second, while we see the majority of lincRNAs silent in any given accession, many of them retain the potential to be expressed, and must therefore be actively silenced instead of having been inactivated by mutations. In fact, an analysis spanning across the different tissues and mutants

with deactivated TE-silencing pathways in Col-0 showed that over 50% of our annotated lincRNAs can be expressed in Col-0 (Suppl. Fig. S31), in contrast to 4-10% normally expressed in one sample (Fig. 2F). Thus, it appears any genome is capable of expressing a large fraction of the numerous lincRNAs it harbors but they are actively silenced, presumably largely via TE-silencing pathways.

Figure 8

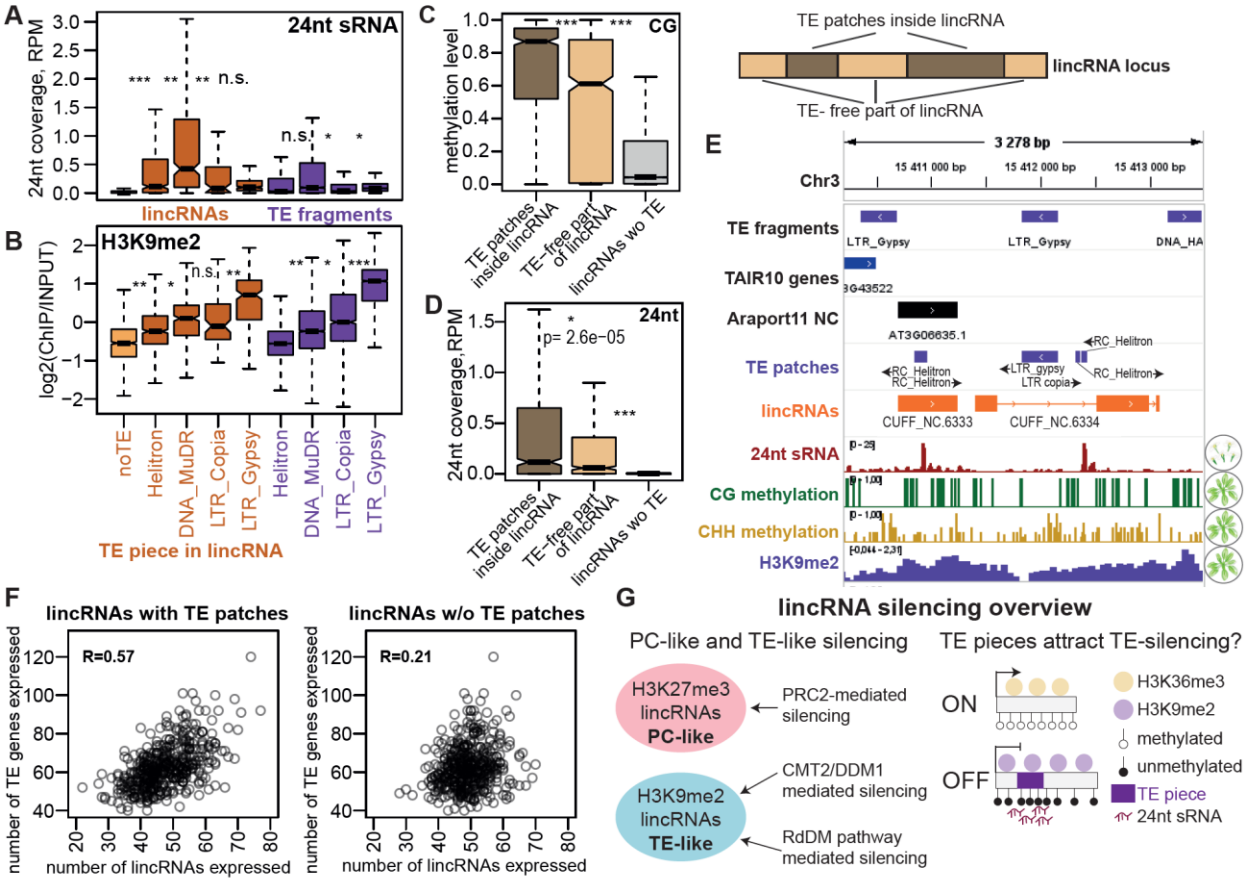


Fig. 8. TE pieces appear to attract silencing to lincRNA loci.

A-B. 24nt sRNA level in Col-0 flowers (**A**) and H3K9me2 level in rosettes (**B**) for lincRNAs with pieces of TEs from 4 superfamilies and TAIR10 TE fragments from the same superfamilies. Only lincRNAs with TE pieces from one superfamily are plotted. Light-orange box indicates lincRNAs without TE pieces. **C-D.** The boxplots show CG methylation level (C) and 24nt sRNA coverage (D) for TE patches inside lincRNAs, TE-patch-free parts of TE-containing lincRNA loci and lincRNA loci without TE patches. Outliers not plotted. P-values calculated using Mann-Whitney tests: ***$p<10^{-10}$, *$p< 0.01$. **E.** The IGV screenshot shows an example of lincRNAs with TE patches that have higher level of CG methylation and 24nt sRNA coverage over TE patches than over the rest of the locus. **F.** The scatter plot shows the number of TE genes expressed in rosettes of 460 different accessions (Kawakatsu et al., 2016) as a function of the number of lincRNAs with TE pieces (left) and without TE pieces (right) expressed in the same accession. Pearson correlation coefficient is displayed. **G.** The scheme summarizes lincRNA silencing pathways. PC-like lincRNAs that show H3K27me3 repressive histone marks are likely silenced by PRC2, while TE-like lincRNAs that display H3K9me2 are silenced by CMT2/DDM1 and RdDM pathways. TE piece presence likely attracts TE silencing and repressive chromatin to the lincRNA locus.

TE pieces appear to attract silencing to lincRNA loci

We have seen that the presence of TE pieces inside lincRNA loci is associated with increased epigenetic silencing (Fig. 5) and that TE silencing pathways predominantly affect lincRNAs with TE pieces (Fig. 7I-K). We hypothesized that TE pieces might be decisive for TE-like silencing of lincRNAs by attracting the silencing machinery to the locus. To investigate this we made use of the fact that different TE types show different silencing patterns. In particular, the RdDM pathway has been shown to be more prevalent for DNA elements (Class II TEs), which are heavily targeted by 24nt siRNAs, while retrotransposons (Class I TEs) such as LTR elements are more affected by the DDM1/CMT2 pathway showing heterochromatic patterns with high H3K9me2 levels (Sasaki et al., 2019; Sigman and Slotkin, 2016). If TE pieces inside lincRNA loci are decisive for their silencing, we would expect that lincRNAs with pieces of different types of TEs would show silencing patterns resembling the corresponding TEs. Our analysis confirmed this: lincRNA loci with pieces of DNA TEs, especially those of MuDR elements, showed significantly increased levels of 24nt sRNAs (Fig. 8A), and lincRNAs with pieces of LTRs, especially those of Gypsy elements, showed significantly increased H3K9me2 levels (Fig. 8B). While class I TEs are more prevalent in the chromosome arms and LTRs are enriched closer to the centromeres (Quesneville, 2020), the observed trends were preserved when controlling for chromosomal position (Suppl. Fig. S32).

Although TE patches usually constitute only a portion of a lincRNA locus (Fig. 5E), they are associated with full-length silencing (Fig. 5G). We analyzed repressive chromatin on the TE patch and TE-patch-free parts of lincRNA loci and found that while TE patches showed higher repressive chromatin epigenetic modification, there was a very significant increase in repressive chromatin also outside of TE patches (Fig. 8C-D, Suppl. Fig. S33), consistent with spreading of silencing (Sigman and Slotkin, 2016). While H3K9me2 generally covered the whole TE-containing locus, 24nt sRNAs and DNA methylation were more restricted to the TE pieces inside the loci (Fig. 8C-E, Suppl. Fig. S33).

