The Plant Cell, Vol. 32: 1845–1867, June 2020, www.plantcell.org © 2020 ASPB.

LARGE-SCALE BIOLOGY ARTICLE

Characterization of *Arabidopsis thaliana* Promoter Bidirectionality and Antisense RNAs by Inactivation of Nuclear RNA Decay Pathways

Axel Thieffry,[a,b] Maria Louisa Vigh,[a] Jette Bornholdt,[a,b] Maxim Ivanov,[c] Peter Brodersen,[a,1] and Albin Sandelin[a,b,1]

[a] Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
[b] Biotech Research and Innovation Centre, University of Copenhagen, DK-2200 Copenhagen N, Denmark
[c] Department of Plant and Environmental Sciences, University of Copenhagen, DK-1871 Frederiksberg C, Denmark

ORCID IDs: 0000-0002-6717-2785 (A.T.); 0000-0001-6271-6520 (M.L.V.); 0000-0002-3494-6721 (J.B.); 0000-0001-7548-3316 (M.I.); 0000-0003-1083-1150 (P.B.); 0000-0002-7109-7378 (A.S.)

In animals, RNA polymerase II initiates transcription bidirectionally from gene promoters to produce pre-mRNAs on the forward strand and promoter upstream transcripts (PROMPTs) on the reverse strand. PROMPTs are degraded by the nuclear exosome. Previous studies based on nascent RNA approaches concluded that Arabidopsis (*Arabidopsis thaliana*) does not produce PROMPTs. Here, we used steady-state RNA sequencing in mutants defective in nuclear RNA decay including the exosome to reassess the existence of Arabidopsis PROMPTs. While they are rare, we identified ~100 cases of exosome-sensitive PROMPTs in Arabidopsis. Such PROMPTs are sources of small interfering RNAs in exosome-deficient mutants, perhaps explaining why plants have evolved mechanisms to suppress PROMPTs. In addition, we found ~200 long, unspliced and exosome-sensitive antisense RNAs that arise from transcription start sites within parts of the genome encoding 3′-untranslated regions on the sense strand. The previously characterized noncoding RNA that regulates expression of the key seed dormancy regulator, *DELAY OF GERMINATION1*, is a typical representative of this class of RNAs. Transcription factor genes are overrepresented among loci with exosome-sensitive antisense RNAs, suggesting a potential for widespread control of gene expression via this class of noncoding RNAs. Lastly, we assess the use of alternative promoters in Arabidopsis and compare the accuracy of existing TSS annotations.

INTRODUCTION

The vast majority of promoters of mammalian protein-coding genes initiates transcription by RNA polymerase II (RNAPII) on both strands, producing a sense pre-mRNA transcript and a shorter, unstable RNA on the reverse strand, denominated the promoter upstream transcript (PROMPT) or upstream antisense RNA (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). Transcription of PROMPTs and pre-mRNAs initiates in opposite directions from two separate core promoters, where PROMPT and pre-mRNA TSSs are located at each edge of the nucleosome-depleted region (NDR; Ntini et al., 2013; Andersson et al., 2014a, 2014b). Unlike pre-mRNAs, PROMPTs are rapidly degraded by the nuclear exosome (Preker et al., 2008), a central player in RNA degradation. The core exosome is a highly conserved 9-subunit complex consisting of a central, hexameric barrel composed of the proteins ribosomal RNA (rRNA)-processing (RRP) 41-43, RRP45-46, and mRNA TRANSPORT REGULATOR3 (MTR3; as a trimer of RRP41-RRP42, RRP45-RRP46, and RRP43-MTR3 heterodimers), and a 3-subunit lid containing the proteins RRP4, RRP40, and CEP1 SYNTHETIC LETHAL protein4 (CSL4). Nuclear 3′-5′ exonucleases associate with the core exosome, and the resulting exosome machinery constitutes the main 3′-5′ exonuclease activity both in the nucleus and in the cytoplasm (Chlebowski et al., 2013; Kilchert et al., 2016). The exosome sensitivity of PROMPTs correlates with the frequency of occurrence of specific DNA sequence elements downstream of the PROMPT transcription start sites (TSSs): Compared to the coding strand, the PROMPT region is enriched in TSS-proximal polyadenylation (poly[A]) sites and depleted of splice donor sites (Almada et al., 2013; Ntini et al., 2013; Core et al., 2014).

The pattern of bidirectional transcription initiation is not exclusive to promoters of coding genes, but appears to be generic for RNAPII initiation in metazoans: It is also detectable at long noncoding RNA (lncRNA) promoters (Andersson et al., 2014b) and in particular at active enhancers (Kim et al., 2010). We and others have previously shown that enhancer regions can be accurately predicted as sites of bidirectional initiation of unstable transcripts on the basis of either RNA sequencing (RNA-seq; Wu et al., 2014), sequencing of capped RNA 5′-ends (Andersson et al., 2014a), or nascent RNA approaches (Core et al., 2014).

[1] Address correspondence to pbrodersen@bio.ku.dk and albin@binf.ku.dk.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Albin Sandelin (albin@binf.ku.dk) and Peter Brodersen (pbrodersen@bio.ku.dk).
www.plantcell.org/cgi/doi/10.1105/tpc.19.00815

IN A NUTSHELL

Background: Initiation of transcription is a vital and highly regulated process. Recent research on animal cells has shown that transcription initiation of most genes is accompanied by divergent non-coding transcription on the reverse strand (bidirectional transcription initiation). These reverse strand RNAs are rapidly degraded, making them hard to detect unless relevant degradation pathways are inactivated. Bidirectional transcription has been observed in mammals, insects and fungi, prompting the hypothesis that this is a universal feature of eukaryotes. Surprisingly, plants may be an exception, since some studies found no evidence of bidirectional transcription initiation in plants. However, these studies relied on methods to capture RNAs as they are transcribed and did not use mutants impaired in degradation of aberrant nuclear RNA, including divergent non-coding transcripts.

Question: We investigated the existence of bidirectional transcription initiation in Arabidopsis using mutants in major nuclear RNA degradation pathways. We employed a combination of transcriptomics approaches to map transcription initiation sites at nucleotide resolution, and to uncover the nature of transcripts that accumulate in these mutant backgrounds.

Findings: Our data on Arabidopsis seedlings confirms that plants do not generally employ bidirectional transcription initiation: only around one hundred Arabidopsis genes exhibit this feature. However, our findings strongly indicate that if allowed to accumulate, divergent non-coding transcripts are processed to become small silencing RNAs. This process would depend on RNA-dependent RNA Polymerase, an enzyme found in plants, many fungi, and some invertebrates, but not in mammals or insects. We hypothesize, therefore, that reverse strand transcription initiation in Arabidopsis is selected against, since the resulting RNAs may cause ectopic silencing, potentially including epigenetic silencing via the plant-specific RNA-directed DNA methylation pathway. In addition, our approach revealed a large set of long non-coding anti-sense RNAs whose transcription initiates in the 3'-end regions of protein-coding genes.

Next steps: Our works raises three important questions. 1) Which mechanisms, probably acting at the transcriptional level, eliminate accumulation of divergent non-coding transcripts? 2) Do small interfering RNAs produced from divergent non-coding transcripts really possess detrimental silencing activity, as proposed? 3) Do these non-coding antisense RNAs, or their transcription, play a role in plant gene regulation?

The similarity of transcription initiation patterns between different RNA classes has spurred the hypothesis that transcription by RNAPII is subject to the same rules irrespective of the RNA that is produced, so that bidirectionally transcribed NDRs constitute a generic transcription initiation block, and the fate and function of RNAs produced are determined after transcription initiation (Andersson et al., 2015b; Chen et al., 2017; also reviewed by Andersson and Sandelin, 2020). It is currently an open question whether this framework is shared between eukaryotes. The majority of enhancers and a substantial fraction of gene promoters were recently shown to be divergently transcribed in *Drosophila melanogaster* S2 cells (Rennie et al., 2018) and larvae (Meers et al., 2018). In yeast species, the number of enhancers is small due to their compact genomes. Promoter bidirectionality was observed in budding yeast (*Saccharomyces cerevisiae*; reviewed by Jensen et al., 2013), and in at least a subset of promoters in fission yeast (*Schizosaccharomyces pombe*; Shetty et al., 2017; Thodberg et al., 2019a).

In flowering plants, only a handful of studies on a limited number of species has investigated this question in depth, using approaches in which nascent RNA is either labeled and affinity-purified from isolated nuclei (global run-on sequencing, GRO-seq; Core et al., 2008; and global run-on sequencing of capped RNAs [GRO-cap]), or isolated as RNAPII-associated RNA by immunoprecipitation (native elongating transcript sequencing, NET-seq). Analysis of Arabidopsis (*Arabidopsis thaliana*) seedlings by GRO-seq provided little support for divergent transcription from promoters of protein-coding genes (Hetzel et al., 2016). Likewise, analysis of average GRO-seq and NET-seq signals around TSSs also led Zhu et al. (2018) to conclude that such promoters are unidirectional in Arabidopsis. One study in maize (*Zea mays*) also found little evidence for bidirectional transcription using GRO-seq (Erhard et al., 2015). However, a precision nuclear run-on sequencing analysis of cassava (*Manihot esculenta*) and a re-analysis of the maize GRO-seq data generated by Erhard et al. (2015) showed many instances of clear bidirectional transcription at intergenic loci that may be candidate enhancers (Lozano et al., 2018). In addition, the same study identified some cases of bidirectional transcription from promoters of protein-coding genes in cassava (Lozano et al., 2018). Although both GRO-seq and NET-seq data clearly identify RNA populations distinct from the steady-state pool, several concerns may be raised on negative findings based on these techniques. First, GRO-seq can have limited power to detect weakly transcribed loci such as PROMPTs if the sequencing depth is low (Andersson et al., 2015a), and indeed, many of the published plant GRO-seq studies suffer from

a lack of sample replicates. Second, RNA species may escape detection if their degradation is rapid and occurs cotranscriptionally, even if the techniques successfully capture many other species of nascent RNA. For these reasons, it is not completely clear whether plant promoters are truly unidirectional, in particular because of conflicting conclusions reached in maize (Erhard et al., 2015; Lozano et al., 2018). The extent of bidirectional transcription in genomic regions other than gene promoters is similarly a question worthy of further investigation.

Although nascent RNA techniques also support bidirectional transcription at gene promoters in mammals (Andersson et al., 2015a), one of the most instrumental techniques in uncovering this aspect of RNAPII function has been the analysis of steady-state RNA in reference cells compared with cells in which components of the nuclear RNA exosome machinery were depleted (Preker et al., 2008; Ntini et al., 2013). In Arabidopsis, loss of function mutants of the core exosome subunits *RRP4* and *RRP41* are not viable. In contrast, null alleles of the third lid subunit *CSL4* displays surprisingly mild, if any, growth phenotypes, and few molecular phenotypes (Chekanova et al., 2007). These findings initially inspired Chekanova et al. (2007) to complete a thorough genome-wide survey of RNA exosome substrates in Arabidopsis using inducible RNA interference lines targeting *RRP4* and *RRP41* in combination with whole-genome tiling arrays. This pioneering study did notice that, upon exosome depletion, a group of polyadenylated RNA species referred to as upstream noncoding transcripts (UNTs) exhibited a conspicuous accumulation around the 5′-ends of a subset of protein-coding genes. However, the poorly resolved TSS locations available at the time made it unclear to what extent UNTs generally overlap with 5′-ends of coding transcripts. Indeed, whereas PROMPTs in mammals were originally also found using comparative hybridization to whole-genome tiling arrays and appeared to originate from both forward and reverse strands, the identification of exosome-sensitive UNTs in Arabidopsis has not been followed up by more powerful RNA-seq analyses. Thus, a comparative Arabidopsis study using steady-state RNA-seq and RNA 5′-tag sequencing approaches on wild type and mutants defective in nuclear RNA decay components would be a useful complement to nascent RNA studies, not only because of the potential limitations of these techniques, but also because of the original identification of UNTs upon exosome knockdown. Moreover, because of the much higher sensitivity and precision of modern-day sequencing approaches compared with tiling arrays used earlier (Chekanova et al., 2007), such approaches may also inform on other classes of cryptic transcripts in Arabidopsis, and on the nuclear RNA degradation pathways—including the exosome—involved in their rapid turnover.

Exosomal RNA decay in vivo depends crucially on adaptor proteins of the DEAD-box RNA helicase family that funnel RNA substrates to the exosome (Lykke-Andersen et al., 2009). Plants encode at least three such adaptors: Similar to yeast and mammals, SUPERKILLER2 (SKI2) is required for exosomal degradation of cytoplasmic RNA substrates, including endonucleolytic mRNA cleavage fragments (Branscheid et al., 2015). However, in contrast to *S. cerevisiae* and mammals, two different nuclear DEAD-box helicase adaptors are encoded in plant genomes: MTR4 localizes to the nucleolus and has specific functions in rRNA processing (Lange et al., 2011), while HUA ENHANCER2 (HEN2) localizes to the nucleoplasm and is required for the degradation of a wide range of RNA substrates, including many noncoding RNAs (Lange et al., 2014). In addition to the RNA exosome, less well-characterized nuclear 5′-3′ exonucleolytic pathways exist. These pathways may involve the nuclear 5′-3′ exonucleases EXORIBONUCLEASE2 (XRN2) and XRN3. XRN2 has documented functions in rRNA cleavage (Zakrzewska-Placzek et al., 2010), while XRN3 has more identified nuclear substrates, including RNAPII-associated RNA and cleavage products by DICER-LIKE enzymes, the former giving rise to defects in transcription termination in *xrn3* mutants (Krzyszton et al., 2018). In addition, the heptameric nuclear LIKE Smith (Lsm) protein complex, Lsm2-8, a well-established pre-mRNA splicing factor (Bouveret et al., 2000; Tharun et al., 2000), also plays roles in nuclear RNA decapping and, ultimately, decay in Arabidopsis, presumably via 5′-3′ exonucleolysis (Golisz et al., 2013). In yeast, mammals, and plants, Lsm2-8 shares six of its subunits, Lsm2 to Lsm7, with the cytoplasmic decapping activator Lsm1-7, such that Lsm8 is the only subunit specific to the nuclear Lsm complex (Perea-Resa et al., 2012).

In this study, we used a null allele in *hen2*, a hypomorphic loss-of-function mutant in the core exosome subunit *RRP4*, and an *lsm8* null allele for transcriptome-wide studies. We applied transcriptome-wide sequencing of capped 5′ ends of RNAs (CAGE; Takahashi et al., 2012), small RNA-seq, and RNA-seq using RNA isolated from mutant and wild-type seedlings to identify transcripts subject to preferential nuclear degradation. In agreement with the results of nascent RNA analyses, we found that the majority of promoters of coding genes only have CAGE signal on the sense strand, even when the exosome complex or its associated helicase system was inactivated. However, unambiguous exceptions to these observations exist: We found 96 mRNA TSSs with robust PROMPT-like configurations that produce exosome-sensitive transcripts, supported by CAGE and RNA-seq, and, in some cases, by earlier GRO-seq datasets. Interestingly, our small RNA-seq data showed that regions with detectable PROMPTs in exosome mutants became sources of 21- to 22-nucleotide small interfering RNAs (siRNAs), pointing to the potential danger of widespread PROMPT production in plants. We also found 113 noncoding regions featuring bidirectional transcription of exosome-sensitive RNAs reminiscent of active enhancers in vertebrates. While exosome mutants did not reveal widespread bidirectional transcription initiation as in mammalian genomes, a striking feature of the transcriptional output in Arabidopsis not present to the same degree in mammalian systems was a frequent occurrence of exosome-sensitive RNAs transcribed antisense to genes. Such transcripts commonly initiated in 3′-untranslated regions (UTR) of cognate genes were ~1,000 nucleotides long, typically unspliced, and, in many cases, occurred in genes encoding transcription factors (TFs). In addition, we show that the set of active TSSs in wild-type plants presented here has higher accuracy than current TSS annotations such as The Arabidopsis Information Resource (TAIR10; Berardini et al., 2015) and, in particular, ARAPORT11 (Cheng et al., 2017). We also demonstrate the prevalence of alternative TSSs in the wild-type transcriptome. Our data, therefore, constitute an essential resource for gene regulation and RNA degradation research in Arabidopsis.

RESULTS

Generation of a Comprehensive Map of Accurate Arabidopsis TSSs

Investigations of bidirectional transcription are crucially dependent on highly accurate measurements of TSS locations and activity. As a baseline for further analysis, we prepared CAGE libraries from RNA purified from 14-d–old Arabidopsis Columbia (Col-0) wild-type seedlings in biological triplicates (Figure 1A). CAGE reads were mapped to the TAIR10 genome with an average of 20.8 million (83.4% of total) uniquely mapped reads per replicate. As described previously, gene TSSs in complex genomes, including Arabidopsis (Tokizawa et al., 2017), are often locally dispersed, a phenomenon often referred to as "broad promoters" (Carninci et al., 2006). It is therefore helpful to cluster CAGE tags with closely spaced 5′-ends on the same strand into CAGE tag clusters (TCs). We created TCs across all libraries and calculated the normalized expression of each library as tags per million mapped reads (TPMs). Because a TC may be composed of multiple adjacent TSSs, we will use the term TC for CAGE initiation events rather than just TSS, which we will use for annotated 5′-ends and data from other methods. Figure 1B shows an example of CAGE tags identifying the annotated TSS for the gene encoding *TRANSMEMBRANE PROTEIN97* (At2g32380).

While CAGE mapping is independent of gene annotation, it is often useful for downstream analysis to assess the overlap of TCs and gene annotation. As reported by Thodberg et al. (2019a, 2019b) and Thodberg and Sandelin (2019), we devised a hierarchical annotation scheme Figure 1C) based on gene models from either TAIR10 or ARAPORT11 and counted the number of wild-type CAGE TCs located in each type of annotated region (Figure 1D, left). As expected, the vast majority of TCs fell into annotated promoter regions ($\pm$100 bp around the 5′-ends of gene models from ARAPORT11 or TAIR10), although ARAPORT11 promoters accounted for a smaller fraction of TCs than TAIR10. All other categories accounted for only a minor fraction of TCs when using TAIR10, but a much larger fraction of TCs mapped to 5′-UTR regions when using ARAPORT11. As observed previously in human, mouse (*Mus musculus*), and *S. pombe* (Bornholdt et al., 2017; Boyd, 2018; Thodberg et al., 2019a), CAGE TCs falling into either promoter or 5′-UTR regions were substantially more expressed than other categories (Figure 1D, right).

For the TAIR10 annotation, the CAGE signal peaked sharply at annotated TSSs. By contrast, for ARAPORT11, the distribution of CAGE signal around annotated TSSs was much broader (Figure 1E), pointing to possible inaccuracies in TSS annotation in ARAPORT11. Indeed, the CAGE TCs that were differentially categorized in ARAPORT11 versus TAIR10 and fell within the annotation category "promoter" in TAIR10 were almost exclusively annotated as 5′-UTR region in ARAPORT11, while CAGE tags falling into the "proximal" region in TAIR10 (–100 to –400 relative to annotated TSS) were frequently annotated as "promoter" in ARAPORT11 (Figure 1F; all categories are defined as in Figure 1C, based on respective gene models). These observations indicate that TAIR10 TSSs are in general agreement with CAGE TCs, while ARAPORT11 TSSs are, on average, located upstream of CAGE TCs/TAIR10 TSSs.

The TSS Annotation in TAIR10 Is Substantially More Accurate than in ARAPORT11

Because TSS accuracy is a crucial feature of any genome annotation, we decided to assess rigorously which of the TSS annotations (TAIR10 or ARAPORT11) was more accurate. To this end, we designed a series of quality control measures, and also considered two previously published genome-wide TSS datasets in our analyses: paired-end analysis of transcription sequencing (PEAT-seq) of 7-d–old seedling roots (Morton et al., 2014), and parallel analysis of RNA 5′-ends (nanoPARE) of flower buds (Schon et al., 2018; see Supplemental Table 1 for details about tissues and growth conditions of the different large-scale experiments used). The majority (62%) of TCs/TSSs was supported by at least two methods, with nanoPARE having the highest fraction of called TCs/TSSs without support from the other methods (25%), compared with CAGE (12%) and PEAT (1%; Figure 2A; see Methods). The substantial overlap is noteworthy, given that the tissues analyzed differ between studies: whole seedlings (CAGE), roots (PEAT-seq), and flower buds (nanoPARE). Notably, the number of TSSs defined by PEAT-seq was much lower than in the other approaches and 95% were supported by CAGE TCs (Figures 2A and 2B). Closer inspection showed that PEAT-seq TSSs had an overall higher expression (as measured by CAGE tags, see "Methods") than CAGE TCs. This indicates that the two methods likely detected the same initiation events, although CAGE was more sensitive and therefore uncovered TSSs of more weakly expressed genes (Figure 2B).

