Supplemental Information for

Novel mRNAs 3′ end-associated *cis*-regulatory elements with epigenomic signatures of mammalian enhancers in the *Arabidopsis* genome

Hsiao-Lin V. Wang and Julia A. Chekanova*

*To whom correspondence should be addressed. Tel: 1-816-516-7992; Email: chekanovaj@gmail.com.

This document includes:

SI Text

- Identification of putative enhancer-like elements (PEs) in *Arabidopsis*
- Analysis of RNA Pol IV and V occupancy in the PE regions
- Analysis of EASE-like module
- Epigenomic signatures of PEs and eILs
- eILs, the regions targeted by PEs, are conserved in 1135 *Arabidopsis* accessions
- Analysis of the genome-wide occupancy of H3K4 monomethylation confirmed that H3K4me1 occupies less than a third of all protein-coding genes
- Analysis of the chromatin signatures at the 3′ ends of protein-coding genes that do and do not house PEs
- Analysis of the enhancer trap insertion sites that are in close vicinity to PEs
- Analysis of convergent gene pairs and PEs

Figures S1 to S15

Captions for Datasets S1 to S14

SI Methods

SI References

SI TEXT

Identification of putative enhancer-like elements in *Arabidopsis*

The scheme designed to identify and annotate PEs in *Arabidopsis thaliana* genome version TAIR10 is described in **Fig. 1B**. This scheme included the following steps:

Step 1: The entire *Arabidopsis thaliana* genome (~135 Mb) was compared with previously published open chromatin data (Zhang et al. 2012), which mapped DNase I hypersensitive sites (DHSs) genome-wide using short-read sequencing (DNase I-seq). The DNase I-seq data for both wild-type whole seedlings (2 weeks old) and inflorescence tissues were published previously and deposited under accession number GSE34318 (Zhang et al. 2012). The file ″GSE34318_dhsites_region.bed″ in GEO GSE34318 series contains all identified DHSs (including data generating using whole seedlings and flowers). Only DHSs labeled with ″wtleaf″, which represent the open chromatin regions genome-wide in whole seedlings, were extracted (38,290 DHSs) (**Fig. 1B** step 1). The DHSs published by Zhang *et al.* had an average length of 389 nt, with a minimum of 11 nt and a maximum of 2188 nt. To analyze the methylation status of histone H3K4 in neighboring chromatin, an additional 1 kb up- and down-stream from each DHS was also analyzed (the average length of the extended DHSs is 2389 nt).

Step 2: The extended DHSs were compared with peaks enriched for H3K4 mono-methylation (identified using published H3K4 mono-methylation ChIP-chip data (Zhang et al. 2009)). The step-2 analysis identified 6001 DHSs that co-localize with peaks enriched for H3K4 mono-methylation (**Fig. 1B** step 2). Fisher's exact test was used to test the statistical significance of data associations with a set of randomized control; the genomic interval based BEDtools Fisher version 2.24 (Quinlan and Hal 2010), which compared the simulated random set of similar genomic regions (control, TAIR10 genome-based) versus the observed enrichment, was used.

Step 3: The ratio of mono- to tri-methylation of H3K4 (Zhang et al. 2009) was calculated for each of these 6001 DHSs. The HMM posterior probability scores of mono- and tri-methylation of H3K4, which represent the signal intensities per specific histone modification, were tabulated for each of these 6001 DHSs. Then the ratio of mono- to tri-methylation of H3K4 in each DHS was calculated, and only 1919 DHSs had a ratio of mono- to tri-methylation > 2. To analyze the resulting 1919 DHSs uniformly, all the DHSs were converted to 3-kb in length

by adding 1.5-kb up- and down-stream from the midpoints of identified DHSs. The resulting 3-kb long DHSs were termed **Putative enhancer-like elements** (**PEs**) and analyzed further (**Fig. 1B** step 3, and **Dataset S1**).

The identified PEs were subjected to meta-gene analysis to profile the distribution of chromatin modifications across the PE regions (**Fig. 1D)**. The results of the analysis were also depicted as heatmaps, and the enrichment and distribution data are shown for H3K4 methylation (mono-, di-, and tri-), acetylation and trimethylation of H3K27, RNA Pol II occupancy, and the transcriptional activity as both poly(A)+ and ribo-minus RNA expression (**SI Appendix, Fig. S2**). All PEs had a greater Pol II posterior probability scores compared with the genome-wide median (the median of the HMM posterior probability scores per 35-nt bin is 2e-6). The HMM posterior probability scores of acetylation and trimethylation of H3K27 were tabulated for each of 1919 PEs, and 1461 PEs had a greater H3K27ac posterior probability scores compared to H3K27me3.

Analysis of RNA Pol IV and V occupancy in the PE regions

Arabidopsis has two additional RNA polymerases, IV and V, which are involved in directing repressive epigenetic modification through RNA-mediated transcriptional silencing (RNA-directed DNA methylation pathway, RdDM). To determine if PEs are associated with Pol IV and Pol V loci, the ChIP-seq datasets of previously determined RNA Pol IV (Law et al. 2013) and Pol V (Wierzbicki et al. 2012) peaks were extracted and analyzed. The RNA Pol IV peaks were annotated with *Arabidopsis* TAIR8 genome version (928 peaks, Supplemental Table 6 by Law *et al.* (Law et al. 2013)), and the genomic coordinates were converted to the TAIR9 genome version with Perl scripts provided by TAIR (The Arabidopsis Genome Initiative 2000) (update_coordinates.pl with file: tair9.ns.indels.subs.assembly_updates.txt). The high-confidence Pol V peaks were previously identified by Wierzbicki *et al.* (Wierzbicki et al. 2012) and a total of 1157 peaks was provided in the previously published Supplemental Table 1. BEDTools version 2.24 (Quinlan and Hal 2010) was used to determine if there were any intersection between identified PE regions and RNA Pol IV and Pol V peaks. The statistical significance was tested using Fisher's exact test (implemented in BEDTools Fisher function, version 2.24 (Quinlan and Hal 2010)). We found that very few PEs are associated with previously determined/identified Pol IV and Pol V loci (Pol IV: 1.7%, p = 5e-3; Pol V: 4%, p = 0.95, statistically insignificant; Fisher's exact test). This result is in agreement with the knowledge that Pol IV/V are generally involved in silencing transposons and heterochromatin

regions, while transcriptional enhancers are located in open chromatin regions. Also, as shown in **SI Appendix, Fig. S14B**, 5mC (DNA methylation status) is largely absent in PE regions, in contrast to transposon and heterochromatin (**SI Appendix, Fig. S14A**).