Finally, we noticed that the numbers of lincRNAs and TE genes expressed in a given accession were quite well correlated (Suppl. Fig. S34A). TE silencing can vary across accessions and indeed, we found that the number of TE genes expressed across accessions varied nearly three-fold. The

correlation was much stronger for lincRNAs that contained TE pieces (Fig. 7N) supporting our hypothesis of shared silencing mechanisms. The number of TEs and lincRNAs expressed correlated in every tissue (Suppl. Fig. S34B), indicating the organism-wide success or failure of silencing. Interestingly, while the number of expressed loci correlated well between accessions, the correlation between the mean expression levels across expressed lincRNAs and TE genes was much lower (Suppl. Fig. S34C, D) indicating that the two loci types might share the silencing machinery, but likely not the general transcription apparatus and factors. We tried to identify genetic factors associated with the number of TE genes and lincRNAs expressed using GWAS, but could not see any clear association, except for one nearly significant peak on chromosome 2 near the XERICO gene (AT2G04240), encoding a Zinc Finger Protein domain, (Suppl. Fig. S35), which is interesting as ZFPs are thought to participate in TE silencing (Yang et al., 2017).

In sum, we have seen that lincRNAs display 2 distinct silencing mechanisms (Fig. 7P): PC-like silencing via H3K27me3 that is normally deposited by PRC2 (Hansen et al., 2008) and TE-like silencing, achieved via DDM1-CMT2 and RdDM silencing pathways (Fultz et al., 2015). The presence of TE pieces inside lincRNAs seems to induce their TE-like silencing (Fig. 7P).

Discussion

Extended *A. thaliana* lncRNA annotation

Unlike annotations based only on the reference accession Col-0, we used almost 500 natural accessions and several developmental stages and identified more than 10,000 novel lncRNA loci in the TAIR 10 reference genome. We conclude that over 10% of the genome can express lncRNAs, but that most are not expressed in any particular accession or tissue, preventing a comprehensive lncRNA identification from few accessions or tissues. Analyzing more accessions allows identification of more lncRNA loci—and there is little evidence of saturation even at several hundred accessions (Fig. 1F, Suppl. Fig. S5A). We provide an extended lncRNA annotation (Supplemental_Annotations) as a resource for the *Arabidopsis* research community. Our results also suggest that other plant lncRNA annotations could similarly be extended by population-wide studies.

The largest part of the lncRNA transcriptome consists of lncRNAs that are antisense to PC genes. Apart from the general problem of natural variation impeding lncRNA identification described above, identifying antisense lncRNAs crucially depends on having high-quality stranded RNA-seq data and a careful analysis that avoids artifacts (Suppl. Fig. S1). We were able to annotate almost 9,000 antisense lncRNAs with nearly 30% of PC genes having an antisense partner, which greatly extends the scope of antisense transcription. This is an important finding since most functional lncRNAs reported in *A. thaliana*, such as COOLAIR (Csorba et al., 2014), asDOG1 (Fedak et al., 2016), SVALKA (Kindgren et al., 2018), and recently SEAIRa (Chen et al., 2023), are antisense lncRNAs, and the massive extension of AS lncRNA annotation reported here thus opens a broad field for functional studies. A deeper investigation into antisense lncRNAs and their function is beyond this study, but we provide a list of 14 AS lncRNAs that show striking negative correlation in expression with their partner PC gene (Suppl. Table S9, Suppl. Fig. S36) and thus are excellent candidates for being regulatory.

The second largest class of lncRNAs were intergenic lncRNAs that do not overlap any PC genes, and these are the main focus of this paper. This type of lncRNAs is very actively studied in mammals, with many functional examples reported (Rinn and Chang, 2020). We focused on them

because they showed extreme expression variability and an interesting position intermediate between PC genes and TE genes in terms of expression, epigenetic features and variation (Fig. 2,3). The bimodal distribution of CG methylation levels was particularly striking (Fig. 3C). We also observed a clear dichotomy between H3K27me3 and H3K9me2 silencing (Fig. 5A) that was recently also reported by Zhao et al (Zhao et al., 2022). K27- and K9-silenced lincRNAs were distinct in many features, most strikingly TE piece content, which made them be similar to PC genes and TE genes, respectively, thus allowing us to distinguish two lincRNA subclasses: PC-like and TE-like.

TE pieces

We found that about a half of *A. thaliana* lincRNA loci contained sequences similar to TAIR10 annotated TEs, which we referred to as TE pieces, or patches when they held similarity to more than one TE superfamily. Strikingly, TE pieces were nearly 20 times more common inside lincRNA loci than inside PC genes and about 3 times more common than in random intergenic regions (Fig. 5C). It is unclear why lincRNA loci are so dramatically enriched in TE pieces. While the enrichment over PC genes is understandable as TE insertions can be more deleterious for PC genes than for lincRNAs, the enrichment over random intergenic regions is very interesting. As lincRNAs are simply expressed intergenic regions of the genome without protein-coding capacity, the enrichment suggests that having a TE piece inside increases the probability of transcription. While our analyses suggest that TE pieces are associated with silencing, they might also provide the ability to be expressed when silencing fails. TEs are known to be the source of novel promoters in various organisms (Sundaram and Wysocka, 2020). Thus, we can hypothesize that for many lincRNAs, TE pieces inside provide the potential for being transcribed, as well as contribute to it being silenced, albeit imperfectly, leading to our ability to detect these loci in our population-wide annotation. This hypothesis would go along with the extreme expression variability of TE-containing lincRNAs (Fig. 5K) and the very high variability in the overall level of TE gene and lincRNA expression (Fig. 7O) that indicates high TE-silencing variability. Alternatively (and arguably more obscurely), the enrichment of TE pieces inside lincRNAs is caused by their transcriptional activity if actively transcribed regions loci are more attractive for insertions compared to non-transcribed intergenic regions.

One very interesting group of TE-containing lincRNAs are lincRNAs with antisense LTR_gypsy elements. LincRNAs showed significant enrichment in LTR_gypsy pieces or often full elements in the antisense direction (Suppl. Fig. S22C). Why LTR_gypsy elements show antisense transcription more commonly than other elements remains to be investigated. One speculative hypothesis might be that these elements increase their mobilization chances by transcribing from another strand as strandness does not matter for the transposition of retroelements since it involves a double-stranded DNA step.

Another major topic that our results raise is the nature and origin of the TE pieces we find in lincRNA loci. Some of these TE pieces are simply parts of intact TE fragments that are overlapped by the lincRNA locus, and some are full TE fragments in the direction antisense to the direction of lincRNA transcription. In these cases, the nature of the TE sequence inside lincRNA is clear but the question of what came first — the expression or the TE — remains. Most intriguing are the many cases of short, and sometimes very short, independent pieces of TEs inside lincRNA loci, the nature of which is puzzling. First, these TE pieces might represent insertions into the loci. However, their small size (Fig. 5D) raises the question of how they were able to mobilize and get inserted into the lincRNA loci. Non-autonomous TEs (Quesneville, 2020), in particular, DNA-TE derived MITEs (miniature inverted-repeats transposable elements) (Oki et al., 2008) and LTR-TE derived SMARTs (small LTR-retrotransposons) have been studied (Mhiri et al., 2022), yet those still have a length of a few hundred bp, while our pieces are often around 100bp or shorter. It has also been suggested that small non-autonomous TEs can transpose with a piece of a nearby genomic sequence, thus shuffling it around, but there is little understanding of how this might work (Quesneville, 2020).

As many lincRNAs are known to originate from TEs (Kapusta et al., 2013), such as the famous XIST lncRNA (Colognori et al., 2020), it is also possible that the TE pieces we find inside lincRNAs are not insertions but rather remnants of decaying TEs. One approach to distinguishing this would be to study the structural variation of TE pieces: variability of the presence of that precise piece would clearly indicate insertion/excision rather than the decay of a larger TE. What we can assess within the scope of this paper is whether multiple TE pieces within one lincRNA locus resemble one or multiple TE families. If a locus contains pieces of different TEs this would be evidence against the TE decay hypothesis. Among lincRNAs with more than one TE piece in

TAIR10, 74% have TE pieces from different superfamilies and 24% from both ClassI and ClassII TEs. Further research and an analysis of full genomes from multiple accessions are crucial for understanding the nature, evolutionary history, and population dynamics of TE pieces inside lincRNAs.

Finally, some of the lincRNAs we detect might actually be previously unannotated active or recently active TEs. The TE-like lincRNAs showed many similarities to TEs and the presence of a patch similar to annotated TEs might resemble the occasional sequence likeness between annotated TEs of different families. However, to definitively conclude that a lincRNA locus is in fact a TE, we would need evidence of mobilization between accessions or species, and evidence of the TE piece inside the lincRNA locus being an integral part of it rather than an insertion – thus not showing variability between accessions. These analyses represent future directions and are out of scope for this study.