We reasoned that the most accurate TSS annotation set should be best at recapitulating known promoter biology in terms of accessible DNA, sequence characteristics, and data from nascent RNA assays. To make the analyses comparable between sets and focus on TSSs expressed in our samples, we only assessed TCs/TSSs that had a CAGE expression $\geq$1 TPM in at least two replicates in the $\pm$100-bp region around TCs peaks or annotated TSSs (depending on analysis; see below). First, we assessed core promoter patterns by constructing sequence logos (Schneider and Stephens, 1990) from DNA sequences around annotated TSSs from TAIR10 and ARAPORT11, TC peaks from CAGE, and TSS sets from PEAT-seq and nanoPARE (Figure 2C). Sequence logos constructed based on CAGE TCs resulted in the expected patterns observed in mammals (Carninci et al., 2006) and Arabidopsis (Yamamoto et al., 2009; Morton et al., 2014): a strong pyrimidine–purine (PyPu) signal at positions $\pm$1 relative to the TC peak, corresponding to the central part of the initiator motif, and a TA-rich pattern at positions –35 to –27, corresponding to the TATA-box (Figure 2C). TSSs defined by PEAT-seq and nanoPARE showed a highly similar pattern, whereas TSSs defined by TAIR10-annotated TSSs had weaker TATA-box and PyPu dinucleotide patterns. Remarkably, both signatures were almost absent around ARAPORT11-annotated TSSs, strongly indicating that ARAPORT11 poorly captures these central elements of promoter biology. Second, we assessed the average signal of nascent RNA approaches relative to TCs/TSSs. Three datasets were used: GRO-seq from two different laboratories (Hetzel et al., 2016; Zhu et al., 2018; from 6-d–old seedlings and inflorescence tissues, respectively; Supplemental Table 1) and 5′ GRO-cap from one laboratory (Hetzel et al., 2016). We observed a consistent shift of

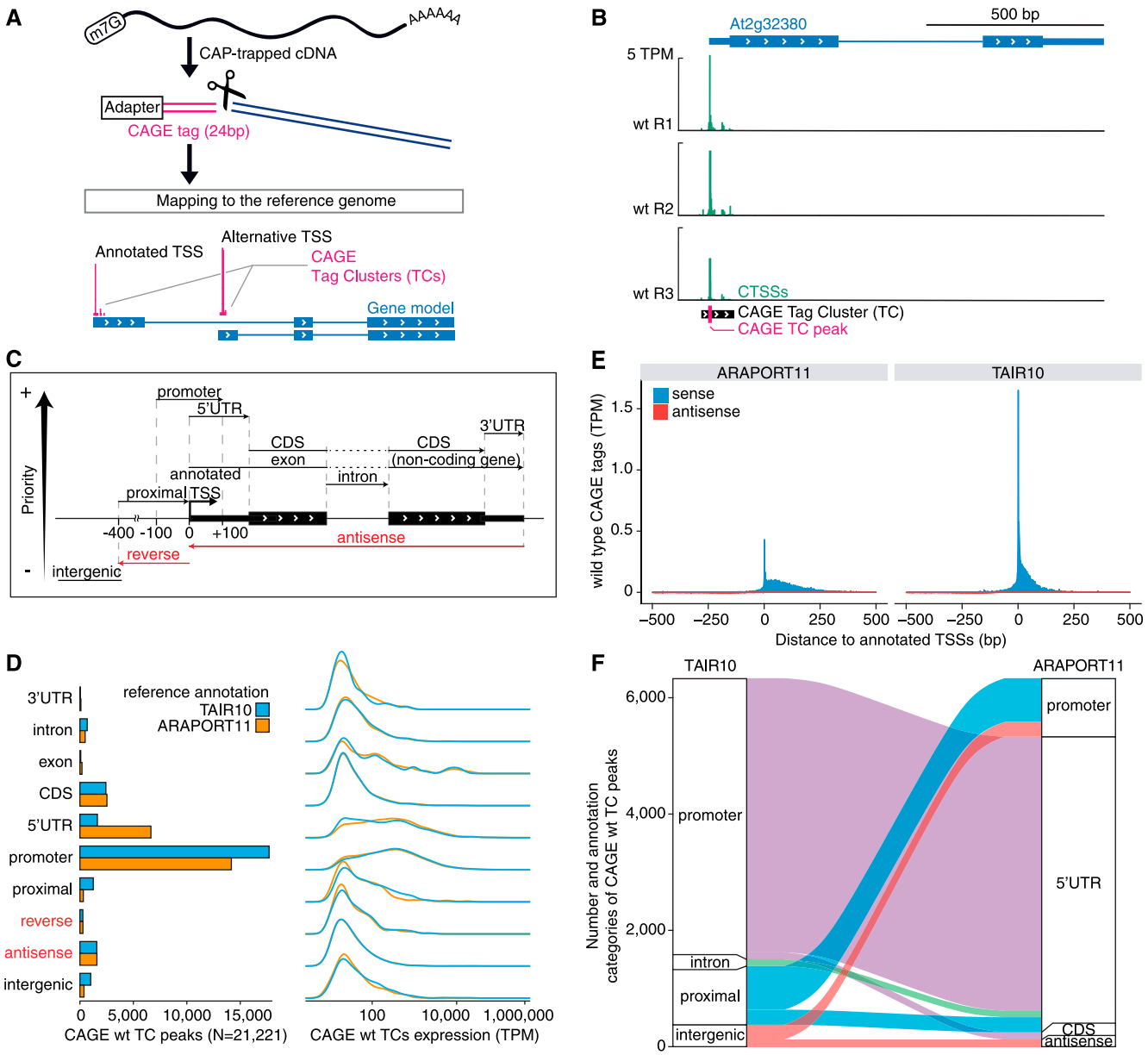


Figure 1. Experimental Design and Annotation of CAGE TCs.

(A) Conceptual overview of CAGE protocol.

(B) Typical genome browser view of CAGE data, using the At2g32380 locus as an example. Blue line, intron. CAGE TPM signals are shown in green on the plus strand. The At2g32380 locus had no antisense CAGE signal. Black block, CAGE TC; pink tick marker, CAGE TC peak. wt, wild-type.

(C) Hierarchical annotation strategy. Schematic representation of the annotation of CAGE TCs with respect to the reference genome. CAGE TC peaks are used as a proxy for CAGE-detected TSSs. When a CAGE TC peak overlaps several genomic features, the highest feature in the hierarchy is given priority, as indicated by the arrow (left margin). CAGE TCs falling outside of the indicated categories were annotated as intergenic. See "Methods" for details.

(D) Annotation and expression of CAGE TCs. (Left) Number of wild-type CAGE TCs (X-axis) falling into distinct genomic categories (Y-axis), following the hierarchical annotation strategy defined in **(C)**. (Right) Wild-type CAGE expression (X-axis) for the different annotation categories (Y-axis). Blue, TAIR10 annotation; orange, ARAPORT11 annotation. Red categories denote regions on the opposite strand of the gene, as in **(C)**.

(E) CAGE footprint at annotated TSSs. The X-axis shows the distance relative to ARAPORT11-annotated TSSs (left) and TAIR10-annotated TSSs (right) in bp. The Y-axis shows wild-type CAGE TPM-normalized signal per bp. Blue color and positive values indicate CAGE sense signal. Red color and negative values indicate CAGE antisense signal.

(F) CAGE TCs fall into different annotation categories when using TAIR10 or ARAPORT11 reference annotation data. Colors show the link between annotation categories. Only the pairs of categories with >120 CAGE TCs are shown.

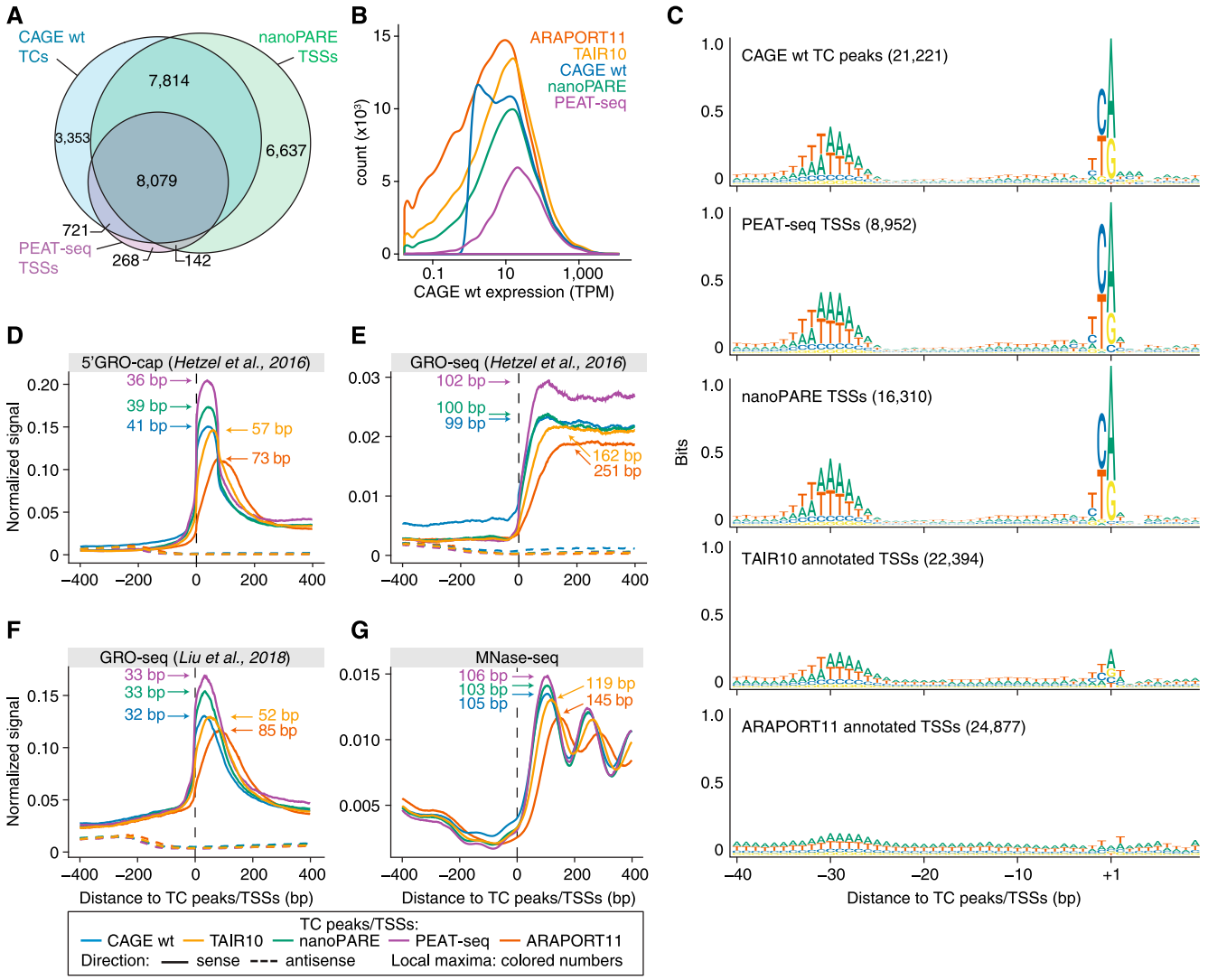


Figure 2. Comparison among TSS Datasets.

(A) Comparison of CAGE, PEAT-seq, and NanoPARE TCs/TSSs. Venn diagrams showing the number of TCs/TSSs identified by each dataset are based on the nonredundant set of all TCs/TSSs (see "Methods"). CAGE TCs are from whole seedlings, PEAT-seq TSSs originated from seedling roots, and nanoPARE TSSs are from flowering buds. A detail of tissues is available in Supplemental Table 1. wt, wild-type.

(B) Expression of TCs/TSSs and annotated TSSs across datasets. The X-axis shows the CAGE expression of TCs/TSSs as TPM (see Methods). The Y-axis indicates the number of TCs/TSSs. Line colors indicate datasets. wt, wild-type.

(C) Sequence patterns around TSSs/TC peaks. Sequence logos of the genomic sequences surrounding CAGE TC peaks, TSSs defined by PEAT-seq and nanoPARE, and annotated TSSs from TAIR10 or ARAPORT11 are shown. The Y-axis shows the information content in bits. The X-axis represents the distance relative to TC peaks/TSSs in bp. wt, wild-type.

(D) to **(F)** Nascent RNA metaplots around TSSs/TC peaks for all datasets listed in **(B)**. The X-axis shows the position relative to the annotated TSSs from TAIR10 or ARAPORT11, the TSSs from PEAT-seq or nanoPARE, or to TC peaks defined by CAGE. The Y-axis shows the average normalized signal of 5′ GRO-cap from Hetzel et al. (2016), whole seedlings **(D)**, and GRO-seq from inflorescence tissues (Hetzel et al. 2016; Liu et al., 2018). Solid lines, forward strand; dashed lines, reverse strand. Positions of local maxima are indicated with colored text and arrows. wt, wild-type.

(G) Nucleosome phasing in seedling leaves around TSSs/TC peaks in all datasets listed in **(B)** reported as average normalized MNase-seq signal (unstranded data from Zhang et al., 2015). Local maxima indicated as in **(D)** to **(F)**. wt, wild-type.

GRO-seq and GRO-cap signal maxima in accordance with the results discussed above: TCs/TSSs defined by CAGE, PEAT-seq, and nanoPARE yielded highly similar locations of GRO-seq maxima 35-bp to 40-bp downstream of the TSS (Figures 2D to 2F). For TAIR10-annotated TSSs, we observed a slight 3′-shift in the GRO-cap maximum, and this shift became substantial for ARAPORT11-annotated TSSs (Figure 2D). The same trend was observed using GRO-seq signal from two different laboratories

(Figures 2E and 2F). Third, to get a measure of nucleosome occupancy of genomic DNA around TSSs defined by the different annotations, we analyzed micrococcal nuclease-seq (MNase-seq) data from young Arabidopsis flower buds (Zhang et al., 2015; Supplemental Table 1). MNase-seq showed similar peaks and amplitudes when centered on TSSs defined by CAGE, PEAT-seq, and nanoPARE: In all of these cases, the first peak corresponding to the center of the +1 nucleosome was located, on average, 105-bp downstream from the TSS (+105 bp; Figure 2G). As above, we observed a small 3′-shift when centering on TAIR10-annotated TSSs (+119 bp), and a substantially shifted peak maximum and lower amplitude when using ARAPORT11-annotated TSSs (+145 bp; Figure 2G). Taken together, these results indicate that CAGE, PEAT-seq, and nanoPARE are the most accurate TSS sets, closely followed by TAIR10-annotated TSSs. By contrast, ARAPORT11 TSSs are misplaced by 128-bp upstream on average, possibly due to the overestimation of transcript lengths, as previously suggested by Schon et al. (2018). Because TSSs annotated in TAIR10 were in better agreement with the TSS definition by CAGE, PEAT-seq, and nanoPARE than those annotated in ARAPORT11, we used the TAIR10 annotation in all subsequent analyses.

Most Arabidopsis Genes Expressed in Unchallenged Wild-Type Seedlings Only Use One Promoter

The use of alternative promoters or TSSs is an important process for generating RNA isoforms in complex genomes (Davuluri et al., 2008; Valen et al., 2009; Pal et al., 2011). We define alternative TSSs as transcription initiation events that are distant from each other but within the same gene. Therefore, local dispersions of TSSs at a single promoter (so-called “broad TSS distributions”; Carninci et al., 2006) are not considered alternative TSSs in this analysis. Alternative promoters can generate mRNAs lacking one or more coding exons or upstream open reading frames and may have an impact on protein abundance and function. To assess the extent of alternative promoter usage, we counted the number of wild-type TCs overlapping each TAIR10-annotated gene, only retaining those that contributed ≥10% of the total CAGE expression across the gene to filter out minor events, as reported by Thodberg et al. (2019a). Our analysis showed that the vast majority of genes only used one TC (90%, Figure 3A), located in the annotated promoter (Figure 3B), although we also found a substantial number of genes with multiple TCs. For instance, 1,632 genes (9% of all expressed genes) had two TCs, each contributing at least 10% to the total expression of those individual genes (Figure 3A). When two TCs or more were observed, they mostly occurred either in the annotated promoter or in 5′-UTRs (Figure 3B). The few TCs occurring within protein-coding exons were typically not the most highly expressed TCs within the gene. A full list of genes and corresponding coordinates of TCs has been assembled in Supplemental Dataset 1, and two illustrative examples are shown in Figures 3C and 3D. In Figure 3C, the most used TC coincides with an annotated shorter protein-coding transcript, while Figure 3D illustrates a minor TC occurring at the very end of the gene, which is unlikely to yield a protein-coding RNA. These results indicate that in unchallenged wild-type Arabidopsis seedlings, the use of alternative promoters—resulting in

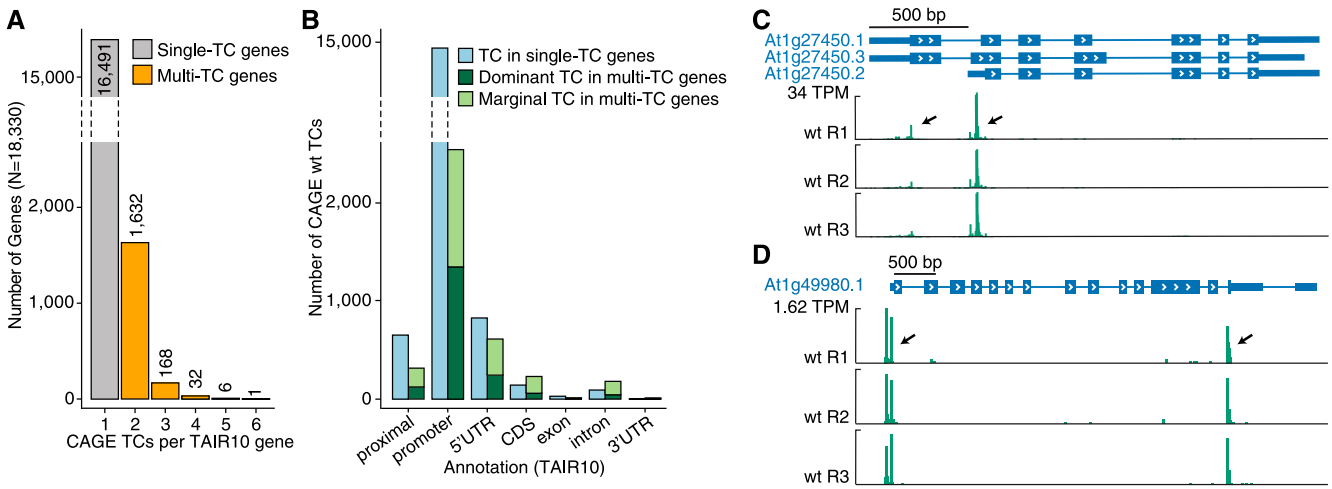


Figure 3. Detection of Alternative TSS Usage by CAGE.

(A) Alternative TSS occurrence in wild-type seedlings. The X-axis shows the number of sense TCs per TAIR10 gene. Only TCs accounting for at least 10% of the total number of CAGE tags mapping to the gene are counted. The Y-axis shows the number of genes. Bar colors separate single-TC (gray) from multi-TC genes (orange). The scaling of the Y-axis is split to facilitate visualization.

(B) Location of wild-type TCs within annotation categories based on the TAIR10 annotation. The Y-axis shows the number of TCs by category. Bar colors separate single-TC (blue) from multi-TC (green) genes. Dominant TCs (most expressed within a gene) in multi-TC genes are indicated in dark green, and remaining TCs are colored light green. TCs were filtered to contribute at least 10% of the total gene expression.

(C) and **(D)** Examples of genes with multiple TCs. Genome-browser images (organized as in Figure 1B), for At1g27450 (*ARABIDOPSIS THALIANA ADENINE PHOSPHORIBOSYLTRANSFERASE1*; **(C)**, and At1g49980 (DNA/RNA polymerases superfamily protein; **(D)**. Tracks show the wild-type CAGE data replicates on the same strand, represented as normalized TPM values. Black arrows, CAGE-detected TCs. wt, wild-type.

mRNAs encoding different proteins—is uncommon. By contrast, several studies have documented the widespread occurrence and functional importance of alternative transcription initiation sites in different tissues or upon perception of environmental change, most notably light quality and availability (Yamamoto et al., 2009; Ushijima et al., 2017; Kurihara et al., 2018).