Analysis of EASE-like module

The egg apparatus-specific enhancer (EASE), which was identified by Yang *et al.* (Yang et al. 2005) from enhancer trap lines that were generated in the *Arabidopsis* accession Landsberg *erecta* (L*er*) by Sundaresan et al. 1995. The 77-bp EASE module (GenBank accession no. AX100536, ccacgatgcaaatatatcgataacgttattaaaaaaagtaaccgcatgatatattctctttcgtatgatattaaggc) was isolated and identified from the *Arabidopsis* enhancer-detector line ET253, which shows spatially restricted GUS expression mainly in the egg apparatus. EASE is A/T rich (g or c: 24-nt (31%); a or t: 53-nt (69%); p = 0.0009502, chi-square test). EASE is located in the intergenic regions on chromosome 4 (17638139-17638215, based on BLAST against TAIR10 *Arabidopsis* genome) between AT4G37530 (Peroxidase superfamily protein, ATP37) and AT4G37540 (encoding LBD39, LOB DOMAIN-CONTAINING PROTEIN 39) (Yang et al. 2005; The Arabidopsis Genome Initiative 2000; Krishnakumar et al. 2015). There is also an *RC/Helitron* transposable element (AT4TE85180, ATREP10B) located downstream.

To characterize the chromatin signatures (H3K4me1, H3K4me3, H3K27ac, and H3K27me3) at the EASE locus, the neighboring chromatin was also considered by examining regions ± 200-bp from the borders of EASE locus. BEDTools version 2.24 (Quinlan and Hal 2010) was used to examine the HMM posterior probability (as a measurement of signal intensity) for each ChIP-chip dataset (described above), and the results were analyzed for statistical significance (**Fig. 5A–B**; two-tailed paired t-test: p = 0.03641 for H3K4me1 v.s. H3K4me3, p = 0.43269 for H3K27ac v.s. H3K27me3). An Integrated Genomic Viewer (IGV, version 2.3.67) visualization of the EASE locus and the 200 bp flanking region is shown in **Fig. 5C**.

FIMO (Find Individual Motif Occurrences) (Grant et al. 2011) was used to scan for EASE-like modules within the entire length of PE regions. The forward and reverse sequences of all PE regions were scanned for EASE 77-bp sequences and this identified 3125 sites carrying statistically significant EASE-like modules in the entire population of PE regions (statistical threshold, $p < 0.0001$). Also, 77% (n=1468; p = 9.232e-15, chi-

square test) of total 1919 PEs have at least one EASE-like module. Out of the 1468 EASE-like module-containing PEs, 61% of them (n=899; $p < 0.0001$, chi-square test) have two or more EASE-like modules (39% have one, 31% have two, and 30% have 3-11 EASE-like per PE region).

Similar searches for EASE-like modules were done on regions flanking the TESs (± 1.5-kb) of all protein-coding genes in the *Arabidopsis thaliana* genome and 53% of these TESs (n = 15,146, $p < 0.0001$, chi-square test) have at least one EASE-like module.

Epigenomic signatures of eILs and PEs

The epigenomic signatures of plant transcriptional enhancers and their targets are largely unknown. We then used a panel of 16 epigenetic datasets for young *Arabidopsis* seedlings (10-days to 3 weeks old) (Wang et al. 2015) to examine whether different categories of eILs associated with specific genomic features (TEs or 5′ UTRs, 3′ UTRs and gene bodies of protein-coding genes) have distinct chromatin signatures. Characteristically of TEs, chromatin signatures of TE-associated eILs (**SI Appendix, Fig. S14A**, bottom panel) differed from those of eILs associated with protein-coding genes. The categories of eILs associated with different components of protein-coding genes had very distinct chromatin signatures, and the 3′ UTR-associated eILs differed markedly from the others (**SI Appendix, Fig. S14A**).

As expected, the 3′ UTR-associated eILs lacked H3K4me3 (**SI Appendix, Fig. S14A**), which usually marks promoters. The groups of eILs associated with 3′ and 5′ UTRs had opposite patterns of di- and tri- H3K36 methylation, with H3K36me3 being noticeably less abundant in eILs associated with 3′ UTRs, consistent with the observation that H3K36me3 plays a role in productive Pol II elongation and tends to be restricted to actively transcribed sequences (Zentner and Henikoff 2013; Sein et al. 2015). Additionally, 3′ UTR-associated eILs had a noticeable enrichment of the histone H3.3 variant, which associates with the 3′ end of genes in *Arabidopsis* and animals (Stroud et al. 2012; Wollmann et al. 2012; Ooi et al. 2010).

Comparison of the epigenomic states of eILs and PEs showed that the population of PEs had the characteristic chromatin signatures of 3′ UTRs (**SI Appendix, Fig. S14B**), similar to the 3′ UTR-associated eILs (**SI Appendix, Fig. S14A**), as manifested by the patterns of di- and tri-H3K36 methylation and histone H3.3 enrichment. Notably, H3.3 is involved in gene activation (Ahmad and Henikoff 2002; Wollmann et al.