Silencing

We found that the *A. thaliana* genome has a large potential for lincRNA expression that is massively repressed by silencing. While many lincRNAs are repressed by PC-like H3K27me3 silencing, about as many are repressed by TE-silencing, and this is associated with having a TE piece inside the locus. The presence of a TE piece was correlated with repressive chromatin and silencing, and the higher the TE content of a locus, the stronger the silencing (Fig. 5E-H). We also showed that deactivating TE-silencing pathways in the reference accession Col-0 allows expression of many TE-like lincRNAs that are normally completely silent in this accession. It seems that TE pieces attract silencing to the locus, as we have seen that lincRNAs seem to be preferentially silenced by RdDM- or CMT2-silencing pathways depending on which TE family the TE piece inside the lincRNA locus comes from (Fig. 7E). Interestingly, TE pieces and multiple copy number were associated with the same patterns of silencing in both AS lncRNAs and PC genes (Suppl. Fig. S37), although the relative number of such TE-like genes was much smaller (Fig. 4B). This suggests that a genome-wide mechanism for suppression of TE-like loci exists (Sigman and Slotkin, 2016).

The mechanism by which short TE pieces attract TE-like silencing to a lincRNA locus is unclear. It is known that full-length TEs can induce the silencing of nearby genes by the spreading of repressive chromatin (Sigman and Slotkin, 2016) and we hypothesize that TE pieces are capable of that as well. However, how they themselves obtain repressive chromatin is unclear. One possibility is that 24nt siRNAs produced at TE loci find the TE pieces by homology and initiate silencing at this "TE-like" locus (Fultz et al., 2015). They likely initially target only the TE piece and not the full locus, and we do see that the 24nt siRNA and CG/CHH methylation signal is the highest at the TE patches (Suppl. Fig. S33). However, we also observed a significant increase of 24nt siRNA and CG/CHH methylation level outside of TE patches, which may suggest that spreading also includes small RNAs starting to be produced at the locus.

It is also unclear what causes the failure of silencing of certain lincRNAs in certain accessions. It is possible that the silencing machinery varies in efficiency, and we see some evidence for this in the 3-fold range of variation in the number of TE genes and TE-containing lincRNAs expressed across accessions (Fig. 7N). However, we could not find any gene expression level or SNP that was clearly associated with the overall extent of lincRNA or TE gene transcription. It is also unclear how variation in silencing efficiency could account for such a strong lincRNA landscape variability across accessions with similar overall extent of lincRNA transcription (Suppl. Fig. S38). One possibility here is that this reflects the presence of particular TE loci producing the appropriate siRNAs for TE pieces inside particular lincRNA loci.

Further studies are surely needed. In this study we focused on the reference genome, demonstrating that TE pieces inside lincRNA loci are important for silencing. Direct experiments, like inserting a TE piece into a TE-free lincRNA locus and accessing the resulting expression change were out of the scope for this study. Similarly, an analysis of the full genomes of multiple accessions, including variation for TE and TE-fragment content, should be informative, and such an analysis is underway.

lincRNA expression variation and future directions

Our study initially had two major goals: to create a population-wide map of lncRNA transcription in *A. thaliana* and characterize its natural variation. We discovered that the extent of lncRNA

transcription in *A. thaliana* is much larger than previously thought and that lncRNA expression patterns are largely variable between accessions with half of lncRNAs expressed in one accession being off in another (Fig. 2E). In this paper, we characterized the expression variability of lncRNAs in *A. thaliana*, but among the factors that could explain the expression variation across accessions we only accessed the epigenetic patterns. We have shown that lncRNAs display extensive epigenetic variation (Fig. 4A,B) and this variation can explain the expression of ~50% of informative lincRNAs and ~20% of informative AS lncRNAs (Suppl. Fig. S20B). While purely epigenetic variation is well-known (Rajpal et al., 2022; Xu et al., 2019), our analysis did not distinguish between this and the case when the epigenetic variation that defines expression variation is itself defined by an underlying genetic or structural variation. We showed that two structural features of lincRNA loci – their TE content and their copy number – are associated with silencing and increased expression and epigenetic variation (Fig. 5, 6), and it is clear that variation in these two features might be responsible for the variation in expression that we observe between accessions. In this paper, we constrained our analysis to the reference genome, because an analysis of structural variation in copy number or TE-piece-presence requires full-genome assemblies of non-reference accessions. We perform those analyses in the upcoming study that investigates the determinants of lincRNA expression across accessions in greater depth.

Conclusions

Analyzing transcriptomes from multiple accessions and tissues of *A. thaliana* allowed us to drastically extend its lncRNA annotation and study the natural variation of lncRNA expression. We found that 10% of the *A. thaliana* genome is covered with almost 12,000 lncRNA loci, however, most of them are silent in any given sample. LncRNAs, particularly intergenic lncRNAs, show very high expression and epigenetic variation. The silencing of lincRNAs is achieved via PC-like and TE-like mechanisms, with the latter being defined by the pieces of TEs present in about half of lincRNAs. We produce a multi-accession transcriptomic and epigenetic resource, as well as a more comprehensive lncRNA annotation useful for the *A. thaliana* community, and provide novel insights into the genome biology and composition of lncRNAs.

Materials and methods

Sample collection

Seeds were sterilized with chlorine gas for ~1 hour, cold-treated at 4°C for ~5 days to induce germination and sown onto soil. Plants were grown in growth chambers at 21°C at long-day conditions (16h light, 8h dark). "14-leaf rosette" (or mature leaves) samples were collected at the 13-16-leaf stage before plants started to bolt. Approximately 8 leaves (avoiding the oldest and the youngest leaves) were collected from 2-3 individuals of the same accession, snap-frozen in liquid nitrogen, and stored at -70°C. Tissue was ground using metal beads while frozen, producing 1-2 ml of tissue powder that was used for RNA-, bisulfate- and ChIP-seq library preparation. For the "9-leaf rosette" samples we collected the full rosette at the 9-leaf stage with one plant harvested per sample. Seedlings were collected at 7 days post sowing (~5 days post germination). The full seedling with the root was harvested with approximately 10 individuals/seedlings harvested per sample. For the "flower" samples we collected flowers and flower buds from approximately 5 individuals per sample. Accessions 1741, 6024, 6244, 9075, 9543, 9638, 9728, 9764, 9888, 9905, 22003, 22004, 22005, 22006, 22007 were cold-treated by placing them into 10°C growth chambers (long-day conditions) for ~ 4 weeks to induce flowering. Accessions 6069 and 6124 did not flower even after the cold treatment. Pollen was collected using the method described in (Johnson-Brousseau and McCormick, 2004) that uses vacuum suction and a series of filters to harvest dry pollen from flowering plants. We used polyester mesh filters of 3 sizes - 150mm, 60mm and 10mm - for the collection. Collected pollen was snap-frozen in liquid nitrogen, stored at -70°C, and ground for RNA-sequencing using ~0.5ml of 0.5mm glass beads (Scientific Industries, Inc.).

RNA sequencing and analysis

RNA was isolated and DNAse I (NEB) treated using a KingFisher Robot with an in-house magnetic RNA isolation kit. RNA was diluted in nuclease-free water (Ambion) and stored at -80°C. Libraries for RNA-seq were prepared using TruSeq Stranded mRNA kit (Illumina) following the manufacturer's protocol with 4-minute RNA fragmentation time and 12 PCR cycles. RNA-seq was performed at the VBC NGS facility on an Illumina HiSeq 2500 machine with the paired-end read mode of 150bp and 125bp. Raw RNA-seq data was aligned to the TAIR10 genome

using STAR (Dobin et al., 2013) with the following options: --alignIntronMax 6000 --alignMatesGapMax 6000 --outFilterIntronMotifs RemoveNoncanonical --outFilterMismatchNoverReadLmax 0.1 --outFilterMismatchNoverLmax 0.3 --outFilterMultimapNmax 10 --alignSJoverhangMin 8 --outSAMattributes NH HI AS nM NM MD jM jI XS. Gene expression was calculated using featurecounts from the Subread package with -t exon option and an exonic SAF file as an annotation (Liao et al., 2014).