Analysis of Bidirectional TSS Activity at Promoters of Protein-Coding Genes

We next focused on the critical problem of promoter directionality. To comprehensively map transcripts rapidly degraded by the nuclear exosome, we prepared triplicate CAGE and ribosomal-RNA-depleted total RNA-seq libraries from *hen2-4* and *rrp4-2* mutant seedlings. *hen2-4* is a null allele caused by an exonic insertion of a T-DNA (Lange et al., 2014), while the hypomorphic loss-of-function *rrp4-2* allele encodes a RRP4 protein containing a mutation in a conserved residue adjacent to the N-terminal domain that forms the interface of the core complex with the RRP6 exonuclease (Hématy et al., 2016). Principal component analysis on CAGE TC level showed that the wild type (as analyzed in Figures 1 and 2), *hen2-4*, and *rrp4-2* samples were all separable by the first principal component; *rrp4-2* was the most dissimilar from the wild type (Supplemental Figure 1A).

Our first objective was to use the data to assess the occurrence of PROMPT-like, exosome-sensitive transcripts in Arabidopsis. In vertebrates, TSSs of protein-coding genes tend to be located at the edges of DNase I hypersensitive sites (DHSs), marking open chromatin regions, close to the +1 and –1 nucleosomes. We therefore analyzed the overlap between the DHS regions identified in leaves or flowers (closed buds) of 14-d–old plants (Zhang et al., 2012; see Supplemental Table 1) and our CAGE data: 89.7% of CAGE tags from the wild type fell within identified DHS regions, which echoes results from HeLa cells (Andersson et al., 2014b). This observation thus justifies the use of our CAGE data together with the DHS-seq data of Zhang et al. (2012).

We first selected DHSs that overlapped with annotated TSSs of mRNAs. To distinguish sense from antisense CAGE signal, DHS regions were attributed to the same strand as the mRNA TSS they overlapped with. We then constructed a metaplot of the average CAGE TPM signal using the DNase I maxima (DHS summits, DHSSs) as anchor points. This analysis showed that the sense (mRNA) CAGE signal was on average 25-fold higher than the antisense signal, regardless of whether nuclear exosomal RNA decay systems were functional (Figure 4A). Specifically, we did not detect a high average enrichment of upstream signal on the reverse strand in exosome mutants, as would have been expected if exosome-sensitive PROMPTs were prevalent (Figure 4A). These observations suggest that divergent transcription resulting in exosome-sensitive PROMPTs at promoters of protein-coding genes is not widespread in Arabidopsis, thus agreeing with previous studies (Hetzel et al., 2016; Zhu et al., 2018). To ensure that these results were not specific to the CAGE technique, we also plotted RNA-seq reads from the wild type and *rrp4-2* mutant seedlings. The outcome was similar, albeit with an expected shift downstream of TSSs due to the inherent characteristics of RNA-seq data, which tend to miss RNA fragments at the 5′ and 3′ edges of transcripts (Figure 4B).

Inactivation of Nuclear Decapping Does Not Promote Detection of PROMPTs

Two different scenarios may explain the overall low CAGE and RNA-seq signal in exosome mutants upstream of mRNA TSSs on the reverse strand: (1) There is, on average, little transcription initiation in these regions, in agreement with the lack of GRO-seq signal (Hetzel et al., 2016); or (2) transcription of PROMPTs may occur, but Arabidopsis may employ redundant RNA decay systems for their degradation, precluding their detection by simple inactivation of the nuclear exosome. To explore the second possibility, we extended the RNA-seq experiments to additional mutants defective in nuclear decapping (and hence 5′-3′ exonucleolysis), or both nuclear decapping and exosome activities. To this end, we used the *lsm8-2* mutant (Perea-Resa et al., 2012; Golisz et al., 2013) and constructed a *rrp4-2 lsm8-2* double mutant in which both RRP4 and LSM8 functions were missing. RNA-seq metaplots of both *lsm8-2* single and *rrp4-2 lsm8-2* double mutants were highly similar to the wild type (Figure 4B), strongly indicating that LSM8 is not part of a redundant pathway for general PROMPT degradation.

Clear PROMPTs Can Be Detected at Individual Genes

Despite the low average signal upstream of genes on the reverse strand observed in metaplots, visual genome-browser inspection allowed us to detect cases of exosome-sensitive transcripts with the characteristics of mammalian PROMPTs. To identify such cases systematically, we analyzed the same TSS-overlapping DHSs as above and required CAGE expression on the reverse strand upstream of mRNA TSSs in *rrp4-2* to be twofold higher than in the wild type, and statistically significant ($\log_2$ fold-change ≥ 1, false discovery rate [*FDR*] ≤ 0.05). Despite these conservative selection criteria, we identified 96 unique regions with typical PROMPT/pre-mRNA configuration as described in mammals (see examples in Figure 4C and Supplemental Figure 1D; Supplemental Dataset 1). Similar to mRNA TSSs, PROMPT TCs showed a clear enrichment of TATA around position –30 and of PyPu at position ±1 (Figure 4D). These sequence signatures are consistent with previously proposed models where PROMPT TSSs have separate core promoters (reviewed by Andersson and Sandelin, 2020).

We next asked to what degree previously published nascent transcription data supported these clear PROMPT examples, and defined PROMPT regions with evidence of transcription as regions with 2-fold–more signal than the genome-wide average (see Methods). Using this criterion, approximately half of the identified PROMPTs were supported by nascent transcription data (on average 52%, depending on the nascent RNA dataset, see Supplemental Figure 1B; Hetzel et al., 2016; Liu et al., 2018). Given the high thresholds we employed for CAGE data, these results indicate that many PROMPTs did indeed escape detection by available nascent transcriptome data. However, we cannot exclude the possibility that the different tissues analyzed might impact the lower PROMPT detection rate in nascent RNA data (14-d–old seedlings in our study; 6-d–old seedlings, [Hetzel et al., 2016]; or inflorescences, [Liu et al., 2018]). The detection rates of PROMPT regions by PEAT and nanoPARE datasets were low: 1%

of PROMPTs were detected as a TSS by PEAT-seq, and 22% were detected by nanoPARE (in both cases allowing a window of ±100 bp of the CAGE TC peaks to be interrogated for signal). These low rates of detection likely reflect the necessity of using RNA decay mutants to detect PROMPTs by steady-state sequencing methods, but the different tissues used may also have an impact (see Supplemental Table 1).

Taken together, our data indicate that transcription from most promoters of protein-coding genes appears to be unidirectional, in agreement with previous conclusions. However, clear exceptions exist in which highly exosome-sensitive transcripts are initiated at locations similar to that of the much more widespread PROMPTs in vertebrates. It is therefore possible that initiation of transcription by RNAPII is fundamentally similar in plants and vertebrates, but that plants have evolved mechanisms in addition to selective nuclear RNA decay to ensure a directional promoter output.

Properties of PROMPTs and Their Associated Genes

We next sought answers to three questions to characterize the identified PROMPT-producing regions better. First, because the set of PROMPT-associated genes was small, is PROMPT

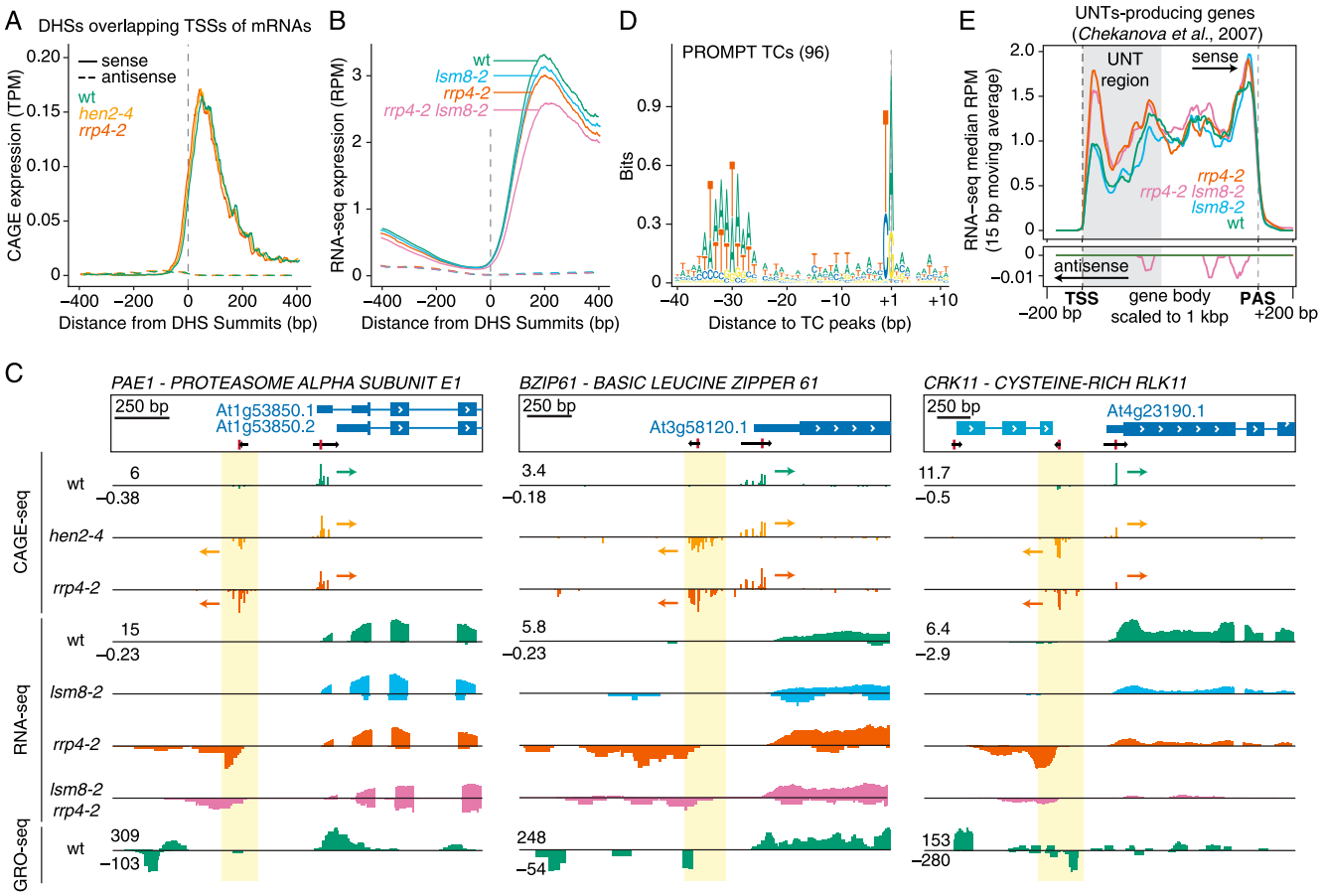


Figure 4. Analysis of Bidirectional Transcription Initiation at Promoters of Arabidopsis Protein-Coding Genes.

(A) The Y-axis shows the average normalized CAGE signal in TPM, smoothed with a 15-bp moving window. The X-axis represents the distance (in bp) relative to the maxima of the DHS peaks that overlap TAIR10 annotated TSSs. Colors show genotypes. Solid line, sense strand; dashed line, antisense strand. wt, wil -type.

(B) RNA-seq signal at DHSs overlapping annotated TSSs of protein-coding genes. The Y-axis shows normalized average RNA-seq signal as reads per million mapped reads (RPM); the X-axis is as in **(A)**. wt, wild-type.

(C) Examples of bidirectional transcription at promoters of protein-coding genes. Three cases of bidirectionally transcribed promoters of protein-coding genes organized as in Figure 1B are shown. (Top) TAIR10 transcript models (blue) and CAGE TCs (black; TC peaks in red). Tracks show normalized CAGE, RNA-seq, and GRO-seq signals (from Liu et al., 2018). Signals on forward strand have positive values; signals on reverse strand have negative values. Arrows indicate direction of transcription for CAGE TCs; yellow highlight indicates PROMPT TSS regions. wt, wild-type.

(D) Sequence patterns around exosome-sensitive PROMPT TCs, organized as in Figure 2C.

(E) RNA-seq metaplot for UNT-producing genes. The Y-axis shows the median sense and antisense normalized RNA-seq signal (10-bp moving average), colored by genotype. Positive values indicate sense signal; negative values indicate antisense signal (reinforced by arrow directions). The X-axis shows scaled position in UNT-producing genes, where gene bodies were scaled to 1 kbp. Vertical dashed lines, TAIR10-annotated TSSs and polyadenylation sites (PASs), flanked by 200 unscaled bp; gray box, UNT region. wt, wild-type.

production linked to a specific biological function of the associated coding genes? Gene-ontology (GO) term analysis showed no significant over-representation of GO terms, suggesting that there is no apparent functional link among PROMPT-associated genes. Second, because enrichment of poly(A) sites and depletion of 5′-splice sites is pronounced in mammalian PROMPT regions compared with mRNAs (Almada et al., 2013; Ntini et al., 2013), do the identified Arabidopsis PROMPT regions share this pattern? In contrast to mammals, we only observed a small enrichment of predicted poly(A) sites (AWTAA consensus) 200 bp or more downstream of PROMPT regions and no difference in the occurrence of predicted 5′-splice sites (see Methods; Supplemental Figure 1C). By contrast, we did observe a significant enrichment of predicted 3′-splice sites (Kolmogorov–Smirnoff test, $D = 0.33$, $P = 0.0009$; Supplemental Figure 1C), although the underlying mechanistic cause of this enrichment remains unclear. The modest enrichment in poly(A) sites may be due in part to the fact that plants use more diverse, and less well-characterized, poly(A) sites than mammals (Li and Du, 2014). Third, is there any relationship between PROMPT-producing regions and the UNT-producing genes identified by (Chekanova et al., 2007)? None of the 96 genes with exosome-sensitive PROMPT-like production belonged to the set of UNT-producing genes. Thus, antisense PROMPT production and sense UNT production are not linked. We note, nonetheless, that UNTs were clearly detected in our experiments as regions with higher RNA-seq signal in *rrp4-2* downstream of TSSs than in the wild type (Figure 4E).

PROMPTs Are Sources of Small Interfering RNAs in Arabidopsis

The inability to detect widespread PROMPT production by either nascent RNA techniques or steady-state RNA-seq methods even when using mutants deficient in nuclear RNA decay suggests that plants, in contrast to animals, may have evolved mechanisms to suppress PROMPT transcription or expression. A key difference in RNA metabolism between plants and mammals is that plants encode canonical RNA-dependent RNA Polymerases (RdRPs) whose activity may route single-stranded RNAs toward RNA interference (RNAi) pathways. RdRPs produce double-stranded RNAs that are converted into small interfering RNAs (siRNAs) by DICER-LIKE (DCL) enzymes. siRNAs may then guide RNA-Induced Silencing Complexes to silence complementary nucleic acids, a process that may include DNA methylation in plants via the RNA-directed DNA methylation pathway (Law and Jacobsen, 2010). Thus, PROMPTs may be substrates for the production of illicit siRNAs that may interfere with proper control of gene expression. To test this possibility, we prepared small RNA libraries from flowers (which express high levels of *RdRPs*) and leaves, using the wild type and *rrp4-2* mutants. For the flower libraries, we also included the *hen2-5* knockout mutant (Lange et al., 2014). Global analysis of small RNAs from these libraries showed that, at a genome-wide scale, small RNAs were no more abundant in regions upstream of mRNA TSSs (potential PROMPT-producing regions) in *rrp4-2* and *hen2-5* than in the wild type (Figure 5A; Supplemental Figures 2A and 2B). A scenario in which PROMPTs are generally produced, but turned over rapidly by RNAi, is therefore unlikely. However, the 96 observed PROMPT regions did produce more small RNAs in exosome mutants than in the wild type. First, these loci, indicated with red dots in Supplemental Figures 2A and 2B, tended to fall above the diagonal in the scatterplots showing small RNA read counts in exosome mutant versus the wild type. Second, regions producing detectable PROMPT TCs in exosome mutants displayed substantially higher small RNA read counts than those that did not (Figure 5A).

We performed two additional analyses to support the notion that the small RNAs mapping to PROMPT-producing regions were bona fide siRNAs, and not merely degradation fragments that may overaccumulate as a consequence of lost exosome function. We reasoned that if simple RNA degradation defects were the basis of small RNA accumulation in exosome mutants, small RNA production should be proportional to the expression level of the longer RNA precursor, and would be equally likely to take place in PROMPT and mRNA regions. We therefore calculated the relative small RNA densities in PROMPT and mRNA regions, using the expression levels from CAGE TCs in our seedling libraries to normalize small RNA read counts. We also split small RNA counts according to size, because Arabidopsis DCLs produce diagnostic size classes of siRNAs: DCL4 produces 21-nucleotide siRNAs, DCL2 produces 22-nucleotide siRNAs, and DCL3 produces 24-nucleotide siRNAs (Xie et al., 2004; Dunoyer et al., 2005; Xie et al., 2005). Regions with detectable PROMPTs in *rrp4-2* young flowers had markedly higher relative levels, particularly of 21- and 22-nucleotide small RNAs, than the wild type, while the corresponding mRNA regions did not (Figure 5B). Similar results, although clear only for 21-nucleotide siRNAs, were observed in leaves (Supplemental Figure 2C). Thus, small RNAs accumulate specifically in PROMPT regions and display a size bias typical of bona fide siRNAs produced by DCL4 and, to some extent, DCL2. This is important because DCL4/DCL2 are most likely to generate siRNAs from substrates whose ultimate source are RNAPII transcripts, as 24-nucleotide siRNAs produced by DCL3 are more tightly linked to RNA Polymerase IV transcripts (Law and Jacobsen, 2010). A typical example of siRNA generation in exosome mutants in regions with detectable PROMPTs is shown in Figure 5C. This example also shows that siRNAs accumulate on both forward and reverse strands, consistent with the involvement of an RdRP-DCL pathway rather than mere degradation. We conclude that PROMPTs indeed give rise to illicit siRNA production when allowed to accumulate in exosome mutants.

Evidence for Bidirectional TSS Activity at Intronic and Intergenic Space

We next examined the possibility of bidirectional transcription at intronic or intergenic loci, as this was shown to be a powerful predictor of enhancer regions and enhancer activity in vertebrates and insects (Andersson et al., 2014a; Rennie et al., 2018). Visual inspection also revealed such cases in Arabidopsis (examples in Figure 6). To identify such regions systematically, we used the same approach as reported by Thodberg et al. (2019a) to locate genomic regions with a balanced bidirectional CAGE signal. A total of 113 bidirectional clusters were identified, of which 78 were intronic, and the remaining 35 were intergenic (Supplemental Dataset 1). On average, such regions were more highly expressed

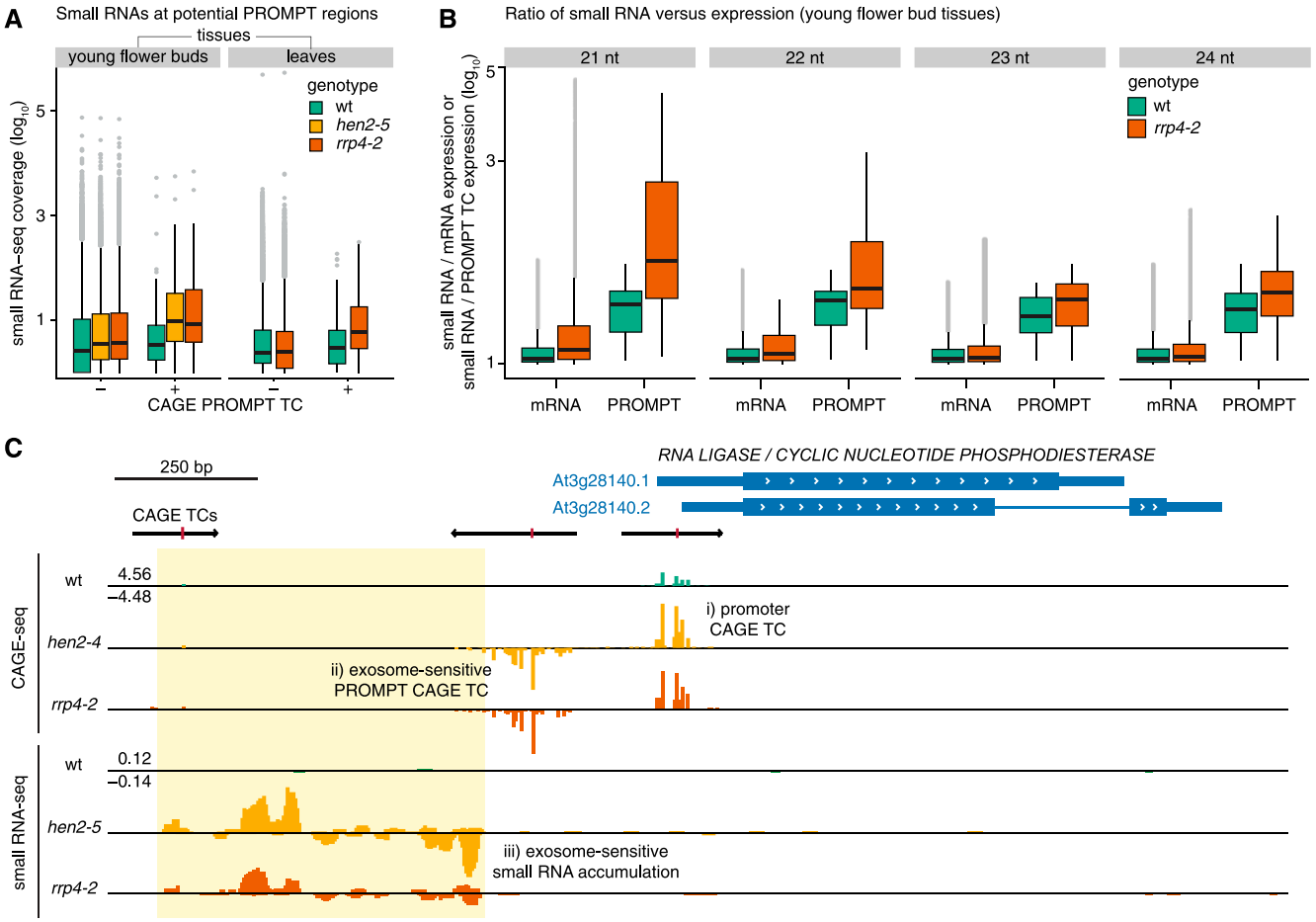


Figure 5. PROMPTs Give Rise to Production of siRNAs in Arabidopsis.