2012; Stroud et al. 2012) and enriched at enhancer regions (Jin et al. 2009; Goldberg et al. 2010). PEs also had the characteristic enrichment of H3K4me1 while lacking H3K4me3. Remarkably, the pool of interactive PEs has more-pronounced 3′ UTR chromatin signatures than the entire population of PEs (**SI Appendix, Fig. S14B,** bottom panel). Moreover, we found that PEs are DNA 5mC hypomethylated (**SI Appendix, Fig. S14B**), in contrast to TE-associated eILs (**SI Appendix, Fig. S14A**) and heterochromatin (Law et al. 2013; Wierzbicki et al. 2012).

eILs, the regions targeted by PEs, are conserved in 1135 *Arabidopsis* accessions

To determine if eILs, the regions targeted by PE regions, are conserved in the genomes of recently sequenced natural variants of *Arabidopsis thaliana*, all possible AVs in all 1135 accessions at corresponding eILs were extracted from the VCF file (1001genomes_snp-short-indel_only_ACGTN.vcf.gz). The AVs in each eIL region (2-kb) per genome were analyzed independently, using the same methods as analyzing PEs. The genomic average of AV is 200-nt per 2-kb. We found that more than 80% (570 eILs; p = 4.675e-05, chi-square test) of total eILs have AVs less than 200 nt per 2-kb, with the average being 165 nt per 2-kb eIL region, suggesting that eILs are more conserved than the genomic average.

Analysis of the genome-wide occupancy of H3K4 monomethylation confirmed that H3K4me1 occupies less than a third of all protein-coding genes

To analyze the genome-wide occupancy of H3K4me1, the previously published data from chromatin immunoprecipitation followed by high-density/resolution whole-genome tiling microarray (ChIP-Chip) of H3K4 monomethylation (GSE13613 (Zhang et al. 2009); sample GSM343141 in WT, performed for 3-week-old *Arabidopsis* seedlings) were downloaded from GEO. The raw microarray data (CEL files) were analyzed using TileMap implemented in CisGenome to identify ChIP-chip peaks (Ji et al. 2006; 2008) (see the detailed methods in the **Microarray data analysis** section in **Methods**). This identified a total of 13,747 H3K4me1 ChIP-chip peaks (**SI Appendix, Fig. S1A**). Similarly to the previously published study by Zhang *et al.* (Zhang et al. 2009), we found that the majority of the H3K4me1 ChIP-chip peaks are located at protein-coding gene regions (n= 13,367; **SI Appendix, Fig. S1A**). The *Arabidopsis* genome version 10 (TAIR10) has a total of 27,416

annotated protein-coding genes. After carefully examining the H3K4me1 ChIP-chip peaks that are located at protein-coding gene regions, we found that less than one third of all protein-coding genes contain a H3K4me1 enriched ChIP-chip peak (*n*= 7,657, 28%, posterior probabilities ≥0.5; **SI Appendix, Fig. S1A**). Thus, our analysis confirmed the data reported by Zhang *et al.* that H3K4me1 occupies only less than a third of the transcribed regions of all protein-coding genes.

Analysis of the chromatin signatures at the 3′ ends of protein-coding genes that do and do not house PEs

Almost all PEs overlap with protein-coding gene structures, and specifically, a total of 961 PEs overlap with the 3′ UTRs of 1,098 protein-coding genes. To profile the chromatin signatures of this group of protein-coding genes that house PEs at their 3′ UTRs, we first analyzed the H3K4 methylation at the gene body of these 1,098 protein-coding genes. Each protein-coding gene may have more than one isoform; to avoid redundant analysis, only one isoform per protein-coding gene was selected for further analysis (the isoform with the longest overlap at the 3′ UTR with PEs and an overlap of at least 100 nt). The previously published ChIP-Chip datasets of H3K4 methylation (GSE13613 (Zhang et al. 2009); sample GSM343141 for H3K4 mono-methylation, GSM343143 for H3K4 di-methylation, and GSM343144 for H3K4 tri-methylation in WT, performed for 3-week-old *Arabidopsis* seedlings) were downloaded from GEO and analyzed. The raw microarray data (CEL files) were analyzed using TileMap implemented in CisGenome to identify ChIP-chip peaks (Ji et al. 2006; 2008). Detailed methods are fully described in **Microarray data analysis** in **Methods**. To perform the metagene scaled gene-body analysis, the gene bodies were defined to be from the start to the stop codons of the protein-coding regions. Each gene body was then divided into 100 intervals/bins, and the accumulation of HMM posterior probability (as a measurement of signal intensity) for each ChIP-chip dataset (described above) was tabulated per interval/bin. The summed signal intensities in each 100 intervals in the entire gene body was graphed in a metagene plot and scaled to each gene body (50% represents the center of the gene body, as indicated by the orange dotted line). We found that in the group of protein-coding genes that house PEs at their 3′ UTRs, H3K4me1 is mostly present along gene bodies and dramatically declines towards the 3′ ends of genes (**SI Appendix, Fig. S1B**). We then compared the H3K4me1 distribution found at PEs, and the result indicated that the H3K4me1

distribution at PEs does not mimic the H3K4me1 distribution in the gene body of this group of protein-coding genes that house PEs (compared to data shown in **Fig. 1D**).

Next, we analyzed the chromatin signatures (H3K4 methylation (mono-, di-, and tri-) and acetylation and trimethylation of H3K27) of the 1.5 kb region flanking the TESs of these protein-coding genes (*n*= 1,089) that house PEs at their 3′ UTR (**SI Appendix, Fig. S10A**), and the distribution of H3K4 methylation (mono-, di-, and tri-) and acetylation and trimethylation of H3K27 were analyzed. The previously published ChIP-Chip datasets of H3K4 methylation (GSE13613 (Zhang et al. 2009); sample GSM343141 for H3K4 mono-methylation, GSM343143 for H3K4 di-methylation, and GSM343144 for H3K4 tri-methylation in WT, performed for 3-week-old *Arabidopsis* seedlings), H3K27 acetylation (GSE15597 (Charron et al. 2009), performed for 5-day-old *Arabidopsis* seedlings), and H3K27 tri-methylation (GSE7093 (Zhang et al. 2007), performed for 10–14-day-old *Arabidopsis* seedlings) were downloaded from GEO and analyzed (methods as described above). To analyze the distribution of these chromatin modifications, the metagene analysis centered at TESs was used. The entire 3-kb region (± 1.5 kb from the TESs) was divided into 100 intervals (30-nt per bin), and the accumulation of HMM posterior probability (as a measurement of signal intensity) for each ChIP-chip dataset (described above) was tabulated per 30-nt bin. The summed signal intensities in each 100 intervals in the entire 3-kb region was graphed in metagene plot relative TESs (orange dotted line). We found that the groups of protein-coding genes that house PEs at their 3′ UTRs (**SI Appendix, Fig. S10A**) resembled the chromatin signatures of H3K4me1 characterized at PEs regions (as shown in **Fig. 1D**).