Transcriptome assembly and lncRNA annotation

Transcriptome assembly was done in several steps described in Suppl. Fig. S1 and the scripts are provided in (Kornienko). In brief, the following RNA-seq datasets were used for transcriptome assembly: 14-leaf-rosette data from 461 accessions (100bp single-end reads) (Kawakatsu et al., 2016), seedling data from Cortijo et al (Cortijo et al., 2019) (75bp paired-end, 14 replicates for each of the 12 samples were pooled), our 14-leaf-rosette data from 28 accessions (2-4 replicates each) (125bp paired-end), our seedling and 9-leaf-rosette data from 27 accessions (150bp paired-end), and our flowers and pollen data from 25 accessions (150bp paired-end). First, we assembled transcriptomes of each sample separately using Stringtie v.2.1.5 (Pertea et al., 2015) with options: -c 2 -m 150 -j 2.5 -a 15 guided by TAIR10 gene annotation (-G). Then we merged the transcriptome assemblies of the same tissue and data types using Cuffmerge (Cufflinks v.2.2.1) (Trapnell et al., 2010) with --min-isoform-fraction 0 and then performed a second merging of the resulting 7 transcriptomes to obtain the cumulative transcriptome annotation. A series of filtering steps were applied, including the transcript-length cutoff of 200nt for multiexon and 400nt for single-exon genes, and then the genes were split into 1. PC genes by exonic overlap with TAIR10 or Araport11 annotated PC genes, 2. TE genes by exonic overlap with Araport11 annotated TE genes, 3. TE fragments by >60% same strand exonic overlap with Araport11 annotated TE fragments, 4. Pseudogenes by exonic overlap with Araport11 annotated pseudogenes, 5. Initial lncRNAs by no overlap with PC genes, TE genes, pseudogenes, and <60% same strand exonic overlap with Araport11 annotated TE fragments. Then lncRNA transcripts (and the corresponding loci containing those transcripts) with protein-coding capacity tested by CPC2 (Kang et al., 2017) were removed, rRNA, tRNA, sn/snoRNA and miRNA precursor lncRNAs were classified by overlap with annotations and the rest of lncRNAs were classified into 1. Antisense lncRNAs by antisense overlap with TAIR10 or Araport11 annotated PC genes, 2. lncRNAs antisense to pseudogenes, 3.

lncRNAs antisense to Araport11 TE genes (AS_to_TE), 4. intergenic lncRNAs (lincRNAs) with no overlap with PC genes, TE genes or pseudogenes. LincRNAs were additionally filtered against loci that started <100bp downstream from annotated genes to avoid read-through transcripts. The number of Araport11 PC genes with an antisense transcript was calculated using Araport11 non-coding and novel_transcribed_region annotations filtered for genes longer than 200bp.

Gene saturation curve

To create the gene saturation curves for accession and tissue number, the annotation pipeline was automated and run many times with different numbers of accessions and tissues. The accession saturation curve was done by inputting 10 to 460 transcriptome assemblies (one assembly is one accession) obtained from the 1001G dataset (Kawakatsu et al., 2016) into the same annotation pipeline used for the main gene annotation. Subsampling of accessions was done randomly with 8 iterations for each accession number. The curve fitting and prediction of the saturation curve behavior with up to 1000 accessions was done by fitting a linear model using the lm function in R: model <- lm(y ~ x + I(log2(x))) (Suppl. Fig. S5A,B). The control for the accession saturation curve was done using the data from Cortijo et al. (Cortijo et al., 2019). We randomly picked from 1 to 12 transcriptome assemblies (corresponding to 12 samples with 14 replicates per sample pooled into one BAM file pre-assembly) and fed those into our standard annotation pipeline counting the number of loci identified as an output. The procedure was repeated 8 times for each assembly number. Then the number of reads was calculated and juxtaposed to the number of reads in the multi-accession saturation curve (Suppl. Fig. S5C). As different datasets had different read modes, we aligned the results by calculating the total read length and multiplying it by the total read number. Tissue saturation curve analysis was performed on 23 accessions that had data from all 4 tissues. Random sampling of accessions was performed with 8 iterations as replicates for each accession number. Tissues were assessed in this particular order: 1. seedling, 2. rosette, 3. flowers, 4. pollen without random sampling (Suppl. Fig. S6).

ChIP sequencing

Chromatin immunoprecipitation was performed with a protocol adapted from (Yelagandula et al., 2014) (full protocol is available at (Kornienko)). Briefly, 1-2 grams of ground frozen leaf tissue

was fixated with formaldehyde at 4°C for 5 min, then nuclei were isolated and lysed, and chromatin was fragmented using Covaris E220 Focused-ultrasonicator. The input sample was then taken out and frozen at -20°C and 5 antibody reactions were processed together. The antibodies used were: H1 (Agrisera) - 3 micrograms per reaction, H3K4me3 (Abcam) - 3 micrograms per reaction, H3K9me2 (Abcam) - 4 micrograms per reaction, H3K27me3 (Millipore) - 4 microgram per reaction, H3K36me3 (Abcam) - 4 microgram per reaction. The immunoprecipitation was performed using pre-washed Dynabeads Protein A magnetic beads (Invitrogen) and ran overnight. Afterwards the samples were washed, followed by elution, overnight reverse-crosslinking, RNAse A (Fermentas) treatment for 30 min at room temperature, and DNA isolation using Qiagen PCR purification kit Qiaquick with 0.3M Sodium Acetate. Next, ChIP-seq libraries were prepared from half of the resulting sample (due to very low amounts we did not measure the DNA concentrations) with the NEBNext Ultra II DNA kit (New England Biolabs) according to the manufacturer's protocol and sequenced with 100 bp single-end read mode on an Illumina NovaSeq 6000 machine.

ChIP-seq analysis

Raw ChIP-seq reads were mapped using STAR (Dobin et al., 2013) adjusted for ChIP-seq with the following options --alignIntronMax 5 --outFilterMismatchNmax 10 --outFilterMultimapNmax 1 --alignEndsType EndToEnd. Only samples with >1mln unique non-duplicated reads were used for the analysis. Aligned BAM files from each ChIP-sample were then normalized by the corresponding input samples using bamCompare from deeptools (Ramírez et al., 2016) with the following options: --operation log2 --ignoreDuplicates --effectiveGenomeSize 119481543 and bigwig and bedgraph files were created. Read coverage over loci and promoters was estimated using bedtools map on the bedgraph files with the "mean" operation. To estimate the variation of the histone modification levels, the chip-seq coverage values were normalized again to achieve the same range of values across accessions. For this we applied a quantile normalization, setting the 20%- and 80%-quantile values for each sample to the same value with the function in R: quantile_minmax <- function(x) {(x-quantile(x,.20)) / (quantile(x,.80) - quantile(x,.20))}. Histone modification variation was then calculated as the standard deviation of quantile-normalized levels averaged across replicates for each accession.

Bisulfite sequencing

Bisulfite sequencing was performed as described in (Pisupati et al., 2022). Briefly, DNA was extracted from the frozen leaf tissue (14-leaf rosettes) using Nuclear Mag Plant kit (Machery-Nagel) and the bisulfate libraries were prepared using a tagmentation method described in (Wang et al., 2013) using an in-house Tn5 transposase (IMBA-IMP-GMI Molecular Biology Services) and EZ-96 DNA Methylation-Gold Mag Prep kit (Zymo Research) for bisulfite conversion. Bisulfate-seq libraries were sequenced on the Illumina NovaSeq 6000 machine with the 100bp paired-end mode.

DNA methylation analysis

Bisulfite sequencing data was used to call methylation in 3 different contexts (CG, CHG and CHH where H stands for A, C or T) using a method described in (Pisupati et al., 2022). The methylation level per locus for each context was determined by dividing the number of methylated reads by the total number of reads covering the cytosines in the CG, CHG or CHH context. Thus, the values of methylation of the locus were ranging from 0 to 1 and roughly correspond to the ratio of methylated to total cytosines in the locus (we did not take the average of the ratios for each cytosine to avoid high error rates caused by low read coverage).

Small RNA sequencing and analysis

RNA was isolated from frozen and ground flower samples using NucleoSpin miRNA kit from Macherey-Nagel following the manufacturer's protocol. SmallRNA-seq libraries were prepared using the QIAseq miRNA Library Kit from Qiagen. Raw fastq files were trimmed with cutadapt (v.1.18) (Martin, 2011) using -a AACTGTAGGCACCATCAAT and --minimum-length 18 options. Trimmed reads were aligned to the TAIR10 genome using STAR (v.2.9.6) adjusted for smallRNA-seq, allowing 10 multimappers and 2 mismatches. 24nt-long reads were extracted and read coverage for each basepair of the genome was calculated using genomeCoverageBed (bedtools v.2.27.1) and normalized by dividing by the unique number of reads in the sample. Final smallRNA coverage was calculated by mapping the normalized read coverage per base over the

loci of interest and calculating the average coverage across all base pairs of each locus. Raw smallRNA-seq data from (Papareddy et al., 2020) was processed using the same pipeline.