(A) Comparison of small RNA coverage at potential PROMPT regions versus detected PROMPT regions. PROMPT regions were defined as the 500-bp stretch upstream of TAIR10 annotated TSSs (see "Methods"). The X-axis indicates whether the PROMPT region hosted one of the CAGE-detected PROMPT TCs or not. The Y-axis shows the normalized small RNA read count (log-scale). Colors indicate genotype. Left and right groupings separate small RNA tissue origins (see Supplemental Table 1). wt, wild-type.

(B) Comparison of small RNA coverage (from 21-nucleotide to 24-nucleotide length) at PROMPT and mRNA regions from young flower bud tissues, normalized by expression levels (see Supplemental Table 1). mRNA, 500-bp stretch downstream of annotated TSSs; PROMPT, potential PROMPT regions host to a PROMPT CAGE TC (within the 500-bp stretch upstream of an annotated TSS). The Y-axis shows the normalized small RNA coverage divided by the CAGE TPM expression in the same region (PROMPT TC or mRNA CAGE TC, respectively; see "Methods") from the same mutants (in log-scale). For mRNA regions, only CAGE TCs annotated as promoters were considered. Colors indicate genotypes. wt, wild-type.

(C) Genome-browser view of a PROMPT region accumulating small RNAs. Organized as in Figure 4C. Text indicates (i) the mRNA promoter CAGE TC corresponding to the annotated TAIR10 TSS; (ii) an exosome-sensitive PROMPT TC on the opposite strand; and (iii) accumulation of small RNAs in the PROMPT region in exosome mutants, on both strands. wt, wild-type.

in *rrp4-2* or *hen2-4* compared with the wild type and exhibited clear average bidirectional CAGE transcription in all mutants except *lsm8-2*, both in CAGE and RNA-seq data (Figure 7A). A similar but less pronounced trend was observed in 5′ GRO-cap data (Figure 7B). RNAPII chromatin immunoprecipitation sequencing (ChIP-seq) signal from 10-d–old seedlings (Cortijo et al., 2017) was also enriched at the edges of NDRs; in intronic regions, these were highly balanced while there was a surprising minus-strand bias in intergenic regions (Figure 7C).

To further characterize these bidirectionally transcribed loci, we assessed their chromatin state using published data. The majority of bidirectional CAGE sites overlapped with a DHS (Figure 7D) and were centered on accessible DNA, as measured by DNase and MNase-seq (Figure 7E). Histones adjacent to the NDRs were enriched for H3K27ac (from 12-d–old seedlings; Chen et al., 2017) and to a lesser degree, H3K4me1 and H3K4me3 (from 3-week–old plants and 4-week–old leaves, respectively; Figure 7F; van Dijk et al., 2010; Wang et al., 2015; see Supplemental Table 1 for details of tissues). These patterns should be assessed with the caveat that the ChIP-seq data for histone modifications and RNAPII were not taken from the same tissues. We note, however, that the patterns observed were highly reminiscent of those observed in

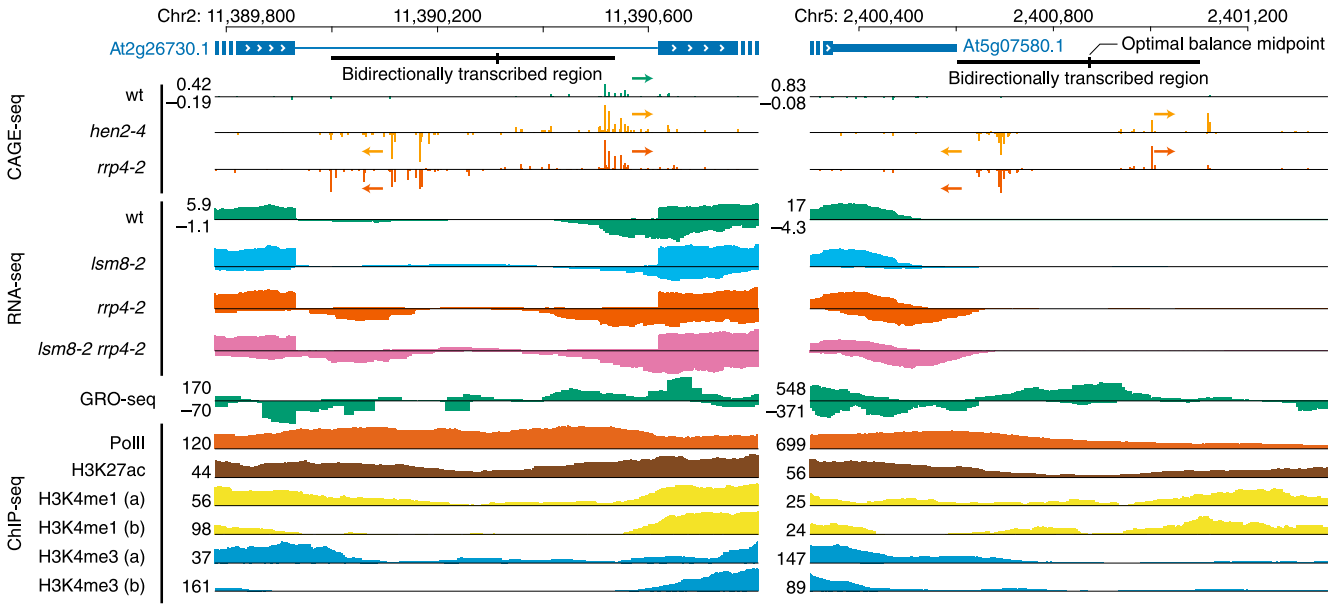


Figure 6. Examples of Bidirectionally Transcribed Intronic and Intergenic Regions in Arabidopsis.

Genome-browser views organized as in Figure 4C. Examples of intronic (left) and intergenic (right) bidirectionally transcribed regions are shown with the TAIR10 gene annotation (blue, top). Tracks from top to bottom: CAGE-seq, RNA-seq, and GRO-seq (Liu et al., 2018); ChIP-seq normalized signals for PolII (Cortijo et al., 2017); and H3K27ac (Chen et al., 2017), H3K4me1 (marked "a"; van Dijk et al., 2010), and H3K4me3 (marked "b"; Wang et al. 2015). Positive signals, sense transcription; negative signals, antisense transcription. Arrows indicate CAGE TCs. wt, wild-type.

mammals, with the exception of H3K4me1, which is highly enriched around enhancer regions in mammals. This may be explained by the association of H3K4me1 with active transcription elongation in Arabidopsis (Nielsen et al., 2019), but not in mammals. Overall, our analysis indicates that some DNase I hypersensitive intergenic and intronic loci in Arabidopsis feature bidirectional transcription initiation at NDR edges. The majority of these loci produce exosome-sensitive RNAs and have enhancer-associated chromatin marks. While such patterns are predictive of enhancer activity in vertebrates, their possible enhancer activity remains to be tested in Arabidopsis.

Identification of Sets of mRNAs Sensitive to Depletion of *RRP4* and *HEN2*

We reasoned that while PROMPTs and potential enhancer RNAs were rarely detected, even when depleting two distinct nuclear RNA degradation systems, a systematic comparison of CAGE and RNA-seq data from exosome mutants and the wild type might identify as-yet unknown transcripts in other regions of the genome. Because we wanted to find exosome substrates, we focused on transcripts that were more abundant in either mutant. We first identified CAGE TCs whose expression was significantly higher in *hen2-4* or *rrp4-2* compared with the wild type ($\log_2$ fold-change ≥ 1, $FDR \leq 0.05$, by the limma method, Ritchie et al., 2015; Supplemental Dataset 2). This resulted in a total of 1,747 upregulated TCs in both mutants combined. All upregulated CAGE TCs showed the expected nucleotide distribution around TSSs with enriched TATA and PyPu dinucleotide patterns at –30 and +1 bp, respectively (Supplemental Figure 3A), making it likely that they represent genuine TSSs. TCs upregulated in *hen2-4* were less numerous, and 92% of them were also significantly changed in *rrp4-2* compared with the wild type (Fisher's exact test, $P = 0$; Figure 8A). The few TCs ($n = 73$) significantly upregulated only in *hen2-4*, but not *rrp4-2* (Figure 8B), may represent transcripts whose exosome targeting requires HEN2, yet their exosomal decay still functions in the hypomorphic *rrp4-2* mutant. Alternatively, HEN2 may have as-yet undescribed functions independent of the exosome. This latter scenario is supported by the fact that only two of the 73 transcripts were found in the set of 848 transcripts upregulated upon inducible RNAi knockdown of RRP4 (Chekanova et al., 2007).

We next classified all TCs upregulated in either *rrp4-2* or *hen2-4* based on their overlap with the TAIR10 reference annotation (as in Figures 1C and 8C). This produced two surprising outcomes. First, a large number ($n = 479$) of TCs corresponding to sense mRNA TSSs were upregulated in the *rrp4-2* mutant. Some ($n = 112$) were also upregulated in *hen2-4*, but the majority ($n = 367$) was only upregulated in *rrp4-2*. Only 19 TCs were upregulated specifically in *hen2-4* (Figures 8D and 8E). Their upregulation might arise either as a direct consequence of reduced exosome activity or indirectly due to transcriptional induction caused by the *rrp4-2* mutation. In the first scenario of transcriptional upregulation, one might expect the mRNAs to encode functionally related proteins, but we found no significant enrichment of GO terms in the group of genes associated with the upregulated TCs. A more direct involvement of the exosome in the decay of these two groups of mRNAs would be of interest: The group of 131 mRNAs upregulated in *hen2-4* may be

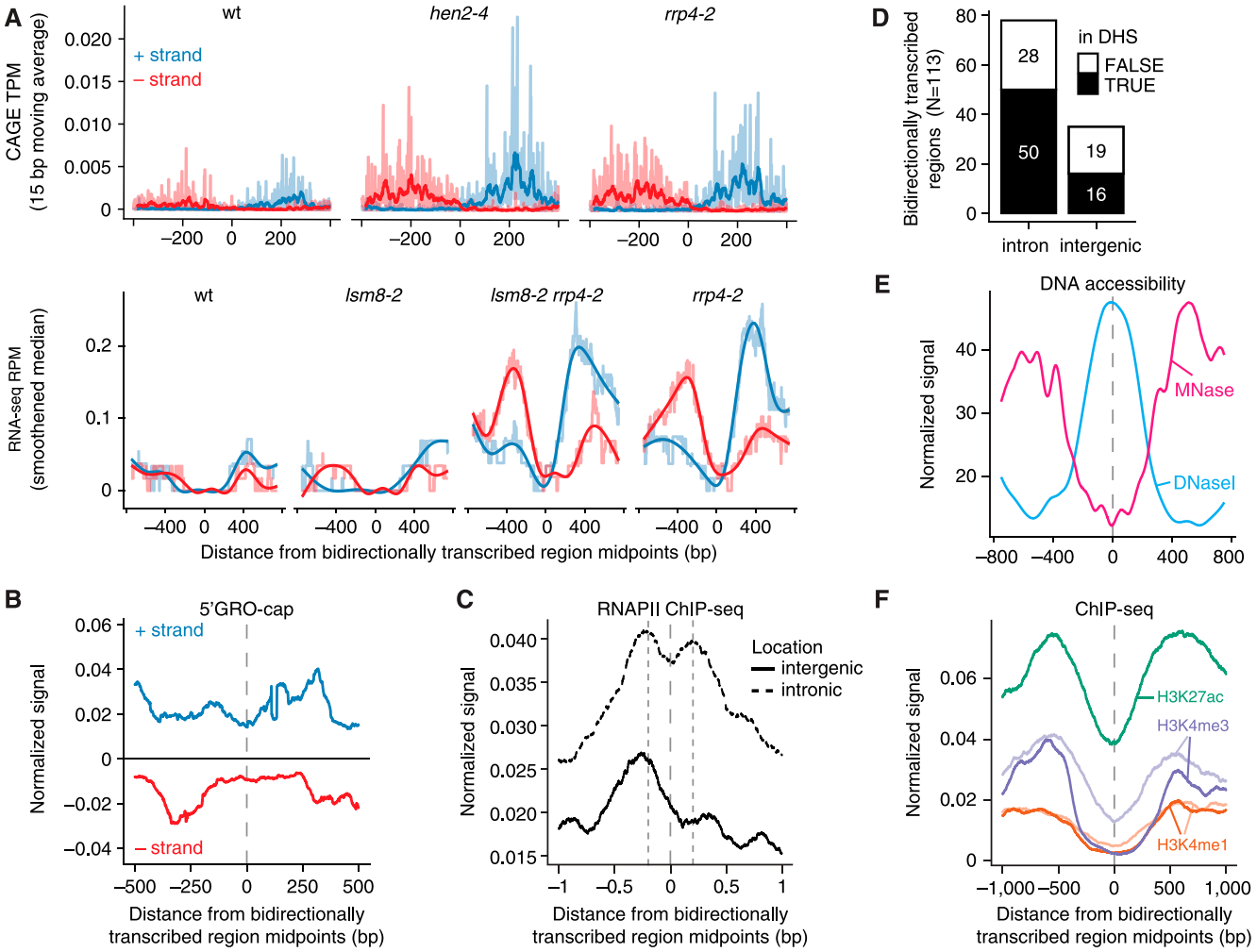


Figure 7. Analysis of Bidirectionally Transcribed Intronic and Intergenic Regions.

(A) CAGE and RNA-seq at bidirectionally transcribed regions. Y-axis shows CAGE (top) and RNA-seq (bottom) signal, colored by strand. CAGE dark line indicates the 15-bp moving-window average. RNA-seq dark line indicates the spline-smoothed signal. Light color, average normalized signal; blue, sense strand; red, reverse strand. The X-axis shows the distance relative to the midpoint position of 113 bidirectionally transcribed regions, in bp. Columns indicate genotypes. wt, wild-type.

(B) 5′ GRO-cap signal at bidirectionally transcribed regions. The Y-axis shows the average 5′ GRO-cap normalized signal from Hetzel et al. (2016). The X-axis is organized as in **(A)**. Blue color and positive values indicate plus strand. Red color and negative values indicate minus strand.

(C) RNAPII occupancy at bidirectionally transcribed regions. The X-axis is organized as in **(B)**. The Y-axis shows the average normalized RNA Pol II ChIP-seq signal from Cortijo et al. (2017). Solid line, intergenic region; dashed line, intronic region; vertical dashed lines, RNAPII maxima.

(D) Overlap of bidirectionally transcribed regions (intronic or intergenic) and DHSs. The X-axis shows the annotation category. The Y-axis shows the number of bidirectionally transcribed regions. Filled box indicates overlap with a DHS; open box indicates no overlap with DHS.

(E) DNA accessibility at bidirectionally transcribed regions. The X-axis is organized as in **(B)**. The Y-axis shows the average normalized MNase-seq (Zhang et al., 2015) and DNase-seq signals (Zhang et al., 2012), separated by color. The MNase-seq signal is proportional to the nucleosome occupancy, whereas the DNase-seq signal denotes NDRs.

(F) Histone marks at bidirectionally transcribed regions. The X-axis is organized as in **(A)**. Y axis shows normalized ChIP signal. Green, H3K27ac from 12-d–old seedlings (Chen et al., 2017); orange, H3K4me1 from 4-week–old leaves; violet, H3K4me3 from 3-week–old plants. For H3K4me1 and H3K4me3, color transparency distinguishes the source of the data (light, van Dijk et al., 2010; dark, Wang et al., 2015).

appreciably affected by nuclear mRNA quality control, akin to the poly(A)-tail exosome targeting pathway described in mammals (Meola et al., 2016), while the upregulation of the larger group of 367 mRNAs specifically in *rrp4-2* may be explained by a substantial contribution of the exosome to their cytoplasmic decay. Notably, the average expression of all of these moderately exosome-sensitive TCs of mRNAs was higher than most other categories, perhaps suggesting that the decay of highly expressed mRNAs tends to implicate the exosome (Supplemental Figure 3B).

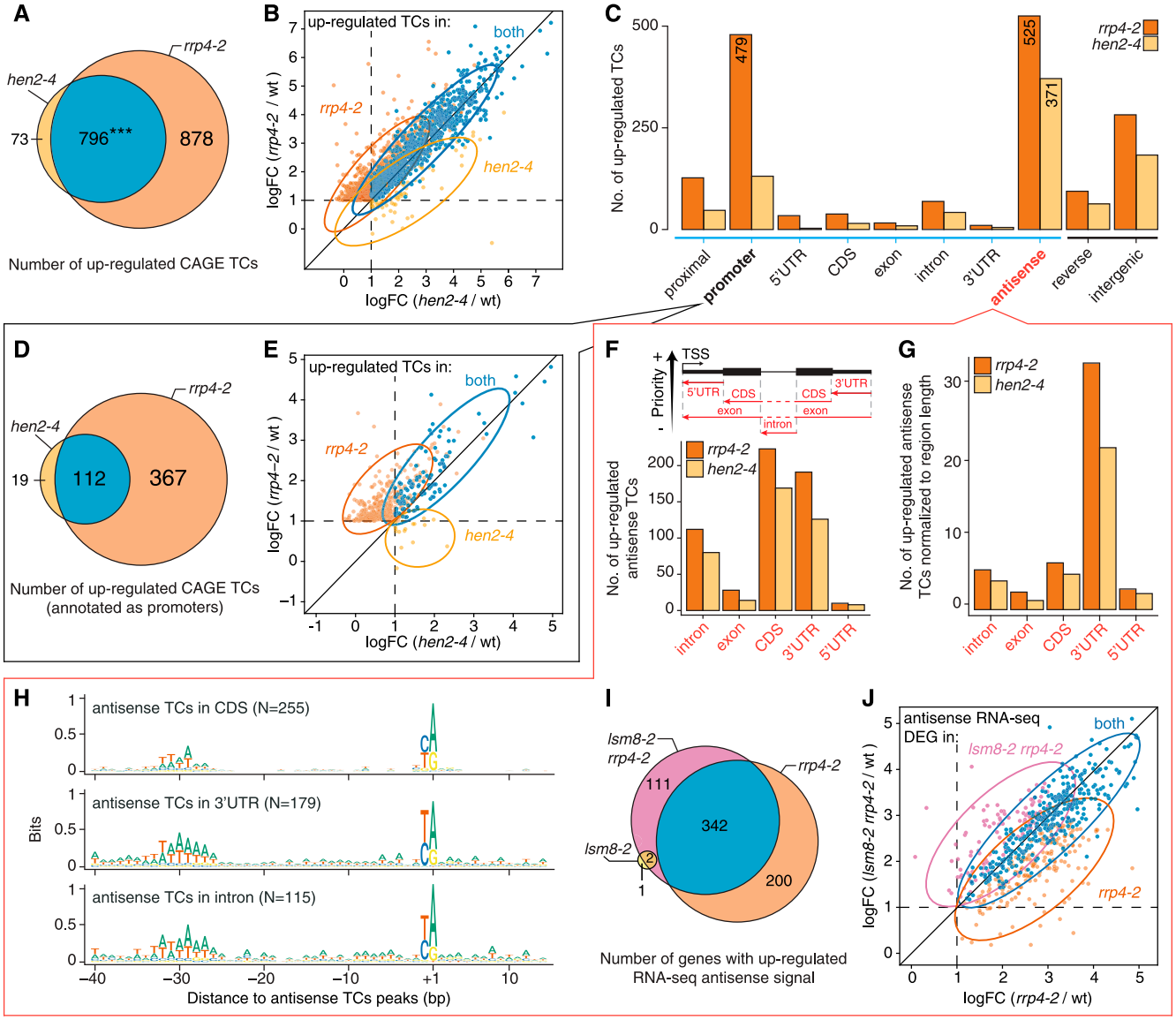


Figure 8. Systematic Discovery and Characterization of Exosome-Sensitive Transcripts in Arabidopsis.