As a comparison to the groups of protein-coding genes that house PEs at their 3′ UTRs, the distributions of chromatin modifications at the 3′ ends of all other protein-coding genes that do not house PEs were also analyzed (*n*=25,350), using the same methods described above. We found that the profiles of H3K4 methylation are different between the protein-coding genes that do not harbor PEs compared to the protein-coding genes that harbor PEs at their 3′UTRs (left panels of **SI Appendix, Fig. S10A and B**); specifically, H3K4me3 is more prevalent in the 3′ ends of the protein-coding genes that do not house PEs (**SI Appendix, Fig. S10B**, left panel). Moreover, the profiles of H3K27ac and H3K27me3 are also different between these two groups of protein-coding genes (left panels of **SI Appendix, Fig. S10A and B**).

We also analyzed and compared the positional and sequence conservation in these two groups of protein-coding genes (the protein-coding genes that do not harbor PEs versus the protein-coding genes that harbor PEs at their 3′UTRs). The collection of sequence variants analyzed by the 1001 *Arabidopsis thaliana* genomes project were utilized (1001 Genomes Consortium 2016), using the method described in **DNA sequence conservation in PE regions** section in **SI Methods** below. We found that the 3′ ends of protein-coding genes that do not house PEs have an allelic variant (AV) average of 9.5% (285 nt per 3-kb region), while the 3′ ends of protein-coding genes that house PEs have an AV average of 8.8% (3′UTRs of these protein-coding genes were analyzed, approximately 264 nt per 3-kb region).

Analysis of the enhancer trap insertion sites that are in close vicinity to PEs

One of the powerful experimental approaches in identifying transcriptional enhancers is the use of enhancer trap lines. Enhancer traps involve using a *GUS* reporter gene with a minimal promoter in a construct that is typically randomly integrated into the plant genome. This minimal promoter and *GUS* gene can only be activated by an enhancer element that is located nearby or overlaps with the integration site. Therefore, we conducted a search for the enhancer trap lines isolated at the developmental stage closest to the stage used in PE identification to examine the relationship between PEs and active enhancers identified in the previously isolated enhancer trap lines.

Numerous enhancer trap lines have been isolated in *Arabidopsis* (Weber et al. 2016). Thomas Jack's population of *Arabidopsis* enhancer-trap lines (CS31086, Arabidopsis Biological Resource Center (ABRC) at Ohio State University (Campisi et al. 1999)) was one of the collections screened by Liu *et al.* (Liu et al. 2005a; 2005b) to identify the enhancer trap lines that are specific to seed germination and young seedlings, similar to the conditions and developmental stages used in PE identification. Multiple enhancer-trap lines from this population exhibited strong GUS signals at the micropylar end of the seed, and these lines were designated Blue Micropylar End (BME) lines (ABRC stock no. CS24362–CS24480 (Liu et al. 2005a)). From 121 independent lines that were previously examined at the seedling germination stage, the position of the insertions was identified for 10 lines using genome-walking PCR experiments by Liu *et al.* (Liu et al. 2005a). Therefore, we examined these 10 positionally mapped enhancer trap lines that were previously isolated by Liu *et al.* and analyzed the genomic regions of these reporter gene insertion sites. We found that among these 10 lines, 2 were

overlapping with the positions of PEs (PE_000643 and PE_001487, **SI Appendix, Fig. S11 A and B**), while one insertion was located within 6 kb of a PE (PE_000778, **SI Appendix, Fig. S11 C**). This result suggests that the PEs overlapping with or located in close proximity to the enhancer trap reporter gene insertion could be active putative enhancers in *Arabidopsis*. Also, the enhancer trap lines provide one of the most direct lines of evidence of measuring an enhancer's activity in its native genomic context, which is a crucial point given the intra-genic nature of PEs.

Analysis of convergent gene pairs and PEs

To examine if PE identification shows any preference towards convergent gene pairs, we examined all convergent gene pairs in the *Arabidopsis* genome. After filtering out pseudogenes, non-protein-coding genes, and transposon-related genes, we found that the *Arabidopsis* genome contains 3,551 protein-coding convergent gene pairs that are separated by less than 1 kb. These convergent gene pairs cover approximately 8% of the genome. However, only a small proportion of these convergent gene pairs overlap with PE regions (451 out of 3,551; 13%), suggesting that there is no bias in PE identification towards convergent gene pairs.