TE-piece analysis

To find TE pieces in various loci we blasted 31,189 TAIR10 annotated TE sequences onto each locus using blastn (blast+ v2.8.1) with options -word_size 10 -strand both -evalue 1e-7. We required >80% sequence identity and did not restrict the length. We then merged same-strand overlapping TE pieces into TE patches. For all of our analyses we grouped TE families into 7 superfamilies: DNA_other, DNA_MuDR, SINE_LINE, RC_Helitron, LTR_Gypsy, LTR_Copia, Unassigned_NA.

Copy number analysis

Copy number was estimated by extracting the sequence of the locus from the TAIR10 genome and blasting it back to the TAIR10 genome using blastn (blast+ v2.8.1) with options -word_size 10 -strand both -outfmt 7 -evalue $1e^{-7}$. We allowed for copies to be disrupted by insertions of no more than 1.5kb and applied a cutoff of >80% on sequence similarity and >80% on length to all regions identified by blastn.

lincRNA reexpression analysis

Raw RNA-seq data from the silencing mutants from (Osakabe et al., 2021; He et al., 2022; Nguyen et al., 2023) was processed using the same RNA-seq processing pipeline as described above. The re-expression in the mutants was generally defined as the lack of expression in the wildtype (TPM<0.5, averaged from all available replicates), the presence of expression in the mutant (TPM>0.5), and additionally – a 3-fold difference between the expression in the mutant and the wildtype (MUT>3*WT). For the ddm1 knockout in stem cells, the mock ddm1 mutant sample was matched with the mock WT and the heat-treated ddm1 mutant was matched with the heat-treated WT sample (heat treatment as described in (Nguyen et al., 2023)). For the methylation mutants, ddcc and met1-9 mutants were 2 weeks old and matched to the 2-week-old WT control, and the mddcc mutant was matched to the 5 week old WT control. The *met1-9* mutant corresponds to the MET1 knockout with loss of CG methylation, *ddcc* mutant corresponds to the quadruple mutant

for DRM1, DRM2, CMT2, and CMT3 with a loss of CHG/CHH methylation and mddcc mutant is a quantiple mutant for MET1, DRM1, DRM2, CMT2, and CMT3 with a nearly full loss of any methylation (He et al., 2022).

Use of public datasets

The summary of the public datasets used in our study and the corresponding mapping statistics are available in the Suppl. Table S2, S3 and S10. The public datasets were downloaded from NCBI GEO using the below specified GEO accession numbers:

1. RNA-seq and bisulfite-seq from mature leaves of 14-leaf rosettes from the 1001 Genomes project (Kawakatsu et al., 2016): **GSE80744** and **GSE43857**.
2. RNA-seq data from Col-0 seedlings from 12 time points with 14 technical replicates each (Cortijo et al., 2019): GEO accession number **GSE115583**
3. Early embryo and flower bud sRNA-seq data from NRPD1 knockouts (Papareddy et al., 2020): **GSE152971**.
4. Rosette RNA-seq data from DDM1 knockouts (Osakabe et al., 2021): **GSE150436**.
5. Stem cell RNA-seq data from DDM1 knockouts with and without heat stress (Nguyen et al., 2023): **GSE223915**
6. Rosette RNA-seq data from DNA-methylation-free knockouts (He et al., 2022): **GSE169497**.

Availability of data and materials

The sequencing data produced in this study is available at the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) as SuperSeries GSE224761. The gene annotations and the 14-leaf rosette RNA-seq dataset from 28 *A. thaliana* accessions are available under the accession number GSE224760 (note that gene annotations are also available in the supplement as Supplemental_Annotations.xlsx), and the corresponding bisulfite sequencing data – under the GEO accession number GSE226560. The 14-leaf rosette ChIP-seq data from 14 accessions is available under accession number GSE226682. The RNA-seq dataset from seedlings, 9-leaf rosettes, flowers, and pollen from 25 to 27 accessions is available under the GEO accession number GSE226691. The flower sRNA-seq data from 14 A.thaliana accessions is available under the GEO

accession number GSE224571. The code used for the analyses as well as the full ChIP protocol are available at our GitHub repository (Kornienko).

Acknowledgements

We would like to thank the Vienna Biocenter Core Facilities GmbH (VBCF) Next Generation Sequencing for NGS services, VBCF Plant Sciences for the excellent growth chambers, IMBA-IMP-GMI Molecular Biology Services for providing the access to instruments, molecular biology reagents and support, and the VBC Ethics, Health and Safety team for their support during the COVID pandemic. We would like to thank Mirjam Bissmeier for help with preliminary ChIP-seq analyses, and Bhagyshree Jamge, Vu Nguyen, Ramesh Yelagandula, Zdravko Lorkovic, Nathalie Durut, and Ortrun Mittelsten Scheid for their advice with experiments and data, and fruitful discussions. We would like to greatly thank Detlef Weigel for helping to secure funding for the "1001 Genomes Plus" project and for his helpful comments on the manuscript.

Funding

This study was funded by a Hertha Firnberg Postdoctoral Fellowship by the Austrian Science Fund FWF (Project: T-1018 "Role of long non-coding RNA variation in Arabidopsis thaliana") and the ERA-CAPS grant (Project: "1001 Genomes Plus").

Author information

Authors and Affiliations

Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter, Dr. Bohr-Gasse 3, 1030, Vienna, Austria: Aleksandra E. Kornienko, Viktoria Nizhinska, Almudena Molla Morales, Rahul Pisupati, Magnus Nordborg

Contributions

A.E.K. and M.N. designed the study. A.E.K. performed most of the experiments and data analyses. A.M.M. helped with sample collection. R.P. performed the bisulfite-seq data processing. V.N. prepared all the libraries for sequencing. A.E.K. and M.N. wrote the paper.

Corresponding authors

Aleksandra E. Kornienko, Magnus Nordborg

Ethics declaration

Ethics approval and consent to participate: Not applicable

Competing interests: The authors declare they have no competing interests.

Supplemental files

Supplemental file 1: Supplemental_Figures.pdf: Supplemental Figures S1 - S38

Supplemental File 2: Supplemental_Tables.xlsx: Suppl. Tables S1 - S10

Supplemental File 3: Supplemental_Annotations.xlsx: Annotations created in this study

Supplemental Information

Figure S1. Cumulative transcriptome annotation pipeline.

Figure S2. Transcriptome annotation supplement.

Figure S3. Antisense lncRNAs supplement.

Figure S4. Genomic distribution of annotated loci.

Figure S5. Gene identification saturation analysis: accessions.

Figure S6. Gene identification saturation analysis: tissues and accessions.

Figure S7. Expression vs. gene presence population frequency.

Figure S8. Inter-accession expression variation in different tissues.

Figure S9. Expression variability controls: absolute expression level and gene length.

Figure S10. Intra-accession expression variability.

Figure S11. Histone mark profiling.

Figure S12. DNA methylation levels supplement.

Figure S13. DNA methylation vs distance to the centromere.

Figure S14. Heterochromatic histone marks vs distance to the centromere.

Figure S15. Histone modifications of silent and expressed genes: supplement.

Figure S16. Histone modifications of silent and expressed genes: supplement.

Figure S17. 24nt sRNA coverage in early embryo and leaves.

Figure S18. Epigenetic patterns in non-reference accessions.

Figure S19. Epigenetic variation supplement 1.

Figure S20. Epigenetic variation supplement 2.

Figure S21. TE patches in lincRNAs.

Figure S22. Sense and antisense TE pieces in lincRNAs: content, family and position.

Figure S23. Expression variation vs TE content: expression control.

Figure S24. lincRNA methylation variation vs. TE content: supplement.

Figure S25. TE pieces affect epigenetics when controlled for chromosomal location.

Figure S26. Copy number supplement.

Figure S27. Copy number affects lincRNA epigenetic pattern: supplement.

Figure S28. H3K27me3 and H3K9me2 dichotomy.

Figure S29. Loss of 24nt targeting of lincRNA loci in nrpd1a KO flowers.

Figure S30. TE-silencing knockouts supplement.

Figure S31. The scope of lincRNA expression potential in Col-0.