(A) Extent of overlap between TCs upregulated in *rrp4-2* and *hen2-4*. A CAGE TC was defined as upregulated if differentially expressed in either mutant compared with the wild type ($\log_2$ fold-change ≥ 1, $FDR \leq 0.05$).
(B) Fold-change relationship between upregulated TCs across mutants. Scatterplot of $\log_2$ fold-change of *hen2-4* or *rrp4-2* mutants compared with the wild type, respectively. Each dot represents an upregulated CAGE TC, colored by whether its upregulation was specific or shared between mutants. Ellipses group the majority of TCs in each class. wt, wild-type.
(C) Annotation of CAGE TCs upregulated in *rrp4-2* and *hen2-4*. Y-axis shows the number of TCs overlapping each TAIR10 annotation category as bars, colored by genotype. X-axis shows gene annotation categories, as in Figure 1D. Red label denotes category on the opposite strand of the gene.
(D and E) Analysis of TCs falling into promoter regions (black callout box).
(D) Overlap of mRNA promoter TCs upregulated in *rrp4-2* and *hen2-4*, organized as in **(A)**.
(E) Fold-change relationship of upregulated mRNA promoter TCs across mutants, organized as in **(B)**. wt, wild type.
(F to J) Analysis of TCs detected on the strand antisense to genes (see red callout box in **F**).
(F) Annotation of antisense TCs upregulated in *rrp4-2* and *hen2-4*, using the antisense annotation hierarchy shown above the bar plot, where strandedness of TAIR10 features was inverted to allow annotation of antisense TC locations.
(G) Organized as in **(F)**, where the X-axis indicates the number of *rrp4-2* and *hen2-4* upregulated antisense TCs normalized to the genome-wide length of the respective regions (in Mbp).
(H) Sequence patterns of upregulated antisense TCs, shown as sequence logos, around antisense CAGE TC peaks, split by their annotation. Only categories with >100 exosome-sensitive antisense TC peaks are shown.

The second surprising outcome was related to a group of 525 CAGE TCs that were upregulated in *rrp4-2* relative to the wild type. The majority of these TCs ($n = 371$) were also upregulated in *hen2-4*. These TCs were neither PROMPTs nor bidirectionally transcribed intergenic or intronic loci analyzed above, but were located on the reverse strand within gene bodies, thus representing antisense RNAs (Figure 8C, red highlight). To annotate these antisense TCs more accurately, we used the same hierarchical procedure as above, but inverted the strand information in the annotation (Figure 8F, top). Exosome-sensitive antisense TCs tended to overlap with coding sequences (CDS) and 3′-UTRs of mRNAs (Figure 8F, bottom). After normalization of the number of TCs to the genome-wide lengths of CDSs and 3′-UTRs, upregulated antisense CAGE TCs were most enriched within 3′-UTRs (Figure 8G). More generally, antisense TCs had similar average expression and fold-change across annotation categories, which in turn tended to be similar to that of other exosome-sensitive TCs (Supplemental Figure 3C). The antisense CAGE TCs may represent genuine initiation events or partially degraded, recapped RNAs (Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009). We reasoned that if antisense TCs represent genuine TSSs, they should have a similar distribution of core promoter elements as other TCs. Indeed, the great majority of antisense TCs, regardless of overlap category, displayed a strong Initiator as well as a TATA motif (Figure 8H), suggesting that the exosome-sensitive, antisense TCs represent bona fide TSSs.

To verify the existence of these antisense RNAs, and to investigate possible redundancy in their degradation between exosome-dependent and LSM8-dependent pathways, we analyzed RNA-seq reads mapping to genes on the antisense strand, comparing mutants with the wild type. This confirmed the existence of the class of antisense RNAs identified by CAGE, but yielded slightly higher numbers of differentially expressed transcripts than the CAGE-based analysis: 656 genes were upregulated ($\log_2$ fold-change ≥ 1, $FDR \leq 0.05$, see Methods) in either *rrp4-2*, *lsm8-2* or *rrp4-2 lsm8-2* mutants compared with the wild type on the antisense strand (Supplemental Dataset 2). For the remaining transcripts, 52% were upregulated in both *rrp4-2* and *rrp4-2 lsm8-2* mutants, while relatively large sets were significantly upregulated only in *rrp4-2* single mutant (30%) or the *rrp4-2 lsm8-2* double mutant (17%; Figure 8I). Plotting the $\log_2$ fold-change of *rrp4-2* versus the wild type against the $\log_2$ fold-change of *rrp4-2 lsm8-2* compared with the wild type revealed that genes found to be differentially expressed in the *rrp4-2 lsm8-2* mutant had the same trend in *rrp4-2* single mutants (Figure 8J). Thus, the RNA-seq data strongly corroborated the existence of exosome-sensitive antisense RNAs. Further, we found no clear cases of redundancy between LSM8-dependent and exosome-dependent pathways involved in their degradation.

Characterization of Exosome-Sensitive Antisense RNAs Initiating Toward 3′-Ends of Genes

Antisense RNAs initiating within 3′-regions of genes encoding developmental regulators can play essential roles in developmental transitions in plants, including germination and flowering (e.g., Fedak et al., 2016). Given this, and based on the unexpected prevalence of exosome-sensitive TSSs in 3′-regions of genes, we characterized this set of TSSs and their cognate RNAs in more detail. A total of 354 CAGE TCs antisense to an annotated 3′-UTR region were found across all samples using the same expression cutoffs as above, of which 197 (56%) were upregulated in at least one of the mutants compared with the wild type (Figure 8D). Interestingly, genes with an upregulated CAGE TC antisense to their 3′-UTR showed a weakly significant enrichment in "sequence-specific DNA binding" GO term (GO:0043565, $FDR = 0.02$, 8.7% of genes), indicating that genes encoding TFs were more prone to have antisense transcription initiation within their 3′-UTRs. Examples of exosome-sensitive antisense transcripts and TCs are shown in Figure 9A.

CAGE TCs antisense to 3′-UTRs showed no specific bias in terms of location within the 3′-UTR, but were substantially more likely to occur in longer 3′-UTRs (Supplemental Figures 4A and 4B; a summary of statistical tests is available in Supplemental Table 3). Because the Arabidopsis genome is compact, we hypothesized that exosome-sensitive antisense TCs might be PROMPTs from closely located downstream genes. However, this phenomenon appears to be rare: Only 16 instances were identified where the maximal distance between a 3′-UTR antisense TC and its closest downstream mRNA TSS was 300 bp, and this number only increased to 39 instances when a maximal distance of 500 bp was allowed (Supplemental Figure 4C). More generally, there was no relationship between the intergenic distance between genes (3′ end of gene to next 5′ or 3′ gene end downstream) and whether the 3′-UTRs had exosome-sensitive antisense CAGE TCs (Supplemental Figure 4D).

Because we had both CAGE and RNA-seq data, we analyzed RNA-seq read density across sense and antisense strands for the 481 genes that had at least one differentially upregulated antisense CAGE TC in *rrp4-2* or *hen2-4* (the set is larger than above, because antisense CAGE TCs within regions other than 3′-UTRs were included). Antisense RNA-seq reads were uniformly distributed across the gene body in the exosome mutants except at the 5′- and 3′-ends, while the antisense signal was on average 14.5-fold lower than that of the sense strand (Figure 9B, right). Conversely, CAGE antisense reads resided in the last ~ 20% of

Figure 8. (continued).

(I) Overlap between genic antisense transcripts upregulated in *rrp4-2*, *lsm8-2*, and *rrp4-2 lsm8-2*. The RNA-seq signal, antisense to annotated genes, was used to identify transcripts upregulated in a mutant compared with the wild type. A gene was defined as upregulated if differentially expressed in either mutant compared with wild-type ($\log_2$ fold-change ≥ 1, $FDR \leq 0.05$).

(J) Fold-change relationship between antisense upregulated genes across mutants. Organized as in **(B)**, but using RNA-seq data antisense to a TAIR10-annotated gene. X-axis show *rrp4-2* mutants versus wild-type fold-change; Y-axis shows the double mutant *rrp4-2 lsm8-2* fold-change. wt, wild-type.

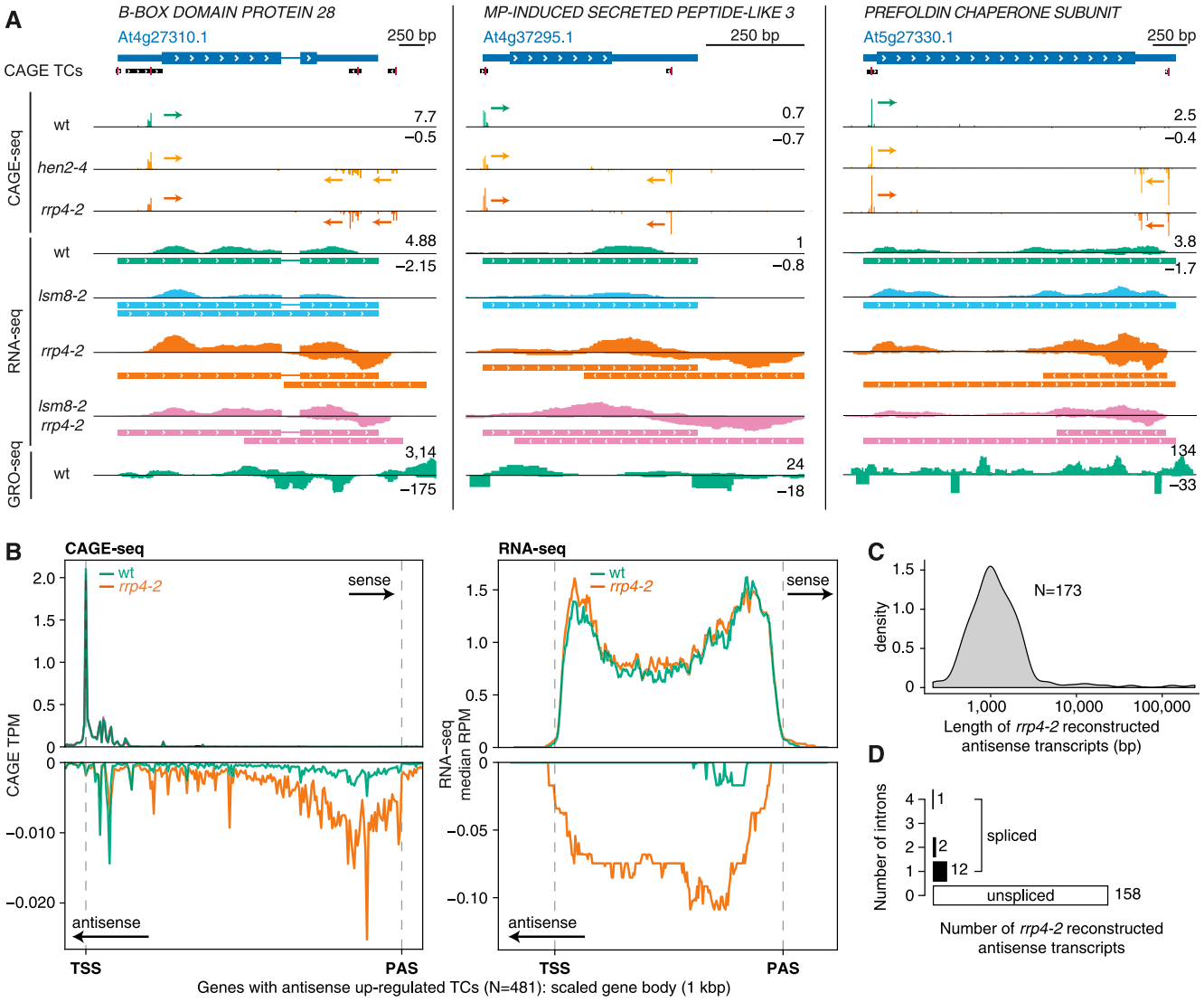


Figure 9. Characterization of Exosome-Sensitive 3′-UTR Antisense TCs.

(A) Genome-browser views of exosome-sensitive 3′-UTR antisense TCs. De novo assembled antisense transcripts using RNA-seq datasets are also shown for the mutants in which the transcripts are detectable. Track colors as in Figure 4. wt, wild-type.

(B) Metaplots of CAGE-seq and RNA-seq signal at genes with upregulated antisense TCs. The Y-axis shows the average or median normalized read count in 5-bp windows across 481 genes with an antisense, exosome-sensitive TC. Gene bodies, shown on the X-axis, were scaled to 1 kbp. Green line, wild-type; orange line, *rrp4-2* mutant. Reads on sense strand have positive values; reads on antisense strand have negative values. Dotted vertical lines indicate positions of annotated TSSs and polyadenylation sites (PASs). wt, wild-type.

(C) Distribution of the lengths of 173 de novo reconstructed transcripts, antisense to a gene, and upregulated in the *rrp4-2* mutant. X-axis shows length in bp, log-scaled; Y-axis shows distribution density.

(D) Number of detected introns in exosome-sensitive, de novo reconstructed antisense transcripts in the *rrp4-2* mutant. X-axis indicates the number of exosome-sensitive de novo assembled antisense transcripts for the *rrp4-2* mutant; Y-axis shows the number of introns detected in these transcripts.

the gene (Figure 9B, left), consistent with their dominant 3′-UTR overlap observed above. Last, we leveraged the RNA-seq data to investigate the length and splicing status of upregulated antisense transcripts in *rrp4-2*. To assess this in as unbiased a way as possible, we used strand-aware de novo transcript assembly for RNA-seq reads (see Methods and Supplemental Dataset 3) in *rrp4-2*, and then plotted the length of novel antisense transcripts (Figure 9C) and their number of introns (Figure 9D). The length of antisense transcripts had a large variance, but their most common length was ~1,000 bp, and the vast majority (91.3%) were unspliced. An important caveat of this analysis is that de novo transcript assembly is only possible if sufficient RNA-seq reads are present. This means that transcript length and splicing status are measured only for more highly expressed transcripts, and

length estimates may be conservative. Taken together, our analysis revealed that a substantial number of Arabidopsis genes exhibit antisense transcription of unspliced, exosome-sensitive antisense RNAs—a property that is more common in genes encoding TFs.

The *DELAY OF GERMINATION1*/*asDELAY OF GERMINATION1* Locus Exemplifies a Complex TSS Organization with a Functional Antisense lncRNA

As a final example of the utility of our data, we analyzed the *DELAY OF GERMINATION1* (*DOG1*) locus in detail. The DOG1 protein is an essential developmental regulator that controls seed dormancy such that *dog1* loss-of-function mutants exhibit uncontrolled seed germination, even during fruit development, while *DOG1* overexpressors cannot break seed dormancy (Fedak et al., 2016). Importantly, *DOG1* expression is regulated in cis by an lncRNA, *asDOG1*, initiated at an alternative 3′-UTR in exon 2 (Fedak et al., 2016). Overlaying our RNA-seq and CAGE data shows the overall complexity and exosome effects at this locus (Figure 10). First, *DOG1* has two alternative sense strand TSSs based on CAGE data: the first corresponds roughly to TAIR10 annotation, while the second is located within exon 2 and produces an uncharacterized transcript. Interestingly, both CAGE and RNA-seq shows that these transcripts have a degree of exosome sensitivity, which is uncommon for mRNAs. Second, we identified an antisense CAGE TC very close to the previously annotated TSS of *asDOG1*: Both CAGE and RNA-seq data show that *asDOG1* is highly exosome-sensitive (notably, TAIR10 annotation suggests that this TSS lies within the coding region of exon 2, while Fedak et al., 2016 showed that this region has a dual function as a 3′-UTR for a shorter mRNA isoform). Consistent with our results above, de novo RNA-seq isoform reconstruction indicated that *asDOG1* was not spliced and extended until the annotated TSS of *DOG1* (estimated length of 1,223 bp), agreeing with the rapid amplicifation of cDNA ends (RACE) experiments from the original study (Fedak et al., 2016). Thus, the well-established cis-regulator of gene expression *asDOG1* is a prominent example of the set of exosome-sensitive antisense RNAs that we identified here. It is therefore plausible that more of these lncRNAs, or the act of their transcription, have regulatory potential. This hypothesis is particularly appealing because of the enrichment of genes encoding TFs in the group of genes that gives rise to exosome-sensitive antisense transcripts.

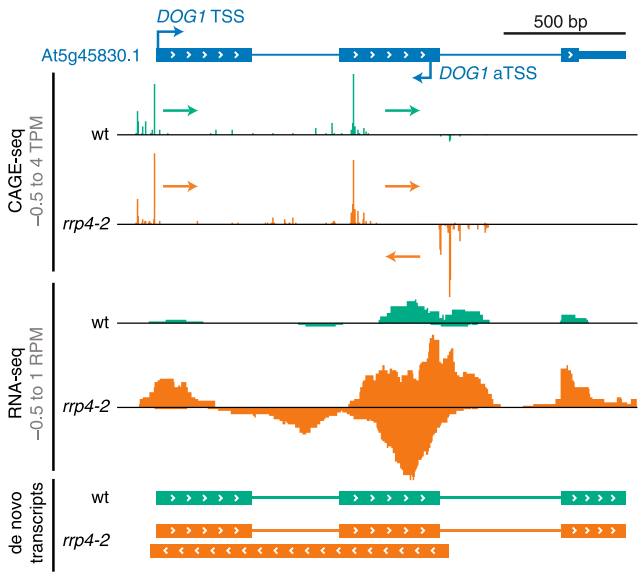


Figure 10. Example of a Complex Antisense Region with a Functional Antisense lncRNA: the *DOG1*/*asDOG1* Locus.

Genome-browser view of the *DOG1* gene, organized as in Figure 4C. The At5g45830 (*DOG1*) gene model is represented at the top with arrows marking the sense-annotated TSS and an antisense TSS (aTSS) in the second exon from Fedak et al. (2016). Tracks show normalized CAGE and RNA-seq signal, and de novo reconstructed transcripts using respective RNA-seq datasets. Arrows indicate CAGE-detected TCs and their transcriptional direction.

DISCUSSION

Divergent Transcription at Genes Is Uncommon, But Not Absent in Arabidopsis

The initial aim of our study was to characterize the extent of bidirectional transcription at promoters of protein-coding genes as well as in intergenic and intronic regions in Arabidopsis, given their frequent occurrence in vertebrates. We did this using two complementary steady-state RNA methods (CAGE and RNA-seq), where critical nuclear RNA degradation pathways were rendered nonfunctional. Our results largely agree with previous efforts using nascent RNA methods (Hetzel et al., 2016): PROMPTs are rare in Arabidopsis, but do exist. Using conservative thresholds, we identified nearly 100 PROMPTs that are clear exosome substrates. Notably, such cases were not reported previously using nascent RNA methods, even though approximately half had nascent RNA support.

It is unclear why so few genes feature distinctive PROMPTs in *rrp4-2* or *hen2-4* mutants. There was no discernible difference in terms of sequence motif predictions previously associated with exosome sensitivity (Almada et al., 2013; Ntini et al., 2013), and no apparent shared role among genes with PROMPTs. Principally, the lack of detected PROMPTs may be either due to (1) the absence of transcription initiation on the reverse strand upstream of Arabidopsis mRNA TSSs, or (2) the efficient degradation of PROMPTs by redundant RNA degradation systems, so that transcripts remain undetectable even after the loss of exosome activity. The lack of signal in nascent RNA assays together with our failure to detect widespread occurrence of PROMPTs using double mutants is consistent with the first hypothesis. Although our results combined with previous nascent RNA studies can most easily be interpreted as evidence for the general absence of PROMPT transcription in Arabidopsis, we cannot completely rule out the possibility that PROMPT transcription is widespread, but coupled with redundant nuclear RNA decay pathways that include exosome- and LSM8-independent mechanisms. In particular, we note that nuclear decapping pathways are relatively poorly understood, and it is formally possible that LSM8-independent

mechanisms of nuclear decapping exist that allow 5′-3′ decay of PROMPTs in the *rrp4-2/lsm8-2* mutant background.