SUPPLEMENTAL FIGURES

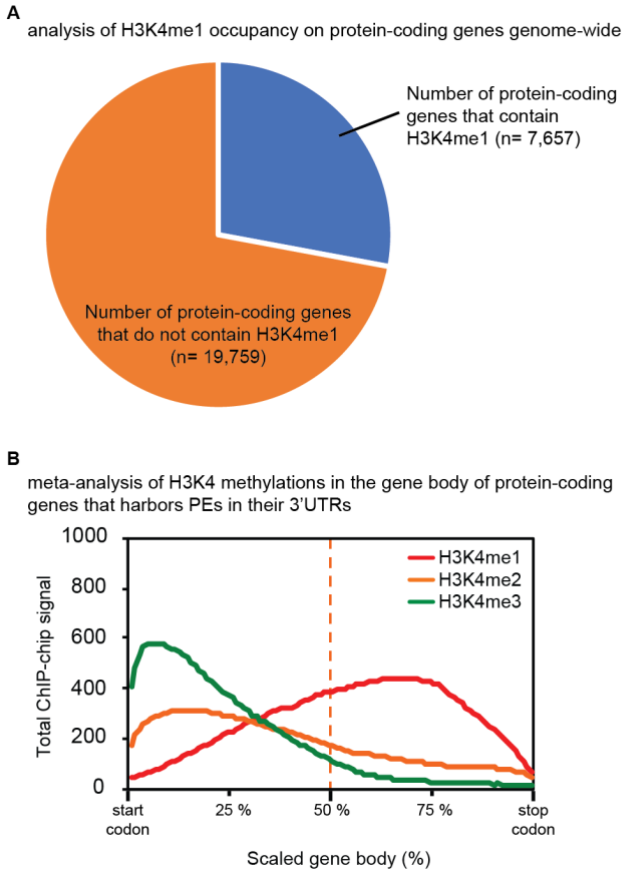

Figure S1. Histone modification H3K4me1 occupies less than a third of the transcribed regions of all protein-coding genes in the *Arabidopsis* genome and the H3K4me1 signal declines towards the 3' ends of protein-coding genes that harbors PEs.
(A) There is a total of 27,416 protein-coding genes annotated in the *Arabidopsis* genome version 10 (TAIR10), and less than a third of all protein-coding genes contain H3K4me1 ChIP-chip peaks (*n*= 7,657, 28%, posterior probabilities ≥0.5). (B) Meta-analysis of histone modification of H3K4 methylation (mono-, di-, and tri-) at gene bodies of protein-coding genes that house PEs at their 3′ UTRs. The protein-coding regions between start and stop codons were analyzed and shown. The y-axis shows the signal intensity from ChIP-chip data; the x-axis depicts the scaled gene body (50% represent the center of the gene bodies, as indicated by orange dotted line).

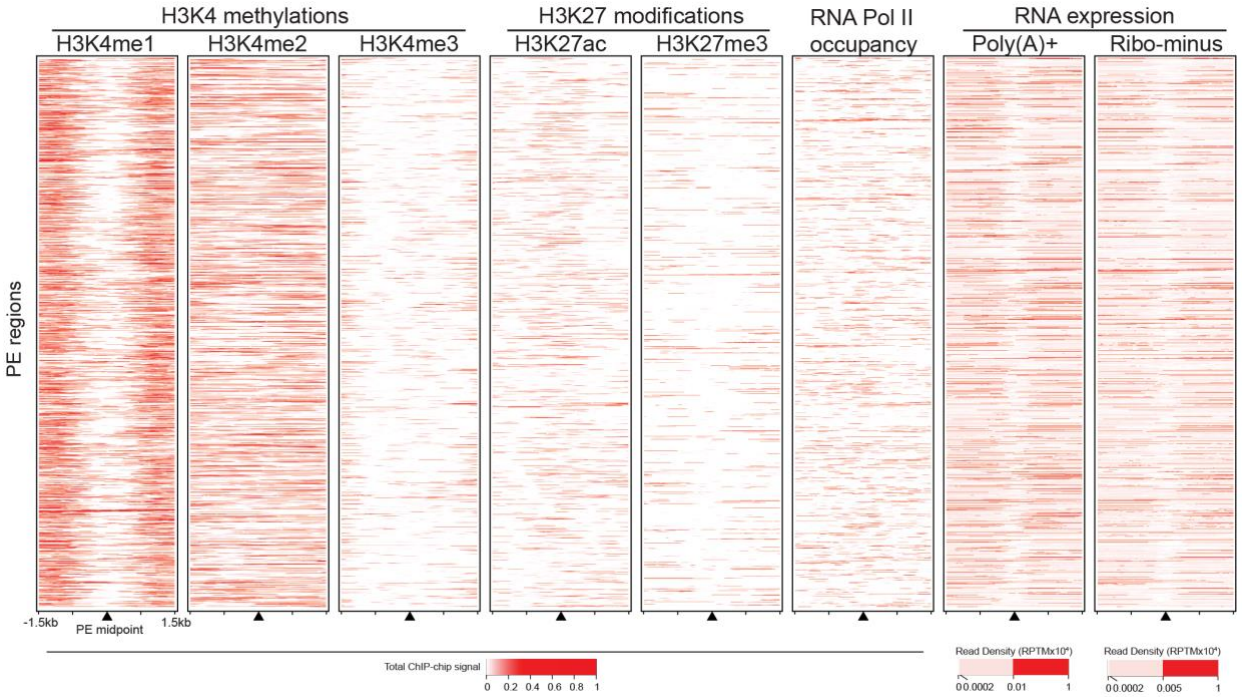

Figure S2. The identified PEs exhibit canonical epigenomic signatures of mammalian enhancers. Related to **Fig. 1 D to G**. The chromatin signatures, Pol II occupancy, and transcriptional activities were analyzed for all 1919 identified PEs and presented as heat maps. The enrichment and distribution data are shown for H3K4 methylation (mono-, di-, and tri-), acetylation and trimethylation of H3K27, RNA Pol II occupancy, and the transcriptional activity as both poly(A)+ and ribo-minus RNA expression (depicted from left to right, respectively). Each heatmap enrichment analysis was centered on the PE midpoint (indicated by black triangles) and 1.5 kb flanking from the midpoint (3-kb window); the PE regions were ordered by their genomic locations. Except for transcriptional activity (as in Read Density), chromatin signatures and Pol II occupancy depicted using the same scale and represented as total ChIP-chip signal. For the meta-analysis of the same epigenomic signatures and their distribution in PEs, please refer to **Fig.1 D to G**.