Figure S32. Genomic position and epigenetics of lincRNAs with pieces of Class I and II TEs.

Figure S33. Spreading of silencing from TE patches.

Figure S34. Variability of the number of TE genes and lincRNAs expressed.

Figure S35. GWAS on expressed TE gene number.

Figure S36. Antisense lncRNA candidates: strong correlation with partner PC genes.

Figure S37. TE pieces affect expression as epigenetic patterns of AS lncRNAs and PC genes.

Figure S38. High expression variability despite similar expressed gene numbers.

Suppl. Table S1: Accessions overview.

Suppl. Table S2: RNA-seq data overview.

Suppl. Table S3: RNA-seq data stats.

Suppl. Table S4: Transcriptome assemblies overview.

Suppl. Table S5: ChIP-seq summary and stats.

Suppl. Table S6: bisulfite-seq samples summary and read number.

Suppl. Table S7: smallRNA-seq summary and stats.

Suppl. Table S8: lincRNAs reexpressed upon TE-silencing knockouts.

Suppl. Table S9: AS lncRNA candidates anticorrelating with partner PC gene.

Suppl. Table S10: Public RNA-seq datasets summary and stats.

References

1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at and 1001 Genomes Consortium (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell **166**: 481–491.

Andergassen, D. et al. (2017). Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. Elife **6**.

Athie, A. et al. (2020). Analysis of copy number alterations reveals the lncRNA ALAL-1 as a regulator of lung cancer immune evasion. J. Cell Biol. **219**.

Batista, P.J. and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell **152**: 1298–1307.

Bewick, A.J. and Schmitz, R.J. (2017). Gene body DNA methylation in plants. Curr. Opin. Plant Biol. **36**: 103–110.

Budak, H., Kaya, S.B., and Cagirici, H.B. (2020). Long Non-coding RNA in Plants in the Era of Reference Sequences. Front. Plant Sci. **11**: 276.

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.

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.

Chen, W., Zhu, T., Shi, Y., Chen, Y., Li, W.J., Chan, R.J., Chen, D., Zhang, W., Yuan, Y.A., Wang, X., and Sun, B. (2023). An antisense intragenic lncRNA SEAIRa mediates transcriptional and epigenetic repression of SERRATE in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **120**: e2216062120.

Choi, J., Lyons, D.B., Kim, M.Y., Moore, J.D., and Zilberman, D. (2020). DNA Methylation and Histone H1 Jointly Repress Transposable Elements and Aberrant Intragenic Transcripts. Mol. Cell **77**: 310–323.e7.

Colognori, D., Sunwoo, H., Wang, D., Wang, C.-Y., and Lee, J.T. (2020). Xist Repeats A and B Account for Two Distinct Phases of X Inactivation Establishment. Dev. Cell **54**: 21–32.e5.

Cortijo, S., Aydin, Z., Ahnert, S., and Locke, J.C. (2019). Widespread inter-individual gene expression variability in Arabidopsis thaliana. Mol. Syst. Biol. **15**: e8591.

Csorba, T., Questa, J.I., Sun, Q., and Dean, C. (2014). Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. Proc. Natl. Acad. Sci. U. S. A. **111**: 16160–16165.

Déléris, A., Berger, F., and Duharcourt, S. (2021). Role of Polycomb in the control of transposable elements. Trends Genet. **37**: 882–889.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.,

and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21.

Fedak, H., Palusinska, M., Krzyczmonik, K., Brzezniak, L., Yatusevich, R., Pietras, Z., Kaczanowski, S., and Swiezewski, S. (2016). Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript. Proc. Natl. Acad. Sci. U. S. A. **113**: E7846–E7855.

Feng, S. and Jacobsen, S.E. (2011). Epigenetic modifications in plants: an evolutionary perspective. Curr. Opin. Plant Biol. **14**: 179–186.

Fultz, D., Choudury, S.G., and Slotkin, R.K. (2015). Silencing of active transposable elements in plants. Curr. Opin. Plant Biol. **27**: 67–76.

Hansen, K.H., Bracken, A.P., Pasini, D., Dietrich, N., Gehani, S.S., Monrad, A., Rappsilber, J., Lerdrup, M., and Helin, K. (2008). A model for transmission of the H3K27me3 epigenetic mark. Nat. Cell Biol. **10**: 1291–1300.

He, L., Huang, H., Bradai, M., Zhao, C., You, Y., Ma, J., Zhao, L., Lozano-Durán, R., and Zhu, J.-K. (2022). DNA methylation-free Arabidopsis reveals crucial roles of DNA methylation in regulating gene expression and development. Nat. Commun. **13**: 1335.

Johnson, R. and Guigó, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA **20**: 959–976.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K.V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim. Biophys. Acta **1840**: 1063–1071.

Kang, C. and Liu, Z. (2015). Global identification and analysis of long non-coding RNAs in diploid strawberry Fragaria vesca during flower and fruit development. BMC Genomics **16**: 815.

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.

Kapusta, A. and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. **30**: 439–452.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. **9**: e1003470.

Kawakatsu, T. et al. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell **166**: 492–505.

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

Kindgren, P., Ivanov, M., and Marquardt, S. (2020). Native elongation transcript sequencing reveals temperature dependent dynamics of nascent RNAPII transcription in Arabidopsis. Nucleic Acids Res. **48**: 2332–2347.

Kornienko, A. Kornienko_et_al_lncRNA_expression_variation_and_silencing (Github).

Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics,

R., Pauler, F.M., and Barlow, D.P. (2016). Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biol. **17**: 14.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. **8**: e1002841.

Leone, S. and Santoro, R. (2016). Challenges in the analysis of long noncoding RNA functionality. FEBS Lett. **590**: 2342–2353.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**: 923–930.

Li, L. et al. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. **15**: R40.

Li, S., Liberman, L.M., Mukherjee, N., Benfey, P.N., and Ohler, U. (2013). Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. Genome Res. **23**: 1730–1739.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., 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, X., Hao, L., Li, D., Zhu, L., and Hu, S. (2015). Long non-coding RNAs and their biological roles in plants. Genomics Proteomics Bioinformatics **13**: 137–147.

Lubelsky, Y. and Ulitsky, I. (2018). Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature **555**: 107–111.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10–12.

Mattick, J.S. et al. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat. Rev. Mol. Cell Biol.

Mattick, J.S. and Rinn, J.L. (2015). Discovery and annotation of long noncoding RNAs. Nat. Struct. Mol. Biol. **22**: 5–7.

Matzke, M.A. and Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat. Rev. Genet. **15**: 394–408.

Melé, M. et al. (2015). The human transcriptome across tissues and individuals. Science **348**: 660–665.

Mhiri, C., Borges, F., and Grandbastien, M.-A. (2022). Specificities and Dynamics of Transposable Elements in Land Plants. Biology **11**.

Necsulea, A. and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. Nature Reviews Genetics **15**: 734–748.

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. Front. Genet. **8**: 52.

Nguyen, V.H., Scheid, O.M., and Gutzat, R. (2023). Heat stress response and transposon control in plant shoot stem cells. bioRxiv: 2023.02.24.529891.

Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., and Tanisaka, T. (2008). A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. Genes Genet. Syst. **83**: 321–329.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell **120**: 613–622.

Osakabe, A., Jamge, B., Axelsson, E., Montgomery, S.A., Akimcheva, S., Kuehn, A.L., Pisupati, R., Lorković, Z.J., Yelagandula, R., Kakutani, T., and Berger, F. (2021). The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W. Nat. Cell Biol. **23**: 391–400.

Papareddy, R.K., Páldi, K., Paulraj, S., Kao, P., Lutzmayer, S., and Nodine, M.D. (2020). Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis. Genome Biol. **21**: 251.

Paytuví Gallart, A., Hermoso Pulido, A., Anzar Martínez de Lagrán, I., Sanseverino, W., and Aiese Cigliano, R. (2016). GREENC: a Wiki-based database of plant lncRNAs. Nucleic Acids Res. **44**: D1161–6.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33**: 290–295.

Pisupati, R., Nizhynska, V., Morales, A.M., and Nordborg, M. (2022). On the Causes of Gene-Body Methylation Variation in Arabidopsis thaliana. bioRxiv: 2022.12.04.519028.

Quesneville, H. (2020). Twenty years of transposable element analysis in the Arabidopsis thaliana genome. Mob. DNA **11**: 28.