We also found evidence of bidirectional transcription at intergenic NDRs producing exosome-sensitive transcripts, which share many of the same signatures as mammalian enhancer regions, including chromatin states, with the caveat that tissues were not always fully matched between experiments. It remains to be seen whether these enhancer-like regions bind TFs and have a role in regulating distal gene transcription initiation.

Reasons for PROMPT Scarcity in Plants

Why Arabidopsis promoters, unlike those of vertebrates, generally initiate predominantly in only one direction, and what might have favored the evolution of promoter directionality, are important questions with ties to fundamental aspects of transcription and RNA metabolism. A genetic study in yeast demonstrated that chromatin-based mechanisms can enforce promoter directionality: The histone chaperone, Chromatin Assembly Factor I, which facilitates replication-coupled nucleosome assembly (Smith and Stillman, 1989), also restricts divergent transcription, while histone exchange promoted by acetylation of Lysine 56 on histone 3 (H3K56ac) promotes divergent transcription (Marquardt et al., 2014). Similar or other independently evolved chromatin-based mechanisms may act in plants to enforce promoter directionality. Insights into such mechanisms could be rewarding for an understanding of RNAPII transcription in plants, and the data presented here should allow the design of genetic screens, akin to those used by Marquardt et al. (2014) toward the isolation of missing factors.

One reason for the repression of PROMPT initiation in Arabidopsis may lie in the high gene density of its genome, whereby PROMPTs may interfere with closely located upstream gene transcription. However, the fact that PROMPTs are also rare in maize that has a less gene-dense and larger genome than many vertebrates (Erhard et al., 2015; Lozano et al., 2018) does not support this hypothesis, and suggests that PROMPT suppression may be more general in plants. What might drive suppression of upstream transcripts specifically in plants within the eukaryotic kingdom? Our analysis of small RNAs in exosome mutants provides a plausible explanation. We observed clear accumulation of 21-nucleotide and 22-nucleotide siRNAs in the regions that produce detectable PROMPTs in exosome mutants, demonstrating that PROMPTs, if allowed to accumulate, are efficiently converted into siRNAs. We did not examine whether these siRNAs cause aberrant gene repression, but note that this understanding of RNA-directed DNA methylation (RdDM) in plants allows for the possibility that PROMPT-derived siRNAs may induce aberrant DNA methylation patterns and unwanted transcriptional repression, perhaps over the course of generations. Although RdDM-loci are associated with 24-nucleotide siRNA production via a RNA Polymerase IV-RDR2-DCL3 module (Singh et al., 2019), loci with initial bursts of 21-nucleotide to 22-nucleotide siRNAs, generally used for post-transcriptional gene regulation, may switch to epigenetic silencing via 24-nucleotide guided RdDM (Marí-Ordóñez et al., 2013; McCue et al., 2015). Thus, stringent transcriptional control of endogenous genes may be jeopardized by PROMPT production in plants, because of their tendency to be converted into siRNAs by RdRPs and DCL enzymes. In contrast to plants, mammals and insects do not encode RdRPs, perhaps explaining why PROMPT degradation solely via the nuclear exosome is sufficient to protect against additional potentially deleterious effects of these lncRNAs in these species.

A Large Class of Exosome-Sensitive Antisense RNAs Initiating in 3′ Regions of Protein-Coding Genes

The large number of antisense TCs residing within the 3′-UTRs of mRNAs was an unexpected finding, and differs from earlier studies showing peaks of RNAPII or GRO-seq signal at or downstream of 3′ ends on the sense strand (e.g., Zhu et al., 2018). These TSSs define the initiation of long, exosome-sensitive unspliced antisense RNAs, which often cover large parts of the cognate mRNA. The function of these transcripts, if any, is unclear. However, it is interesting to note a clear over-representation in genes encoding TFs: Antisense transcripts may be repressing the mRNA by RNA-RNA hybridization, or antisense transcription may repress the transcription of the cognate mRNAs by steric effects such as RNAPII collision, or clearing of DNA binding proteins and chromatin states. Indeed, there is already evidence for regulation of developmental and stress response genes by antisense RNAs (Liu et al., 2010; Henriques et al., 2017; Kindgren et al., 2018), including the *DOG1/asDOG1* example in Figure 10, which is similar to the class of RNAs we show here to be prevalent. This class of exosome-sensitive antisense RNAs is, therefore, an obvious focus for future functional delineation of noncoding RNA and noncoding transcription in plants.

METHODS

Plant Materials

All Arabidopsis (*Arabidopsis thaliana*) plants are of the Col-0 ecotype. The *hen2-4* (At2g06990, SALK_091606), *hen2-5* (At2g06990, SALK_019457), and *lsm8-2* (At1g65700, SALK_048010) mutants were described by Lange et al. (2011, 2014) and Perea-Resa et al. (2012), respectively. Seeds were kind gifts from Dominique Gagliardi (*hen2-4*) and Julio Salinas (*lsm8-2*). The *rrp4-2* hypomorphic mutant is described by Hématy et al. (2016) and was kindly provided by the authors.

Genotyping

DNA was isolated by adding one volume of phenol-chloroform (50:50 [v/v]) to freshly ground leaves in urea buffer (42% [w/v] urea, 312.5 mM of NaCl, 50 mM of Tris-HCl at pH 8, 20 mM of EDTA, and 1% [w/v] N-lauroylsarcosine). Phases were separated by centrifugation (14,000 rpm for 10 min a 4°C) and the supernatant containing DNA was isolated. Nucleic acids were precipitated with the addition of one volume of isopropyl alcohol. DNA was pelleted by centrifugation (as above) and rinsed with 70% (v/v) ethanol. The resulting purified DNA was used as a template for polymerase chain reaction to confirm the T-DNA insertion in *hen2-4*, *hen2-5*, and *lsm8-2*. The G55E single-point mutation in the *rrp4-2* mutant was confirmed by target DNA amplification followed by enzymatic digestion (*Eco*47I). Oligonucleotides used in this study are listed in Supplemental Table 2.

Growth Conditions

For the CAGE samples, wild-type Col-0, *hen2-4*, and *rrp4-2* seeds were surface-sterilized first by incubation in 70% (v/v) ethanol (2 min) followed by 1.5% (w/v) sodium hypochlorite, 0.05% (w/v) Tween-20 (10 min Sigma-Aldrich), and rinsed three times with sterile double-distilled water. Seeds were stratified in complete darkness at 4°C for 72 h, then germinated on full-strength Murashige and Skoog medium at pH 5.7 and supplemented with 1% (w/v) Suc and 0.8% (w/v) agar in sterile conditions in Petri dishes under a 16-h L/8-h D photoperiod (130 μmol photons m^{-2} s^{-1} at 21°C, cat. no. Master TL-D 36W/840 and 18W/840 bulbs; Philips). Intact 12-d–old seedlings were transferred to 8 mL of 1x (4.33 g/L) liquid Murashige and Skoog medium at pH 5.7, supplemented with 1% Suc, in six well-plates (Nunc) and allowed to acclimate for 48 h with mild agitation (130 rpm), under the same light and temperature conditions as above. For the RNA-seq samples, seeds for wild-type Col-0, *lsm8-2*, *rrp4-2*, and the *rrp4-2 lsm8-2* double mutant were subjected to the same sterilization, stratification, and growth conditions as described for the CAGE samples. For the small RNA-seq samples (wild type Col-0, *hen2-5*, and *rrp4-2*), seeds were surface-sterilized as above and spotted directly on soil (Plugg/Såjord [seed compost]; SW Horto), then grown in Percival chambers for a 16-h L/8-h D photoperiod (130 μmol photons m^{-2} s^{-1}, cat. no. Master TL-D 36W/840 and 18W/840 bulbs; Philips) and temperature cycles (21°C during the day, 18°C at night) for three weeks for leaf production, or six weeks for flower production.

Total RNA Extraction

For the CAGE-seq and RNA-seq samples, total RNA was extracted for three biological replicates, each formed of a pool of 10 complete and intact 14-d–old seedlings. For the small RNA-seq samples, total RNA was recovered from a pool of two to three leaves or young flower buds across three different repeats of three independently grown replicates, for a total of nine small RNA samples. Collected plant material was flash-frozen in liquid nitrogen and finely ground before adding 1 mL of TRI-Reagent (Sigma-Aldrich) per 100 mg of ground tissue and vortexed directly. Phase separation was achieved by adding 200 μL of chloroform, vigorous shaking, and centrifugation at 4°C (10 min at 15,000 rpm). The aqueous phase was transferred to a fresh tube and one volume (400 μL) of isopropyl alcohol was added to precipitate RNA at room temperature (30 min). The total RNA was pelleted after centrifugation at 4°C (15 min, 15,000 rpm) and rinsed three times with 70% (v/v) ethanol before solubilization in sterile double-distilled water. To remove contaminants and obtain higher quality RNA material, polysaccharide precipitation was performed as described by Asif et al. (2000). Total RNA was assessed for concentration and purity using a NanoDrop ND-1000 (Thermo Fisher Scientific) and its integrity was checked using a Bioanalyzer 2100 with High Sensitivity RNA chip (RNA 6000 Pico; Agilent Technologies).

Preparation of CAGE Libraries, and Filtering and Mapping of CAGE Sequence Reads

CAGE libraries were prepared as in Takahashi et al. (2012) with a starting material of 5 μg of total RNA. The National High-Throughput DNA Sequencing Centre of the University of Copenhagen performed the sequencing on an Illumina HiSeq 2000 platform. As recommended by Illumina, 30% of Phi-X spike-ins were added to each sequencing lane to balance the low complexity of the 5′ ends of the CAGE libraries. The program FASTX Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit) was used to remove the linker sequences, retain the first 25 nucleotides from the 5′ end, and filter out sequences that fell below a Phil's Read Editor (Phred) score of Q30 in 50% of the remaining bases. The clean reads were mapped to the TAIR10 reference genome (Ensembl release 26) with the program Bowtie v1.1.2 (Langmead et al., 2009) using the parameters -t best -strata -v -k 10 -y -p 6 -phred33-quals -chunksmbs 512 -e 120 -q -un. Uniquely mapped CAGE tags 5′ ends were counted at each genomic position to obtain a set of CAGE TSSs, typically referred to as "CTSSs" in the CAGE literature. CTSS coordinates were offset by 1 bp to account for G-addition bias (Carninci et al., 2006).

Quantification and Clustering of CAGE TCs

Most analyses were performed with the R package CAGEfightR v1.2 (Thodberg et al., 2019b) in the Bioconductor environment (https://www.bioconductor.org). We kept only CTSSs with at least one count in at least three libraries (the smallest sample group size). CTSSs were normalized into TPMs to mitigate differences in sequencing depths. A pan-experiment set of CAGE TCs was established by computing the sum of TPM values over all libraries at each bp, followed by neighbor-clustering of CTSSs on the same strand (within a distance of 20 bp). CAGE TCs detected in the pan-experiment were then quantified individually for each sample (Supplemental Dataset 4). Only TCs supported in at least two libraries with a pan-experiment minimum of 1 TPM were considered for further analysis (Supplemental Dataset 1). TC peaks were defined as the position with maximum signal within the TC.

Annotation of CAGE TCs

TAIR10 annotation was recovered from the TxDb.Athaliana.BioMart.plantsmart28 Bioconductor package (https://www.bioconductor.org/packages/TxDb.Athaliana.BioMart.plantsmart28). ARAPORT11 GFF3 annotation (June 2016) was obtained from https://www.araport.org and converted into a TxDb object (Lawrence et al., 2013) using the makeTxDbFromGFF function from the R package GenomicFeatures (https://www.bioconductor.org/packages/GenomicFeatures). Using TC peaks as proxy (see above), annotation of CAGE TCs was conducted against a hierarchical annotation built from TAIR10 and ARAPORT11 individually (see Figure 1). The promoter region was defined as the 200-bp window centered on the annotated gene TSS. The promoter-proximal region extends 400 bp upstream of the annotated TSS. The same regions, but on the opposite strand, were referred to as the reverse region. The antisense category covered the whole gene body but on the opposite strand. The 5′-UTRs, 3′-UTRs, and CDS regions were defined as in TAIR10 and ARAPORT11. The exonic category corresponds to non-protein-coding exons (e.g., lncRNAs). In the case of multiple overlaps, a CAGE TC peak was annotated using the highest-priority category of the overlapping annotations in the annotation hierarchy.

Alternative TSS Analysis

CAGE TCs attributed to a TAIR10 annotated gene, and contributing at least 10% of the total TPM gene expression across the gene, were selected as a basis for assessing the landscape of alternative TSSs in unchallenged wild-type seedlings.

Recall of TSSs/TCs Across Datasets

TCs and TSSs from CAGE, PEAT-seq (Morton et al., 2014), and nanoPARE (Schon et al., 2018) datasets were merged. Overlapping regions were collapsed to obtain a nonredundant set of TSSs. The number of recalled TCs/TSSs from the nonredundant set was computed for each dataset and plotted as a Venn diagram (Figure 2A).

Average Meta-Profiles

Meta plots/profiles show the average signal from a given experiment across all genes of interest, where all genes are normalized to the same

length. All meta plots were computed on normalized read counts using the R package TeMPO (https://github.com/MalteThodberg/TeMPO) with a trimmed signal (0.01% to 0.99% percentiles).

Quantification and Comparison of TSSs Across Datasets

TSSs from ARAPORT11, TAIR10, nanoPARE (Schon et al., 2018), and PEAT-seq (Morton et al., 2014) were extended by 100 bp in both directions and quantified using the CAGE wild-type TPM signal. To account for the large differences in expression as observed in Figure 2B, TSSs were filtered to have at least one CAGE TPM in at least two samples and were subsequently used as anchors for comparison of nascent RNA signals, MNase signal, and sequence patterns.

Preparation of RNA-seq Libraries, and Mapping and Quantification of RNA-seq Reads

Total RNA was purified from complete and intact 14-d–old seedlings and polysaccharide precipitation was used as described above for the CAGE samples. The resulting polysaccharide-free total RNA was sequenced by Novogene as stranded Illumina paired-end 150-bp reads. Library preparation included an early rRNA depletion; in addition, the fragment size selection limit was lowered to 200 nt to allow for the capture of hypothetical PROMPT transcripts. Adapters were removed from the raw reads using the program Cutadapt v2.3 (Martin, 2011). Mapping to the TAIR10 reference genome (Ensembl release 26) was done with the software STAR v2.7.0 (Dobin and Gingeras, 2015) using default parameters, only considering unique mappers and concordant pairs. The matrix of counts antisense to TAIR10-annotated genes was generated with featureCounts v1.6.3 (Liao et al., 2014) with parameters -O–minoverlap 3–largestOverlap -B -p -s 1 (Supplemental Dataset 5).

Open-Chromatin Data and DHS Annotation

DHS regions, DNase-seq, and MNase-seq were downloaded from PlantDHS.org (Zhang et al., 2016; and see Supplemental Table 1 for details). Position of the highest DNase signal in flower tissue was used to identify the DHSS in each DHS region. DHSSs were extended 400 bp on both directions and annotated based on TAIR10, using the hierarchical strategy as defined in Figure 1C.

Poly(A), 5′, and 3′ Splice Sites at mRNA and PROMPT TSSs

We used the Bioconductor package seqPattern (https://www.bioconductor.org/packages/seqPattern) to scan for AWTAA as well as 5′ and 3′ splice sites. For AWTAA, the consensus sequence was used. Position frequency matrices from Brown et al. (1996) were used with a matching score threshold of 80% for the 5′ and 3′ splice sites.

Differential Expression Analysis, Overlap, and GO-Term Enrichment

We used the edgeR (Robinson et al., 2010) and limma (Ritchie et al., 2015) R packages to conduct differential expression analysis. Counts were normalized using the weighted-trimmed mean of M-values method (Robinson and Oshlack, 2010) and transformed using the voom method (Law et al., 2014) to model the mean-variance relationship before linear modeling, as recommended in Liu et al. (2015). Resulting P-values were corrected for multiple testing by the Benjamini–Hochberg method. Statistical significance of overlap between differentially expressed gene sets was assessed by Fisher's exact test using the testGeneOverlap function in the R package GeneOverlap (Figure 7A). For gene set enrichment analyses, we used the R package gProfileR (Reimand et al., 2007) with the complete set of detected genes as background.

Statistical Tests

For Supplemental Figure 4, *t* tests were conducted on log-transformed genomic distances using the t.test methods in R with settings alternative = "greater" and paired = FALSE. Wilcoxon tests on the untransformed distances were also conducted with the same parameters as detailed above. The complete detail on statistical tests is available in Supplemental Table 3.

Nascent RNA-seq and Support of PROMPTs

Nascent RNA-seq libraries from Hetzel et al. (2016) were reprocessed as in the original article. Briefly, adapter sequences were removed from the 3′-end of reads and mapping to the TAIR10 reference was conducted with STAR (v2.4.2a; Dobin and Gingeras, 2015). Only uniquely-mapped reads were retained, and the genome-wide signal was normalized per million mapped reads. For support of PROMPTs by nascent RNA-seq, we calculated the average signal of 5′ GRO-cap, as well as GRO-seq from two labs (Hetzel et al., 2016; Liu et al., 2018; see Supplemental Table 1 for details of libraries and tissues), across the genome-wide PROMPT regions (defined as the –400-bp antisense stretches from TAIR10-annotated TSSs). Support of exosome-sensitive PROMPT regions by nascent sequencings was considered upon showing 2-fold the average genome-wide signal.

Small RNA-seq Libraries and Analysis

Libraries were constructed from 1 μg of purified RNA using the NEBNext Multiplex Small RNA Library Prep Set (New England Biolabs) following the manufacturer's instructions. Sequencing was done on an Illumina NextSeq platform with 75-bp single-end reads. Adapters were removed from raw reads with the tool Cutadapt v2.3 (Martin, 2011) and filtered for a minimum and maximum length of 19 bp and 25 bp, respectively. Clean reads were mapped to the TAIR10 reference genome with the program STAR v2.7.0f (Dobin and Gingeras, 2015). To distinguish between small RNA read lengths, alignments from SAM output were separated into 21, 22, 23, and 24 nucleotides, using the Concise Idiosyncratic Gapped Alignment Report (CIGAR) matching length code. Raw read counts were normalized to the total number of clean reads.

Coverage at Potential PROMPT Regions

Potential PROMPT regions were defined as the 500-bp stretches upstream of annotated TSSs, not overlapping any other annotated features, independently of strand. mRNA regions to compare with were defined as the 500-bp downstream annotated TSSs. Small RNA-seq coverage was computed over potential PROMPT and mRNA regions. To normalize small RNA coverage to the RNA expression of respective regions, we divided the small RNA-seq coverage values either by the CAGE PROMPT TC expression, or by the sum of CAGE TCs annotated as "promoters" depending on the type of region.

De Novo Reconstruction of Antisense Transcripts

To assemble de novo transcripts from RNA-seq data, we used the program CuffLinks v2.2.1 with default parameters (Roberts et al., 2011), with the TAIR10 general feature format (GFF) annotation provided as a guide (option -g). The Cufflinks tool Cuffcompare was applied to compare the set of de novo transcripts across genotypes. De novo transcripts overlapping a reference exon or intron but on the opposite strand (codes "s" and "x," respectively), and associated with an RNA-seq upregulated gene on the reverse strand, were selected for assessing the length and splicing of exosome-sensitive transcripts.

Accession Numbers

CAGE, RNA-seq, and small RNA-seq libraries generated in this study are available in the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo under accession numbers GSE136356, GSE136362, and GSE142555, respectively. Details of tissues, growth conditions, and re-processed public datasets (nanoPARE TSSs, PEAT-seq TSSs, DNAse-seq, MNase-seq, GRO-seq, 5′GRO-cap, and ChIP-seq) are shown in Supplemental Table 1.

Supplemental Data

Supplemental Figure 1. Principal component analysis and characterization of PROMPT regions.

Supplemental Figure 2. Small RNAs signal at PROMPT regions.

Supplemental Figure 3. Sequence and expression properties of exosome-sensitive TCs.