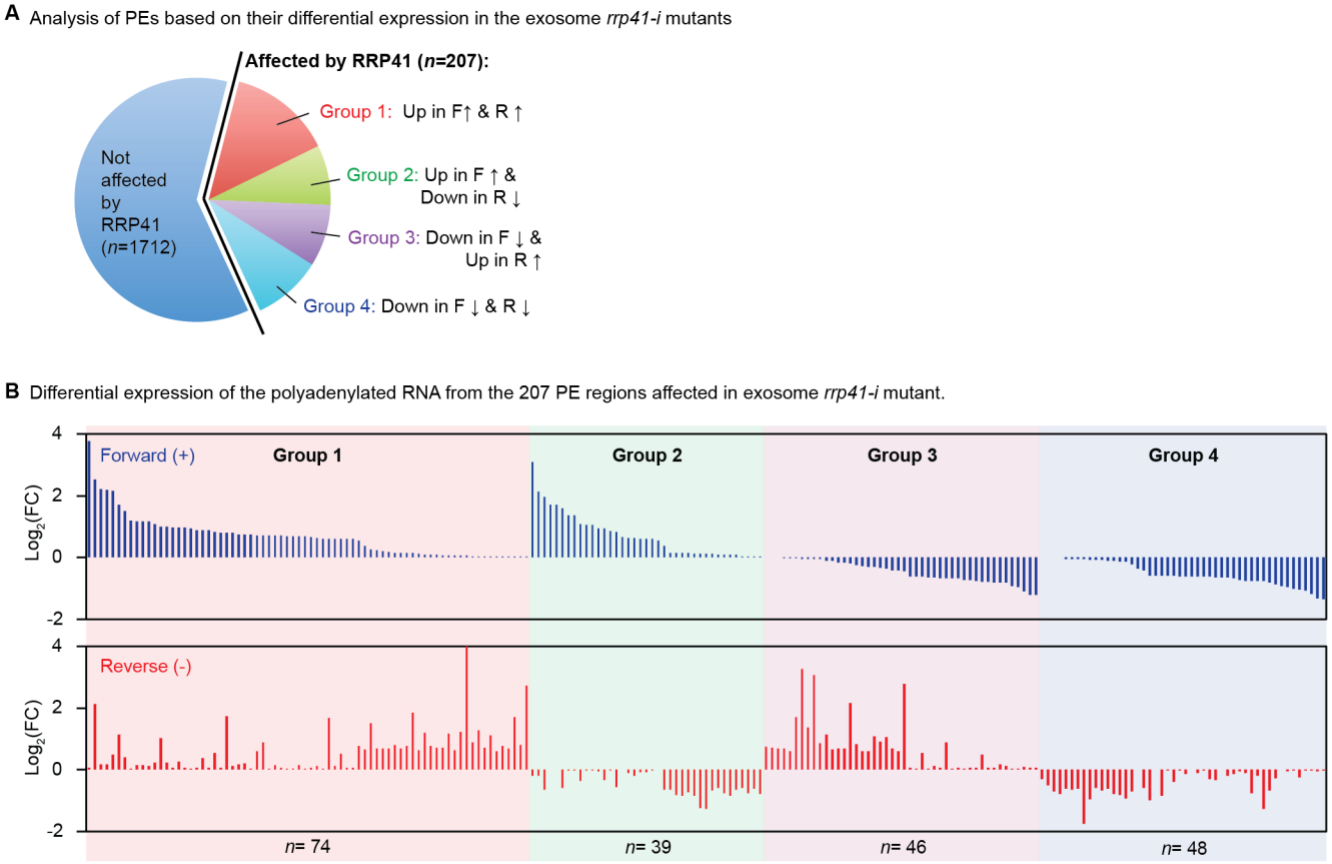

Figure S3. Differential expression of polyadenylated RNA from the 207 PE regions affected by depletion of the exosome subunit RRP41.

(A) Analysis of PEs based on differences in transcriptional activity in the *rrp41-i* lines. F, forward; R, reverse direction. Upwards arrow (↑) indicates up-regulation; downwards arrow (↓) indicates down-regulation. The 207 PE regions that are affected by RRP41 depletion can be divided into 4 groups: group 1: expression is down-regulated in both forward (F) and reverse (R) directions (*n*=48); group 2: expression down in F and up in R (*n*=46); group 3: expression is up in F and down in R (*n*=39); group 4: expression is up in both F and R (*n*=74). (B) Differential expression of all affected PEs is shown. The detailed $\log_2$ (fold change) of all 207 PEs that are either up- or down- regulated upon exosome RRP41 depletion are shown strand specifically (based on expression of poly(A)$^+$ RNA). Up-regulated PEs have $\log_2$ (FC) > 0 and down-regulated PEs have $\log_2$ (FC) < 0. The mild changes in *rrp41* mutants observed in our studies were consistent with the observed effect of the exosome in regulating eRNAs in mammals (Pefanis et al. 2015), which also found that the exosome is most likely only one of many different factors regulating enhancer functions.