Rajpal, V.R., Rathore, P., Mehta, S., Wadhwa, N., Yadav, P., Berry, E., Goel, S., Bhat, V., and Raina, S.N. (2022). Epigenetic variation: A major player in facilitating plant fitness under changing environmental conditions. Front Cell Dev Biol **10**: 1020958.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. **44**: W160–5.

Rinn, J.L. and Chang, H.Y. (2020). Long Noncoding RNAs: Molecular Modalities to Organismal Functions. Annu. Rev. Biochem. **89**: 283–308.

Sasaki, E., Kawakatsu, T., Ecker, J.R., and Nordborg, M. (2019). Common alleles of CMT2 and NRPE1 are major determinants of CHH methylation variation in Arabidopsis thaliana. PLOS Genetics **15**: e1008492.

Sauvageau, M. et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife **2**: e01749.

Sigman, M.J. and Slotkin, R.K. (2016). The First Rule of Plant Transposable Element Silencing:

Location, Location, Location. Plant Cell **28**: 304–313.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell **136**: 461–472.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. **22**: 96–118.

Sundaram, V. and Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. **375**: 20190347.

TAIR10 annotation

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–515.

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. **47**: D135–D139.

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. Nat. Rev. Drug Discov. **12**: 433–446.

Wang, J., Meng, X., Dobrovolskaya, O.B., Orlov, Y.L., and Chen, M. (2017). Non-coding RNAs and Their Roles in Stress Response in Plants. Genomics Proteomics Bioinformatics **15**: 301–312.

Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bähr, M., Wolf, S., Shendure, J., Eils, R., Plass, C., and Weichenhan, D. (2013). Tagmentation-based whole-genome bisulfite sequencing. Nat. Protoc. **8**: 2022–2032.

Wapinski, O. and Chang, H.Y. (2011). Long noncoding RNAs and human disease. Trends in Cell Biology **21**: 354–361.

Whittaker, C. and Dean, C. (2017). The FLC Locus: A Platform for Discoveries in Epigenetics and Adaptation. Annu. Rev. Cell Dev. Biol. **33**: 555–575.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. **42**: D98–103.

Xin, M., Wang, Y., Yao, Y., Song, N., Hu, Z., Qin, D., Xie, C., Peng, H., Ni, Z., and Sun, Q. (2011). Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. BMC Plant Biol. **11**: 61.

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

Xu, J. et al. (2019). Population-level analysis reveals the widespread occurrence and phenotypic

consequence of DNA methylation variation not tagged by genetic variation in maize. Genome Biol. **20**: 243.

Xu, Y. et al. (2020). Identification and comprehensive characterization of lncRNAs with copy number variations and their driving transcriptional perturbed subpathways reveal functional significance for cancer. Brief. Bioinform. **21**: 2153–2166.

Xu, Y.-C., Zhang, J., Zhang, D.-Y., Nan, Y.-H., Ge, S., and Guo, Y.-L. (2021). Identification of long noncoding natural antisense transcripts (lncNATs) correlated with drought stress response in wild rice (Oryza nivara). BMC Genomics **22**: 424.

Yang, P., Wang, Y., and Macfarlan, T.S. (2017). The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. Trends Genet. **33**: 871–881.

Yelagandula, R. et al. (2014). The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. Cell **158**: 98–109.

Yuan, C., Wang, J., Harrison, A.P., Meng, X., Chen, D., and Chen, M. (2015). Genome-wide view of natural antisense transcripts in Arabidopsis thaliana. DNA Res. **22**: 233–243.

Yuan, J., Zhang, Y., Dong, J., Sun, Y., Lim, B.L., Liu, D., and Lu, Z.J. (2016). Systematic characterization of novel lncRNAs responding to phosphate starvation in Arabidopsis thaliana. BMC Genomics **17**: 655.

Zemach, A., Kim, M.Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell **153**: 193–205.

Zhao, L., Zhou, Q., He, L., Deng, L., Lozano-Duran, R., Li, G., and Zhu, J.-K. (2022). DNA methylation underpins the epigenomic landscape regulating genome transcription in Arabidopsis. Genome Biol. **23**: 197.

Zhu, Y., Chen, L., Hong, X., Shi, H., and Li, X. (2022). Revealing the novel complexity of plant long non-coding RNA by strand-specific and whole transcriptome sequencing for evolutionarily representative plant species. BMC Genomics **23**: 381.

Figure legends

Fig. 1. Mapping lncRNA transcription in hundreds of accessions and several tissues reveals thousands of novel lncRNAs.

A. Origins of the *A. thaliana* accessions used for transcriptome annotation and an example photograph of 6 different accessions in the growth chamber. **B.** Overview of the pipeline used for cumulative transcriptome annotation. Tissues from left to right: seedlings, rosette, flowers, pollen. **C.** The distribution of types of loci in the cumulative annotation. **D.** The distribution of lncRNA positional classes. **E.** An example of a novel intergenic lncRNA on chromosome 1. Expression in 7 different *A. thaliana* accessions is shown. **F.** The number of lncRNA and PC loci identified as a function of the number of accessions used, relative to the number identified using 460 accessions. Random subsampling of accessions was performed in 8 replicates and the error bars indicate the standard deviation across replicates.

Fig. 2. lncRNAs display extensive natural expression variability and appear to be largely silent.

A. The fraction of accessions in the 1001 Genomes dataset (Kawakatsu et al., 2016) where the gene is expressed (TPM > 0.5). Only genes that are expressed in at least one accession are plotted. **B.** Coefficient of variance of expression in 461 accessions from 1001 Genomes dataset (Kawakatsu et al., 2016). Only genes with TPM > 1 in at least one accession are plotted. **C.** Expression noise calculated from 14 technical replicates of Col-0 seedlings; expression noise value averaged across 12 samples is displayed (Cortijo et al., 2019). Only genes with TPM > 1 in at least one sample are plotted. Boxplots: Outliers are not shown, and p-values were calculated using Mann-Whitney test on equalized sample sizes: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. **D.** Gene expression levels for different types of genes in 4 tissues for the reference accession Col-0 (6909) and 2 randomly picked accessions. Heatmaps built using “pheatmap” in R with scaling by row. Only genes expressed in at least one sample are plotted. Clustering trees for rows not shown. **E.** Average number of genes expressed in an accession and its randomly selected partner accession from the 1001G dataset, and the number of genes expressed (TPM > 0.5) in both accessions. Percentages indicate the overlap between accessions. **F.** The proportion of genes expressed in one accession in

seedlings, 9-leaf rosettes, flowers, pollen, or all 4 tissues combined (dark bars). The error bars show standard deviation between 23 accessions. The light part of the bars displays the additional proportion of genes that can be detected as expressed when all 23 accessions are considered.

Fig. 3. Epigenetic patterns of lncRNAs in *A. thaliana* indicate ubiquitous silencing.

A. Averaged profiles of the input-normalized ChIP-seq signal for H1, H3K9me2, H3K36me3, H3K4me3 and H3K27me3 over 4 gene types from our cumulative transcriptome annotation. The plots show data from Col-0 rosettes, replicate 2. All genes, expressed and silent in Col-0, are used for the analysis. Profiles were built using plotProfile from deeptools (Ramírez et al., 2016). **B.** H3K9me2, H1 and H3K36me3 histone modifications in Col-0 rosette. The log2 of the gene-body coverage normalized by input and averaged between 2 replicates is plotted. **C.** Left: CG and CHH DNA methylation levels in Col-0 rosette. Right: density of CG methylation level for PC gene, lincRNA and TE gene loci. Methylation level is calculated as the ratio between the number of methylated and unmethylated reads over all Cs in the respective context (CG/CHH) in the gene body and averaged over 4 replicates. **D.** The scheme of the experiment: same tissue was used for RNA-seq, ChIP-seq and Bisulfite-seq in this study. **E.** H3K9me2 input-normalized coverage separately for expressed (ON, TPM>0.5) and silent genes (OFF, TPM<0.5). **F.** H3K27me3 normalized coverage separately for expressed (ON, TPM>0.5) and silent (OFF, TPM<0.5) genes. **G.** Methylation levels for expressed (ON, TPM>0.5) and silent (OFF, TPM<0.5) genes. H. **I.** Coverage of 24nt small RNAs in the gene body, calculated as the number of 24nt reads mapping to the locus divided by the total number of reads and the locus length. **J.** Coverage of 24nt small RNA separately for expressed (ON, TPM>0.5) and silent (OFF, TPM<0.5) genes. P-values were calculated using Mann-Whitney test on equalized sample sizes: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not shown.