Supplemental Figure 4. Investigation of CAGE TCs antisense to an annotated 3′-UTR.

Supplemental Table 1. Summary of public datasets used in this study.

Supplemental Table 2. Genotyping oligonucleotides.

Supplemental Table 3. Summary of statistical tests.

Supplemental Dataset 1. Unidirectional and bidirectional CAGE TCs, and PROMPTs.

Supplemental Dataset 2. Results of differential expression analyses for CAGE TCs, and RNA-seq at gene-level (on antisense strand).

Supplemental Dataset 3. RNA-seq de novo reconstructed transcripts. Due to size issues, this is deposited at FigShare: https://figshare.com/s/8afa8e534ea96989f8e5.

Supplemental Dataset 4. Expression matrix of CAGE TCs (in TPMs).

Supplemental Dataset 5. Expression matrix of RNA-seq (raw counts, antisense to TAIR10-annotated genes).

ACKNOWLEDGMENTS

We thank Kian Hématy (*rrp4-2*), Julio Salinas (*lsm8-2*), and Dominique Gagliardi (*hen2-4*) for providing mutant seeds. We thank Michael A. Schon for providing critical assessment of tag alignments. This work was supported by the Novo Nordisk Foundation (to A.S. and the Hallas Møller Stipend 2010 to P.B), the Lundbeck Foundation (to A.S.), and the Villum Foundation (grant 13397 to P.B.).

AUTHOR CONTRIBUTIONS

A.S. and P.B. designed the research; A.T. conducted experiments; J.B. constructed the CAGE libraries; M.L.V. constructed the small RNA-seq libraries; A.T. did the computational analyses; M.I. reprocessed the public ChIP-seq datasets; A.T. made figures; A.T., P.B., and A.S. interpreted results; A.T., A.S., and P.B. wrote the article with inputs from all authors.

Received October 17, 2019; revised February 3, 2020; accepted March 20, 2020; published March 25, 2020.

REFERENCES

Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project. (2009). Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457: 1028–1032.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature **499:** 360–363.

Andersson, R., et al. (2014a). An atlas of active enhancers across human cell types and tissues. Nature **507:** 455–461.

Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014b). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat. Commun. **5:** 5336.

Andersson, R., Chen, Y., Core, L., Lis, J.T., Sandelin, A., and Jensen, T.H. (2015a). Human gene promoters are intrinsically bidirectional. Mol. Cell **60:** 346–347.

Andersson, R., Sandelin, A., and Danko, C.G. (2015b). A unified architecture of transcriptional regulatory elements. Trends Genet. **31:** 426–433.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. **21:** 71–87.

Asif, M.H., Dhawan, P., and Nath, P. (2000). A simple procedure for the isolation of high quality RNA from ripening banana fruit. Plant Mol. Biol. Report. **18:** 109–115.

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis **53:** 474–485.

Bornholdt, J., et al. (2017). Identification of gene transcription start sites and enhancers responding to pulmonary carbon nanotube exposure in vivo. ACS Nano **11:** 3597–3613.

Bouveret, E., Rigaut, G., Shevchenko, A., Wilm, M., and Séraphin, B. (2000). A Sm-like protein complex that participates in mRNA degradation. EMBO J. **19:** 1661–1671.

Boyd, M., et al. (2018). Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. Nat. Commun. **9:** 1661.

Branscheid, A., Marchais, A., Schott, G., Lange, H., Gagliardi, D., Andersen, S.U., Voinnet, O., and Brodersen, P. (2015). SKI2 mediates degradation of RISC 5′-cleavage fragments and prevents secondary siRNA production from miRNA targets in Arabidopsis. Nucleic Acids Res. **43:** 10975–10988.

Brown, J.W., Smith, P., and Simpson, C.G. (1996). Arabidopsis consensus intron sequences. Plant Mol. Biol. **32:** 531–535.

Carninci, P., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. **38:** 626–635.

Chekanova, J.A., et al. (2007). Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the Arabidopsis transcriptome. Cell **131:** 1340–1353.

Chen, C., et al. (2017). Cytosolic acetyl-CoA promotes histone acetylation predominantly at H3K27 in Arabidopsis. Nat. Plants **3:** 814–824.

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.

Chlebowski, A., Lubas, M., Jensen, T.H., and Dziembowski, A. (2013). RNA decay machines: The exosome. Biochim. Biophys. Acta **1829:** 552–560.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified

architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. **46:** 1311–1320.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science **322:** 1845–1848.

Cortijo, S., Charoensawan, V., Brestovitsky, A., Buning, R., Ravarani, C., Rhodes, D., van Noort, J., Jaeger, K.E., and Wigge, P.A. (2017). Transcriptional regulation of the ambient temperature response by H2A.Z nucleosomes and HSF1 transcription factors in Arabidopsis. Mol. Plant **10:** 1258–1273.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H.-M. (2008). The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. **24:** 167–177.

Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq reads with STAR. Curr. Protoc. Bioinformatics **51:** 11.14.1-11.14.19.

Dunoyer, P., Himber, C., and Voinnet, O. (2005). DICER-LIKE 4 is required for RNA interference and produces the 21-nucleotide small interfering RNA component of the plant cell-to-cell silencing signal. Nat. Genet. **37:** 1356–1360.

Erhard, K.F., Jr., Talbot, J.-E.R.B., Deans, N.C., McClish, A.E., and Hollick, J.B. (2015). Nascent transcription affected by RNA polymerase IV in *Zea mays*. Genetics **199:** 1107–1125.

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. USA **113:** E7846–E7855.

Golisz, A., Sikorski, P.J., Kruszka, K., and Kufel, J. (2013). *Arabidopsis thaliana* LSM proteins function in mRNA splicing and degradation. Nucleic Acids Res. **41:** 6232–6249.

Hématy, K., et al. (2016). The zinc-finger protein SOP1 is required for a subset of the nuclear exosome functions in Arabidopsis. PLoS Genet. **12:** e1005817.

Henriques, R., Wang, H., Liu, J., Boix, M., Huang, L.-F., and Chua, N.-H. (2017). The antiphasic regulatory module comprising CDF5 and its antisense RNA FLORE links the circadian clock to photoperiodic flowering. New Phytol. **216:** 854–867.

Hetzel, J., Duttke, S.H., Benner, C., and Chory, J. (2016). Nascent RNA sequencing reveals distinct features in plant transcription. Proc. Natl. Acad. Sci. USA **113:** 12316–12321.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol. Cell **52:** 473–484.

Kilchert, C., Wittmann, S., and Vasiljeva, L. (2016). The regulation and functions of the nuclear RNA exosome complex. Nat. Rev. Mol. Cell Biol. **17:** 227–239.

Kim, T.-K., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature **465:** 182–187.

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

Krzyszton, M., Zakrzewska-Placzek, M., Kwasnik, A., Dojer, N., Karlowski, W., and Kufel, J. (2018). Defective XRN3-mediated transcription termination in Arabidopsis affects the expression of protein-coding genes. Plant J. **93:** 1017–1031.

Kurihara, Y., Makita, Y., Kawashima, M., Fujita, T., Iwasaki, S., and Matsui, M. (2018). Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in Arabidopsis. Proc. Natl. Acad. Sci. USA **115:** 7831–7836.

Lange, H., et al. (2014). The RNA helicases AtMTR4 and HEN2 target specific subsets of nuclear transcripts for degradation by the nuclear exosome in *Arabidopsis thaliana*. PLoS Genet. **10:** e1004564.

Lange, H., Sement, F.M., and Gagliardi, D. (2011). MTR4, a putative RNA helicase and exosome co-factor, is required for proper rRNA biogenesis and development in *Arabidopsis thaliana*. Plant J. **68:** 51–63.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:** R25.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. **15:** R29.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. **11:** 204–220.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLOS Comput. Biol. **9:** e1003118.

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.

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

Liu, R., Holik, A.Z., Su, S., Jansz, N., Chen, K., Leong, H.S., Blewitt, M.E., Asselin-Labat, M.-L., and Smyth, G.K., et al. (2015). Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. Nucleic Acids Res. **43:** e97.

Liu, W., Duttke, S.H., Hetzel, J., Groth, M., Feng, S., Gallego-Bartolome, J., Zhong, Z., Kuo, H.Y., Wang, Z., and Zhai, J., et al. (2018). RNA-directed DNA methylation involves co-transcriptional small-RNA-guided slicing of polymerase V transcripts in Arabidopsis. Nat. Plants **4:** 181–188.

Li, X.-Q., and Du, D. (2014). Motif types, motif locations and base composition patterns around the RNA polyadenylation site in microorganisms, plants and animals. BMC Evol. Biol. **14:** 162.

Lozano, R., Booth, G.T., Omar, B.Y., Li, B., Buckler, E.S., Lis, J.T., Jannink, J.-L., and Del Carpio, D.P. (2018). RNA polymerase mapping in plants identifies enhancers enriched in causal variants. bioRxiv 376640.

Lykke-Andersen, S., Brodersen, D.E., and Jensen, T.H. (2009). Origins and activities of the eukaryotic exosome. J. Cell Sci. **122:** 1487–1494.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. Nat. Genet. **45:** 1029–1039.

Marquardt, S., Escalante-Chong, R., Pho, N., Wang, J., Churchman, L.S., Springer, M., and Buratowski, S. (2014). A chromatin-based mechanism for limiting divergent noncoding transcription. Cell **158:** 462.

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

McCue, A.D., Panda, K., Nuthikattu, S., Choudury, S.G., Thomas, E.N., and Slotkin, R.K. (2015). ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. EMBO J. **34:** 20–35.

Meers, M.P., Adelman, K., Duronio, R.J., Strahl, B.D., McKay, D.J., and Matera, A.G. (2018). Transcription start site profiling uncovers divergent transcription and enhancer-associated RNAs in *Drosophila melanogaster*. BMC Genomics **19:** 157.

Meola, N., et al. (2016). Identification of a nuclear exosome decay pathway for processed transcripts. Mol. Cell **64:** 520–533.

Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A., Benfey, P.N., Ohler, U., and Megraw, M. (2014). Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. Plant Cell **26:** 2746–2760.

Nielsen, M., Ard, R., Leng, X., Ivanov, M., Kindgren, P., Pelechano, V., and Marquardt, S. (2019). Transcription-driven chromatin repression of Intragenic transcription start sites. PLoS Genet. **15:** e1007969.

Ntini, E., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat. Struct. Mol. Biol. **20:** 923–928.

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L.C., Dahmane, N., and Davuluri, R.V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. Genome Res. **21:** 1260–1272.

Perea-Resa, C., Hernández-Verdeja, T., López-Cobollo, R., del Mar Castellano, M., and Salinas, J. (2012). LSM proteins provide accurate splicing and decay of selected transcripts to ensure normal Arabidopsis development. Plant Cell **24:** 4930–4947.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science **322:** 1851–1854.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. **35:** W193–W200.

Rennie, S., Dalby, M., Lloret-Llinares, M., Bakoulis, S., Dalager Vaagensø, C., Heick Jensen, T., and Andersson, R. (2018). Transcription start site analysis reveals widespread divergent transcription in *D. melanogaster* and core promoter-encoded enhancer activities. Nucleic Acids Res. **46:** 5455–5469.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43:** e47.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-seq. Bioinformatics **27:** 2325–2329.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26:** 139–140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. **11:** R25.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: A new way to display consensus sequences. Nucleic Acids Res. **18:** 6097–6100.

Schon, M.A., Kellner, M.J., Plotnikova, A., Hofmann, F., and Nodine, M.D. (2018). NanoPARE: Parallel analysis of RNA 5′ ends from low-input RNA. Genome Res. **28:** 1931–1942.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science **322:** 1849–1851.

Shetty, A., Kallgren, S.P., Demel, C., Maier, K.C., Spatt, D., Alver, B.H., Cramer, P., Park, P.J., and Winston, F. (2017). Spt5 plays vital roles in the control of sense and antisense transcription elongation. Mol. Cell **66:** 77–88.e5.

Singh, J., Mishra, V., Wang, F., Huang, H.-Y., and Pikaard, C.S. (2019). Reaction mechanisms of Pol IV, RDR2, and DCL3 drive RNA channeling in the siRNA-directed DNA methylation pathway. Mol. Cell **75:** 576–589.e5.

Smith, S., and Stillman, B. (1989). Purification and characterization of CAF-I, a human cell factor required for chromatin assembly during DNA replication in vitro. Cell **58:** 15–25.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat. Protoc. **7:** 542–561.

Tharun, S., He, W., Mayes, A.E., Lennertz, P., Beggs, J.D., and Parker, R. (2000). Yeast Sm-like proteins function in mRNA decapping and decay. Nature **404:** 515–518.

Thodberg, M., and Sandelin, A. (2019). A step-by-step guide to analyzing CAGE data using R/Bioconductor. F1000 Res. **8:** 886.

Thodberg, M., et al. (2019a). Comprehensive profiling of the fission yeast transcription start site activity during stress and media response. Nucleic Acids Res. **47:** 1671–1691.

Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R., and Sandelin, A. (2019b). CAGEfightR: Analysis of 5′-end data using R/Bioconductor. BMC Bioinformatics **20:** 487.

Tokizawa, M., Kusunoki, K., Koyama, H., Kurotani, A., Sakurai, T., Suzuki, Y., Sakamoto, T., Kurata, T., and Yamamoto, Y.Y. (2017). Identification of Arabidopsis genic and non-genic promoters by paired-end sequencing of TSS tags. Plant J. **90:** 587–605.

Ushijima, T., et al. (2017). Light controls protein localization through phytochrome-mediated alternative promoter selection. Cell **171:** 1316–1325.e12.

Valen, E., et al. (2009). Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res. **19:** 255–265.

van Dijk, K., et al. (2010). Dynamic changes in genome-wide histone H3 lysine 4 methylation patterns in response to dehydration stress in *Arabidopsis thaliana*. BMC Plant Biol. **10:** 238.

Wang, Z., Casas-Mollano, J.A., Xu, J., Riethoven, J.-J.M., Zhang, C., and Cerutti, H. (2015). Osmotic stress induces phosphorylation of histone H3 at threonine 3 in pericentromeric regions of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **112:** 8487–8492.

Wu, H., Nord, A.S., Akiyama, J.A., Shoukry, M., Afzal, V., Rubin, E.M., Pennacchio, L.A., and Visel, A. (2014). Tissue-specific RNA expression marks distant-acting developmental enhancers. PLoS Genet. **10:** e1004610.

Xie, Z., Allen, E., Wilken, A., and Carrington, J.C. (2005). DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **102:** 12984–12989.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. **2:** E104.

Yamamoto, Y.Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., and Obokata, J. (2009). Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. Plant J. **60:** 350–362.

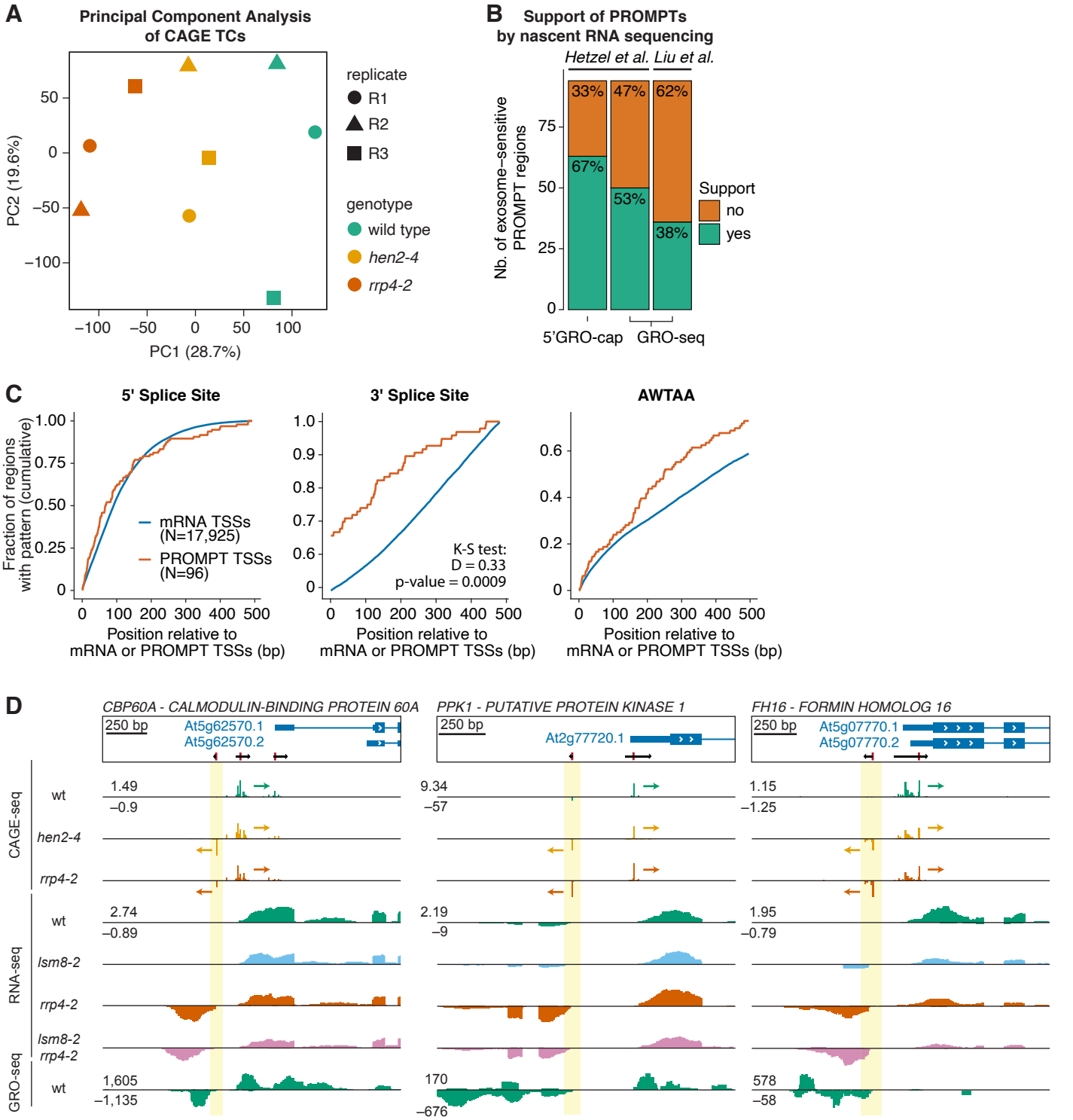
Zakrzewska-Placzek, M., Souret, F.F., Sobczyk, G.J., Green, P.J., and Kufel, J. (2010). *Arabidopsis thaliana* XRN2 is required for primary cleavage in the pre-ribosomal RNA. Nucleic Acids Res. **38:** 4487–4502.

Zhang, T., Marand, A.P., and Jiang, J. (2016). PlantDHS: A database for DNase I hypersensitive sites in plants. Nucleic Acids Res. **44** (D1): D1148–D1153.