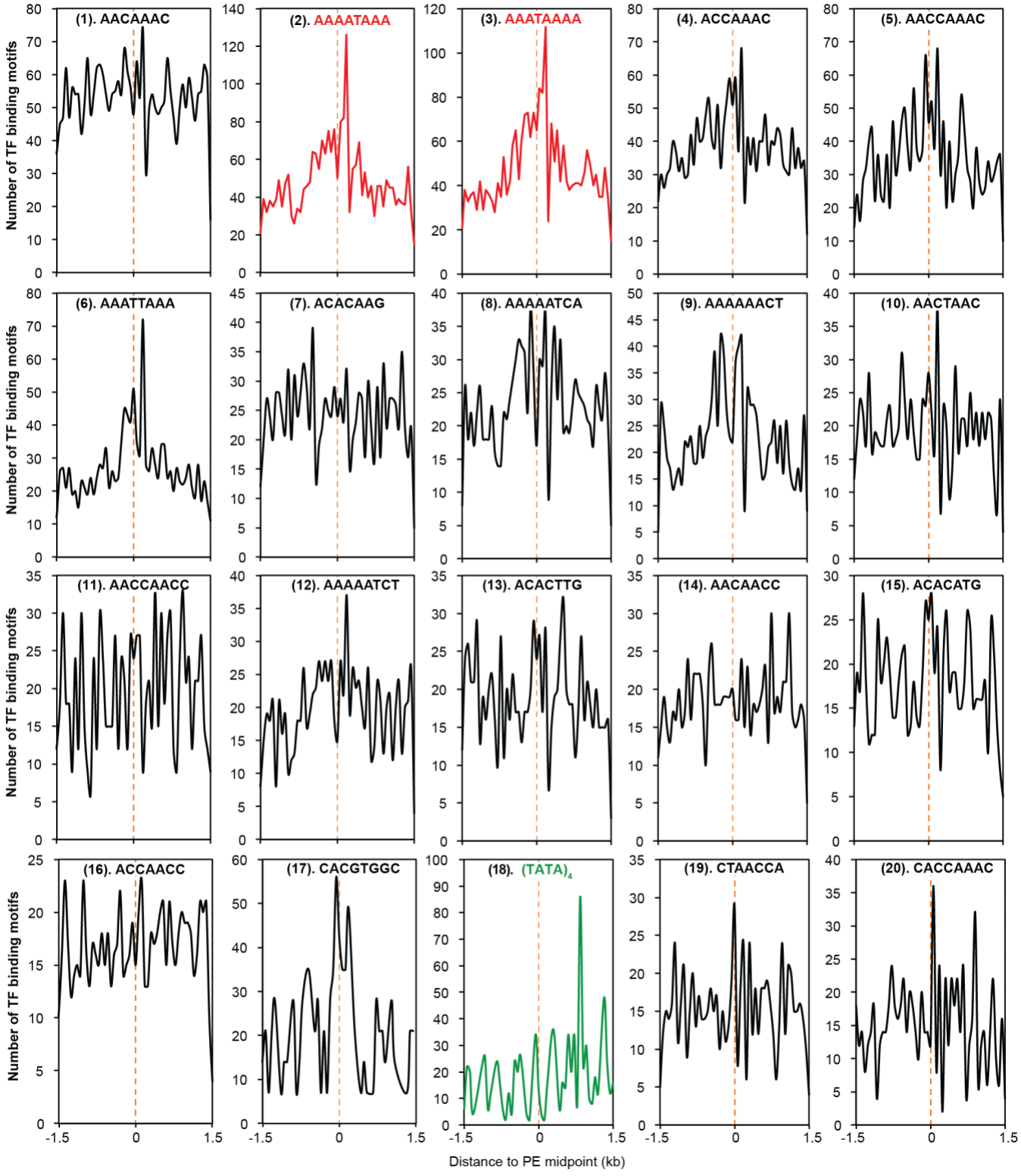

Figure S4. Distribution of specific TF motifs in the PE regions.
Only the top 20 most frequent TF motifs are shown (for the same pool of TFs as shown in **Fig. 2A–D**).

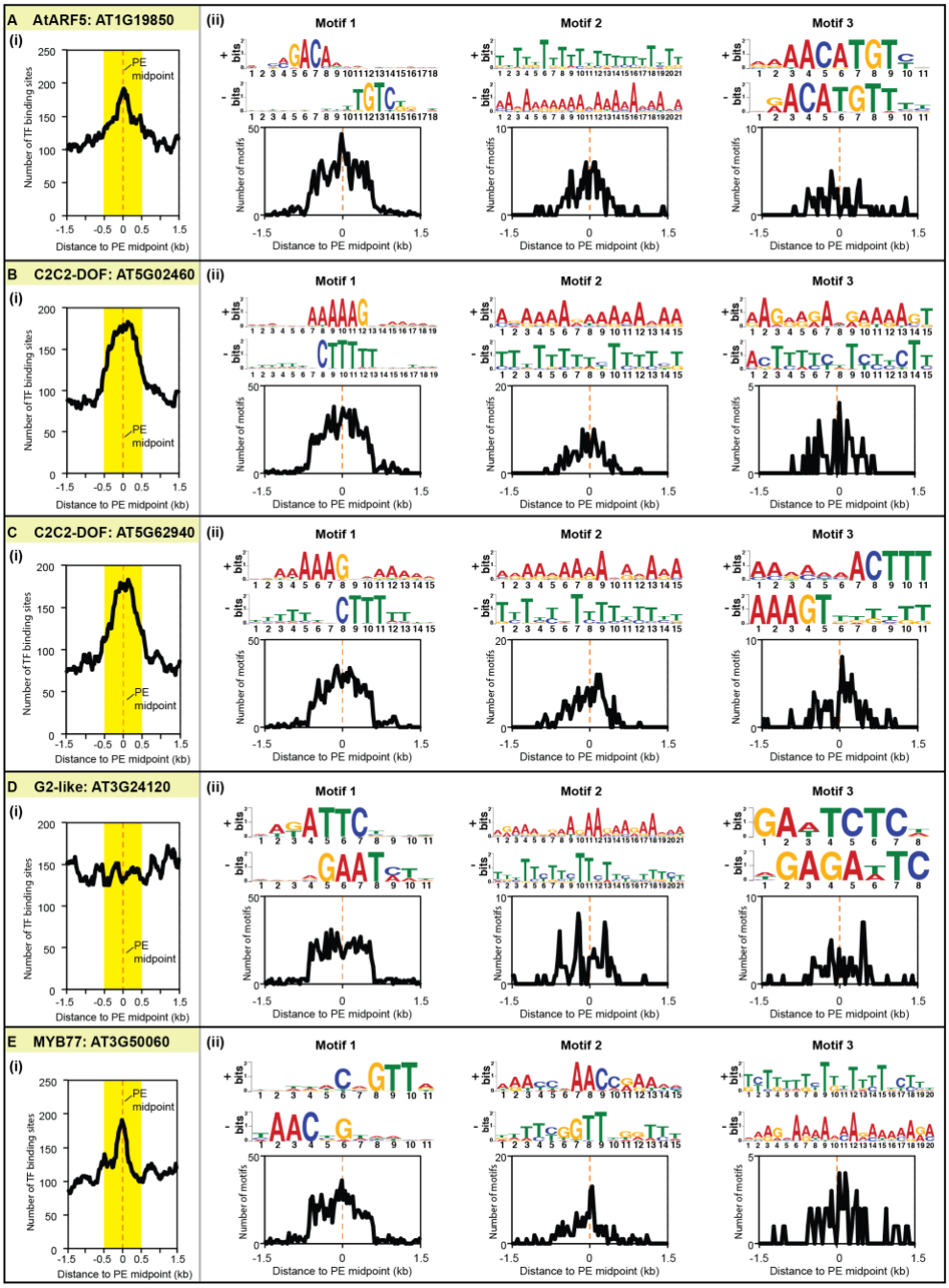

Figure S5. Binding motifs of the five TFs that bind to a large group of PE midpoint-proximal regions
Related to **Figure 2** (also see **Dataset S7**). Rows (A–E), analysis of the binding sites and motifs for: AtARF5 (AT1G19850), C2C2-DOF (AT4G02460), C2C2-DOF (AT5G62940), G2-like (AT3G24120, and MYB77 (AT3G50060), respectively. Each row contains two panels. Panel (i), the distribution of the sites bound by each TF in the PE regions. The midpoint-proximal region of PE (in yellow). Panel (ii), the three most-enriched binding motifs in PE midpoint-proximal regions were determined by interrogating all binding sites specifically at PE midpoint-proximal regions for each TF (E-value ≤ 0.05). The motif in both forward (+) and reverse-complementary (–) orientations and their distribution are shown. These midpoint-specific motifs differ from motifs enriched in the entire 3-kb PE region (**Fig. S6A–E**, panels (ii)). The midpoint-specific motifs are A/T rich and distinct from the general binding motifs calculated in a genomic context-independent manner (shown in **Fig. S6A–E**, panels (i)).