Fig. 4. LncRNAs display increased epigenetic variation that explains expression variation of many lncRNAs.

A. Standard deviation of input and quantile normalized coverage (see Methods) of H3K9me2 (left) and H3K27me3 (right) in rosettes across 13 and 12 accessions respectively. **B.** Standard deviation

of CG (left) and CHH (right) methylation levels across 444 accessions (rosettes, 1001G dataset, (Kawakatsu et al., 2016)). P-values were calculated using Mann-Whitney test on equalized sample sizes: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not shown. **C.** The summary of lncRNAs for which expression can be explained by methylation (Suppl. Fig. S20B). The colored circles show the overlap between loci for AS lncRNAs (green) and lincRNAs (orange) that were found to be defined by CG or CHH methylation level. **D.** An example of a lincRNA defined by CG and CHH methylation. The figure shows RNA-seq signal (forward strand), CG and CHH methylation levels in rosette and the 24nt sRNA signal in flowers in 2 accessions. **E.** The plots show the expression level as a function of CG/CHH methylation for the example lincRNA across 444 accessions (Kawakatsu et al., 2016). The results of the Mann-Whitney tests used for defining the explanatory power of CG/CHH methylation are shown. **F.** Expression in rosettes vs. H3K9me2 level in rosettes of the example lincRNA in 13 accessions. **G.** Expression in rosettes vs. 24nt sRNA coverage in flowers of the example lincRNA in 14 accessions.

Fig. 5. Many lincRNAs contain pieces of TEs that affect their silencing and variation.

A. Outline of TE-content analysis. Top: TAIR10-annotated TEs were blasted to the sequences of lincRNAs (and other loci). Bottom: The mapped pieces of different TEs overlapping in the same direction were merged into “TE patches". The upstream and downstream “borders” of genes were analyzed in the same way. **B.** The fraction of loci containing a TE piece. The intergenic controls for lincRNAs, lincRNA TSS+/-200bp and TES +/-200bp were obtained by shuffling the corresponding loci within intergenic regions (lincRNAs excluded) 3 times and averaging the results. The error bars on controls represent the standard deviation between the 3 shuffling replicates. **C.** The number of TE patches per 1 kb. **D.** Distribution of the length of TE patches (any relative direction) within lincRNA loci. **E.** TE content distribution among lincRNAs. TE patches in any relative direction. The loci with large TE content are those where the TE patches are mapping antisense to the lincRNA locus. **F.** The proportion of TE pieces of different TE families inside different types of loci. **G.** The proportion of expressed lincRNA as a function of their TE-content. The y axis is displayed in log scale. **H,I,J.** Levels of methylation (**H**), H3K9me2 (**I**), and 24nt sRNAs (**J**) for lincRNA loci as a function of TE-content, with TE genes for comparison. **K,L,M.** Expression variability between 461 accessions (Kawakatsu et al., 2016) (**K**), standard

deviation of CG methylation levels across 444 accessions (Kawakatsu et al., 2016) (**L**) and standard deviation of quantile and input normalized H3K9me2 levels in rosettes across 13 accessions (**M**) of lincRNA loci as a function of TE-content, with TE genes for comparison. P-values were calculated using Mann-Whitney tests: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not shown.

Fig. 6. Copy number of lincRNAs affects their epigenetic patterns and variability.

A. The copy number distribution for PC genes, AS lncRNAs, lincRNAs and TE genes from the cumulative transcriptome annotation in the TAIR10 genome. **B.** Top: the scheme of the 4 types of loci, bottom: the distribution of copy number of the 4 types of loci: 1 copy with no TE patch, 1 copy with a TE patch, multiple copies with no TE patch in the original locus and multiple copies with a TE patch in the original locus. **C.** The bar plot shows the proportion of the 4 types of lincRNAs, and 2 types of TE genes expressed (TPM>0.5, green) or silent (TPM<0.5, gray) in Col-0 rosettes. **D-F.** The boxplots show the CG and CHH methylation (**D**), H3K9me2 level (**E**) in Col-0 rosettes and 24nt sRNA coverage in Col-0 flowers (**F**) for the 4 types of lincRNAs and 2 types of TE genes. **G-I.** The boxplots show expression (**G**), CG methylation (**H**) and H3K9me2 (**I**) variability for the 4 types of lincRNAs and 2 types of TE genes. P values in the boxplots are calculated using Mann-Whitney test: ***$p<10^{-10}$, **$p<10^{-5}$, *$p<0.01$, n.s. $p>0.01$. Outliers in the boxplots are not plotted.

Fig. 7. lincRNAs are silenced by PC-like and TE-like mechanisms.

A. H3K27me3 vs. H3K9me2 level on lincRNA loci in Col-0 14-leaf rosettes (average of 2 replicates). K27 genes: red, K27 signal>0, K9 signal<0. K9 genes: blue, K27 signal<0, K9 signal>0. **B.** The distribution of K9 (blue) and K27 (red) genes among the 4 gene types. NA: genes with neither mark (gray, K27 signal<0, K9 signal<0). **C-F.** The boxplots show the relative TE-sequence content (**C**), copy number (**D**), CG (**E**) and CHH methylation level (**F**) of lincRNA loci classified as K27 and K9 genes. Outliers not plotted. P-values calculated using Mann-Whitney tests: ***$p<10^{-10}$. **G.** Relative number of K27 and K9 lincRNAs targeted by 24nt sRNAs (RPM>0.03) in *A. thaliana* embryos ("early heart" stage) (Papareddy et al., 2020). The sRNA coverage is averaged

across 3 replicates. **H.** 24nt sRNA coverage in *A. thaliana* embryos ("early heart" stage) in the wild type (WT, Col-0) and in PolIV-deficient mutants (*nrpd1a*, Col-0 background) (Papareddy et al., 2020). 1,207 lincRNAs that are targeted (RPM>0.03, average of 3 replicates) by 24nt sRNAs in the WT are plotted. **I.** Expression level of the 149 lincRNAs re-expressed upon *ddm1* knockout in Col-0 rosettes (Osakabe et al., 2021). The bars at the bottom show the distribution of K9 (blue) vs. K27 (red) and TE-containing (dark orange) vs. TE-free (light orange) loci among the re-expressed lincRNAs (same for **J** and **K**). **J.** Expression level of the 410 lincRNAs re-expressed in stem cells upon *ddm1* knockout in Col-0 with or without heat stress treatment (Nguyen et al., 2023). **K.** Expression level of lincRNAs re-expressed upon DNA methylases knockouts in Col-0 rosettes (He et al., 2022) (see Methods). Heatmaps built using "pheatmap" in R with scaling by row. No column clustering, row clustering trees not displayed. **L.** An example of a lincRNA epigenetically silenced in Col-0 WT but expressed in the silencing mutants.

Fig. 8. TE pieces appear to attract silencing to lincRNA loci.

A-B. 24nt sRNA level in Col-0 flowers (**A**) and H3K9me2 level in rosettes (**B**) for lincRNAs with pieces of TEs from 4 superfamilies and TAIR10 TE fragments from the same superfamilies. Only lincRNAs with TE pieces from one superfamily are plotted. Light-orange box indicates lincRNAs without TE pieces. **C-D.** The boxplots show CG methylation level (**C**) and 24nt sRNA coverage (**D**) for TE patches inside lincRNAs, TE-patch-free parts of TE-containing lincRNA loci and lincRNA loci without TE patches. Outliers not plotted. P-values calculated using Mann-Whitney tests: ***$p<10^{-10}$, *$p< 0.01$. **E.** The IGV screenshot shows an example of lincRNAs with TE patches that have higher level of CG methylation and 24nt sRNA coverage over TE patches than over the rest of the locus. **F.** The scatter plot shows the number of TE genes expressed in rosettes of 460 different accessions (Kawakatsu et al., 2016) as a function of the number of lincRNAs with TE pieces (left) and without TE pieces (right) expressed in the same accession. Pearson correlation coefficient is displayed. **G.** The scheme summarizes lincRNA silencing pathways. PC-like lincRNAs that show H3K27me3 repressive histone marks are likely silenced by PRC2, while TE-like lincRNAs that display H3K9me2 are silenced by CMT2/DDM1 and RdDM pathways. TE piece presence likely attracts TE silencing and repressive chromatin to the lincRNA locus.