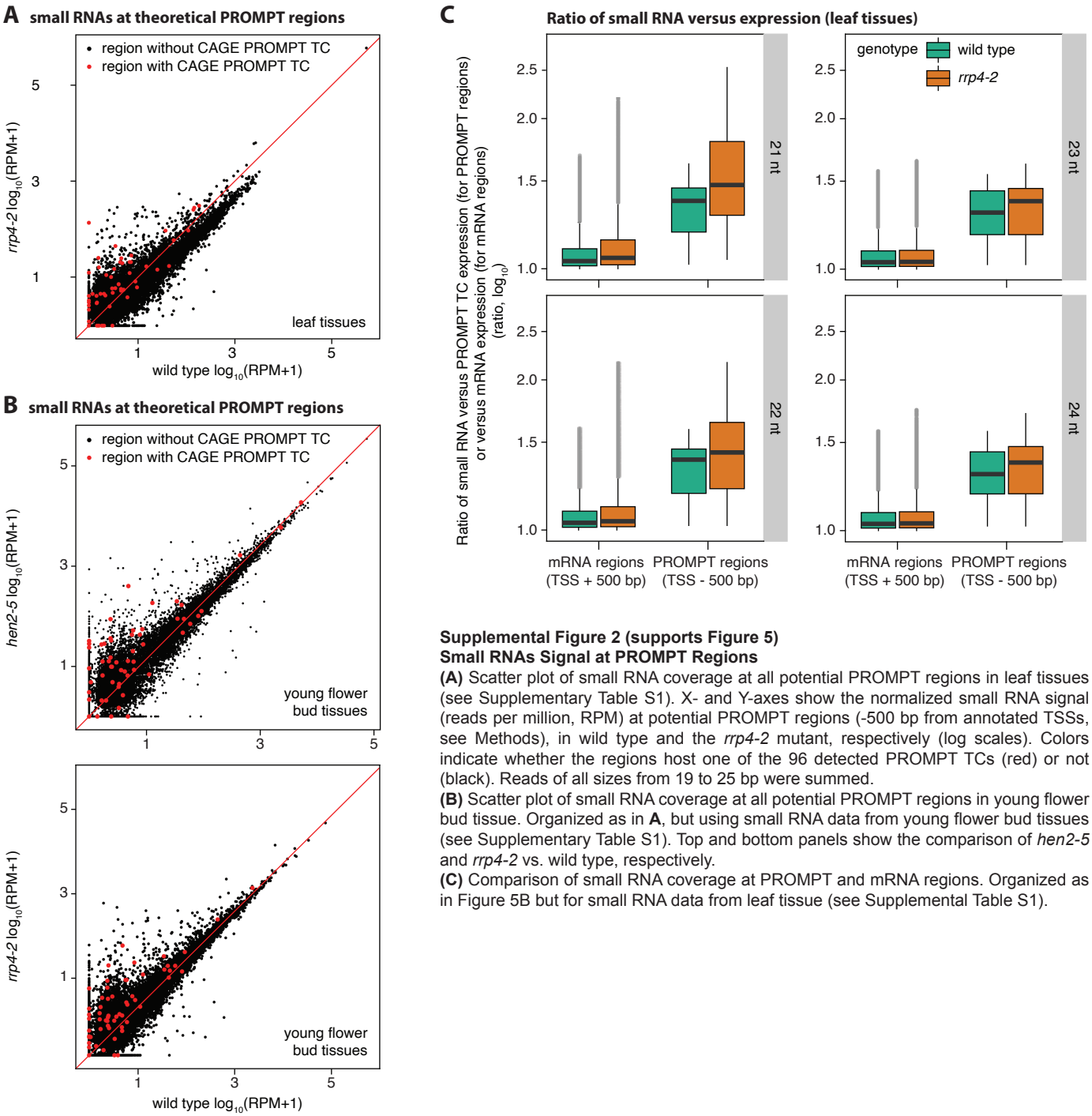
Zhang, T., Zhang, W., and Jiang, J. (2015). Genome-wide nucleosome occupancy and positioning and their impact on gene expression and evolution in plants. Plant Physiol. **168:** 1406–1416.

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

Zhu, J., Liu, M., Liu, X., and Dong, Z. (2018). RNA polymerase II activity revealed by GRO-seq and pNET-seq in Arabidopsis. Nat. Plants **4:** 1112–1123.



Supplemental Figure 1 (supports Figure 4)
Principal Component Analysis and Characterization of PROMPT Regions.
(A) Principal component analysis of CAGE TCs. Axes show the two first principal components and the percentage of explained variance. Each point represents a CAGE library, whose genotype and replicate are indicated by colors and shapes, respectively.
(B) Support of PROMPTs by nascent RNA sequencing data. X-axis shows 5' GRO-cap and GRO-seq from two laboratories, as indicated on top. Y-axis counts the number of exosome-sensitive PROMPT regions detected in this study, colored by whether they were supported by nascent transcription data or not (see Methods).
(C) AWTAA poly(A) and splice sites at mRNA and PROMPT TSSs. X-axis shows distance relative to CAGE TC peaks associated with either mRNA promoters or exosome-sensitive PROMPTs, in bp. Y-axis shows the fraction of regions with a 5' or 3' splice site (left and middle) or AWTAA poly(A) (right) pattern.
(D) Three additionalcases of bidirectionally transcribed promoters of protein-coding genes, organized as in Figure 4C.

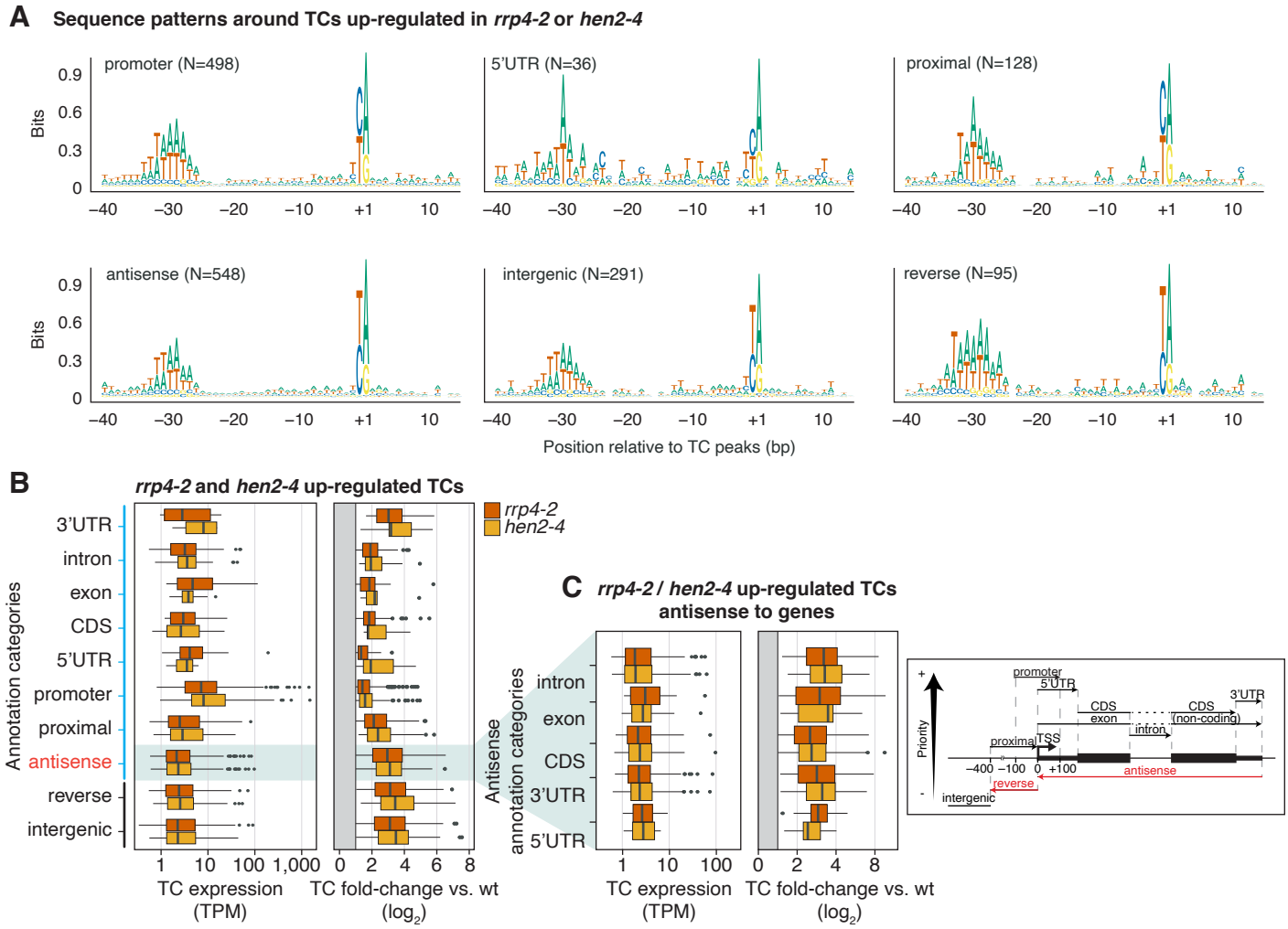


Supplemental Figure 2 (supports Figure 5)
Small RNAs Signal at PROMPT Regions

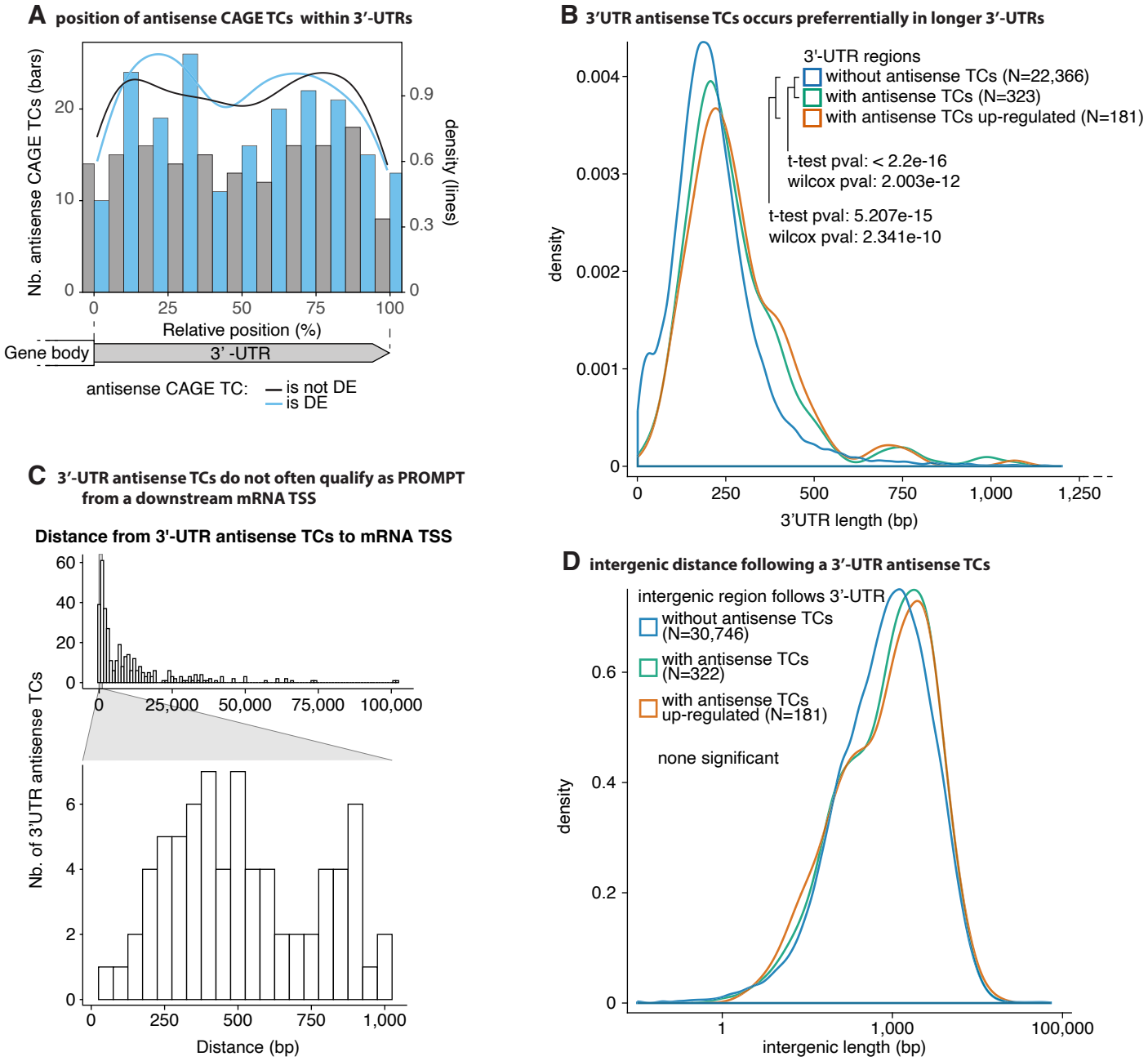
(A) Scatter plot of small RNA coverage at all potential PROMPT regions in leaf tissues (see Supplementary Table S1). X- and Y-axes show the normalized small RNA signal (reads per million, RPM) at potential PROMPT regions (-500 bp from annotated TSSs, see Methods), in wild type and the *rrp4-2* mutant, respectively (log scales). Colors indicate whether the regions host one of the 96 detected PROMPT TCs (red) or not (black). Reads of all sizes from 19 to 25 bp were summed.

(B) Scatter plot of small RNA coverage at all potential PROMPT regions in young flower bud tissue. Organized as in **A**, but using small RNA data from young flower bud tissues (see Supplementary Table S1). Top and bottom panels show the comparison of *hen2-5* and *rrp4-2* vs. wild type, respectively.

(C) Comparison of small RNA coverage at PROMPT and mRNA regions. Organized as in Figure 5B but for small RNA data from leaf tissue (see Supplemental Table S1).



Supplemental Figure 3 (supports Figure 8)
Sequence and Expression Properties of Exosome-Sensitive TCs
(A) Sequence patterns of CAGE TCs up-regulated in *hen2-4* or *rrp4-2*. Organized as in Figure 3C. The annotation hierarchy strategy is recapitulated on the right side of **C**.
(B) Annotation, expression, and fold-change of up-regulated CAGE TCs. Left panel: X-axis shows the expression of CAGE TCs in TPM. Y-axis distinguishes annotation categories based on TAIR10. Colors separate CAGE TCs up-regulated either in *rrp4-2* vs. wild type (orange) or *hen2-4* vs. wild type (yellow). Right panel: X-axis shows $\log_2$ fold-change of respective mutant vs. wild type.
(C) Expression and fold-change of up-regulated CAGE TCs, antisense to a gene. Left and right panel organized as in **B**, but shows expression of CAGE TCs that are antisense to an annotated TAIR10 gene.



Supplemental Figure 4 (supports Figure 9)
Investigation of CAGE TCs Antisense to an Annotated 3'-UTR

(A) Positioning of antisense TCs within 3'-UTRs. X-axis show position within 3'-UTR, scaled from 0 (start of UTR) to 3'-end (100), in percent. Lines show the density of TCs along X-axis, bar overlay shows the numbers. Colors indicate whether TCs are up-regulated (is DE) in *rrp4-2 vs.* wild type or not (not DE).

(B) Relation between 3'-UTR length and presence of antisense TCs. Density distribution of 3'-UTR lengths, where colors indicate the presence of antisense TCs and their up-regulation status. Detail of statistical tests are available in Supplemental Table S3.

(C) 3'-UTR-antisense TCs are not PROMPTs of neighboring promoters. X-axis shows the distance between 3'-UTR-antisense TC peaks and the closest annotated mRNA TSS, located downstream of the PROMPT and on the opposite strand, in bp. Y-axis indicates the number of 3'-UTR-antisense TCs. Bottom insert zooms on distances below 1000 bp.

(D) Relation between intergenic gene distance and presence of antisense TCs. Density distribution of lengths intergenic regions (to next annotated gene) downstream of an annotated 3'-UTR. Colors indicate the presence of antisense TC in the 3'-UTR and up-regulation status.

Supplemental Table 1 – Detail of public dataset used in this study

Author & PMID	Tissues & conditions	Comment	Link to data
GRO-seq studies			
Hetzel et al., 2016 (PMID: 27729530)	6-day old whole seedlings, LS media, 22°C, constant light	Raw reads data were re-processed following the methods described in Hetzel et al., 2016. Adaptors were trimmed from the 3'-end of reads and subsequently mapped using STAR with default parameters. Only uniquely-mapped reads were retained.	Raw data: GSE83108 Re-processed data available on FigShare
Liu et al., 2018 (PMID: 29379150)	Inflorescence tissues grown in soil with 16h/8h light cycle	Processed data were obtained directly from the GEO accessions (bigWig format), see link to data.	GSE108078 GSE100010
5'GRO-cap study			
Hetzel et al., 2016 (PMID: 27729530)	6-day old whole seedlings, LS media, 22°C, constant light	Raw reads were re-analyzed as above, following Hetzel et al, 2016 methods.	GSE83108
TSS studies			
Morton et al., 2014 (PMID: 25035402)	Pool of 7-day old seedling roots grown on full-strength MS supplemented with 1% sucrose at 22°C with 16h/8h light cycle	PEAT-seq dataset (as provided by Schon et al., 2018 re-analysis, see below)	Supplemental Code S1 from Schon et al., 2018 (data tables, PEAK_peaks.bed)
Schon et al., 2018 (PMID: 30355603)	Buds from flowering time, grown at 20-22°C with 16h/8h light cycle	nanoPARE dataset (and re-analysis of PEAT-seq from Morton et al., 2014)	Supplemental Code S1, data tables, fb.W.5p.bed
Open Chromatin study (MNase, DNase-I, and DHSs)			
Zhang et al., 2012 (PMID: 22773751)	Leaf and flower (closed-buds) from 2-week old seedlings, grown in full-strength MS supplemented with 0.5% sucrose at 23°C with 16h/8h light cycle	Zhang et al., 2016 (PMID: 26400163) reprocessed all identified DHSs into a unified dataset for plantDHS.org. This unified dataset has been used for DHS-based analyses in this study.	DHS regions and DNase-I available from plantDHS.org
Zhang et al., 2015 (PMID: 26143253)	Leaf and flower tissues from 2-week old seedlings grown in full-strength MS supplemented with 0.5% sucrose	Processed data were obtained directly from plantDHS.org (bigWig format), see link to data.	MNase-I available from plantDHS.org
ChIP-seq studies			
Cortijo et al., 2017 (PMID: 28893714)	10-day old whole seedlings grown at 17°C in full-strength MS supplemented with 1% sucrose	Original data were re-processed by Nielsen et al., 2019 (PMID: 30707695). Remapping procedure is described in the "Bioinformatics analysis" section in Methods. Code for remapping is available on GitHub (relevant scripts: *> 04-Remapping_Paired-End_ChIP-Seq_and_MNase-Seq.sh* *> 05-Remapping_Single-End_ChIP-Seq.sh*).	GSE79355 RNA Polymerase II
Chen et al., 2017 (PMID: 28947800)	12-day old seedlings, full-strength MS supplemented with 0.5% sucrose at 23°C with 16h/8h light cycle		GSE79524 H3K27ac
Van Dijk et al., 2010 (PMID: 21050490)	4-week old leaves grown in soil with 12h/12h light cycle		GSE11657 H3K4me1 & me3
Wang et al., 2015 (PMID: 26100864)	3-week old plants grown on full-strength MS supplemented with 0.5% sucrose with 16h/8h light cycle		GSE68370 H3K4me1 & me3

Supplemental Table 2 – Genotyping primers

Genotype	AGI	Left primer 5'-3'	Right primer 5'-3'	Comment
hen2-4 (sop1-4)	AT2G06990	TATGGTATTCAGCAACCTCCG	GTTCCTCAAATGCTGCTCTTG	SALK_091606 genotyping
hen2-5 (sop1-5)	AT1G21580	GACTTGTGAAAGCGCTTTTTG	TATGGTATTCAGCAACCTCCG	SALK_019457 genotyping
lsm8-2	AT1G65700	ACTAACTGGCCTCTGAATGGAAG	AAGAAGACCCAAGACTCCGATG	SALK_048010 genotyping
rrp4-2 (sop2-1)	AT1G03360	CTATTCCCGTCAACCATGACG	CATCGACCTCGGAAGTTCCAGGT	DNA amplification before enzymatic digestion
/	/	ATTTTGCCGATTTCGGAAC	/	SALK Right Border primer LBb1.3

Supplemental Table 3 – Detail of statistical tests for Supplemental Figure 4B.
Statistical tests were conducted within the R framework with the following functions and parameters:

T-tests:

```
t.test(x=log(values_1), y=log(values_2), alternative="greater", paired=FALSE)
```

Wilcoxon tests:

```
wilcox.test(x=values_1, y=values_2, alternative="greater", paired=FALSE)
```

Because the data were not normally distributed, the log-transformed values were used for t-tests, whereas the untransformed values were used directly for Wilcoxon tests.

Detail of results when comparing 3'UTR regions with antisense TCs (green) to those without antisense TCs (blue):

T-test:

Data	mean	t-value	p-value	95% CI
log(lengths of 3' UTRs having antisense TC)	5.473			
log(lengths of 3' UTRs having no antisense TC)	5.189	9.312 (df = 340.75)	<2.2E-16	0.234 – Inf

alternative hypothesis: true difference in means is greater than 0

Wilcoxon-test:

Data	W	p-value
lengths of 3' UTRs having no antisense TC		
lengths of 3' UTRs having no antisense TC	4,434,200	2.003E-12

Detail of results when comparing 3'UTR regions with antisense TCs up-regulated (orange,) to those without antisense TCs (blue):

T-test:

Data	mean	t-value	p-value	95% CI
log(lengths of 3' UTRs having up-regulated antisense TC)	5.510			
log(lengths of 3' UTRs having no antisense TC)	5.189	8.410 (df = 186.59)	5.207E-15	0.258 – Inf

Wilcoxon-test:

Data	W	p-value
lengths of 3' UTRs having up-regulated antisense TC		
lengths of 3' UTRs having no antisense TC	2,576,800	2.341E-10