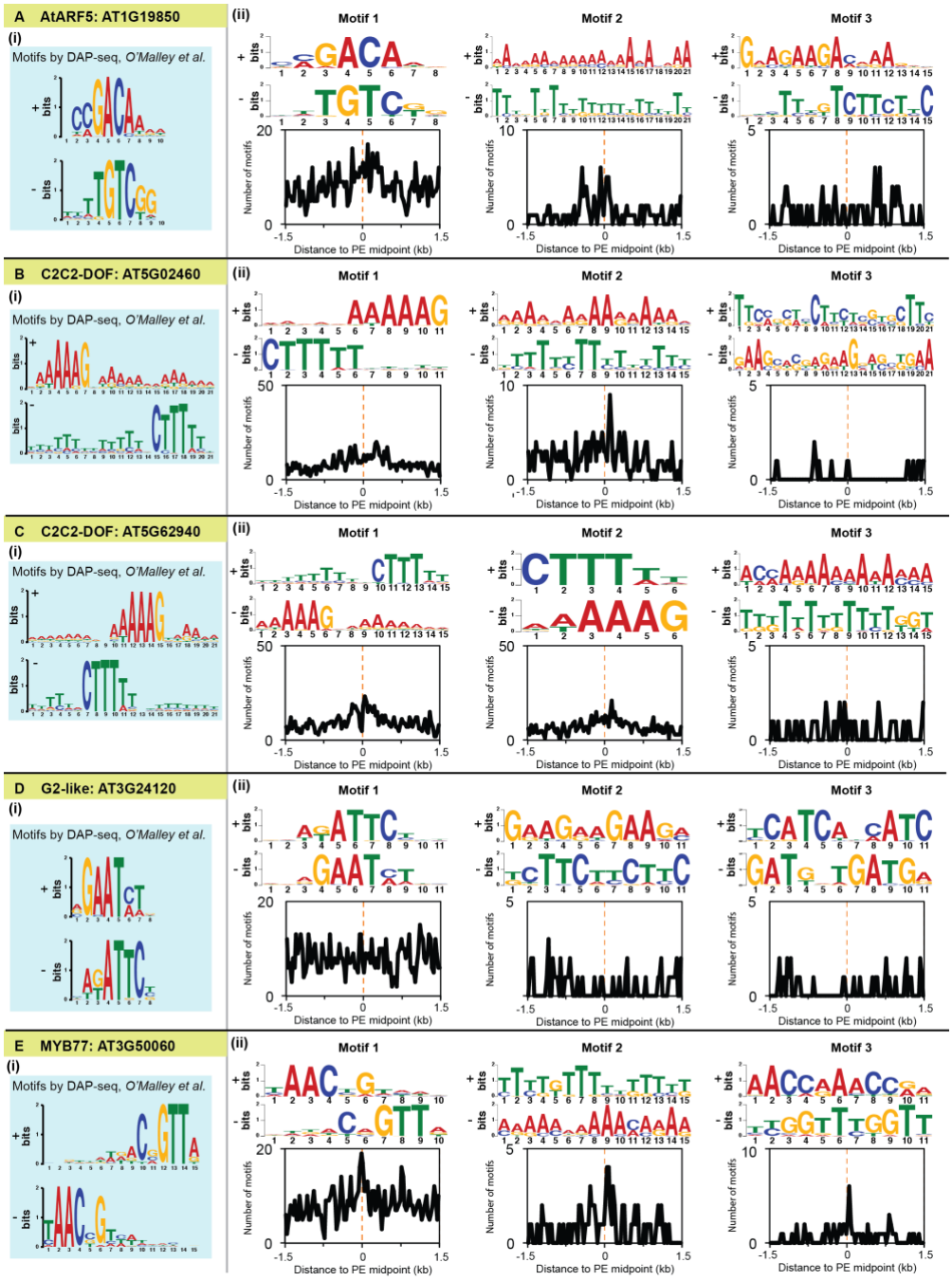

Figure S6. **Binding motifs determined from all binding sites within the entire 3-kb PE regions (±1.5 kb from the PE midpoint) for the five TFs.**

Related to **Figure S5** (also see **Dataset S7**). Rows (A–E) show results of the analysis of the binding sites of each TF: AtARF5 (AT1G19850), C2C2-DOF (AT4G02460), C2C2-DOF (AT5G62940), G2-like (AT3G24120), and MYB77 (AT3G50060), respectively. Each row contains two panels, (i) and (ii): Panel (i) represents the general binding motifs for each TF, which were determined by interrogating all peaks of binding genome-wide; thus, the general binding motifs are not specific to genomic context (O'Malley et al. 2016). Panel (ii) depicts the top three most probable TF binding motifs in the entire 3-kb PE regions (E-value ≤ 0.05). In contrast to the motifs enriched from PE midpoint-proximal regions (shown in **Fig. S5**), all sites for each TF binding in the entire 3-kb PE region were used to find the motif enrichment. The sequence of each motif and its distribution relative to the position of PE midpoint is shown. The identified motif sequences closely resembled the sequences of the general consensus motifs of each TF binding site.

Figure S7. **PEs are distributed across euchromatic regions on all *Arabidopsis* chromosomes.**
Genomic localization of the identified PEs on five *Arabidopsis* chromosomes. The location of each PE on each chromosome is marked using a horizontal line.

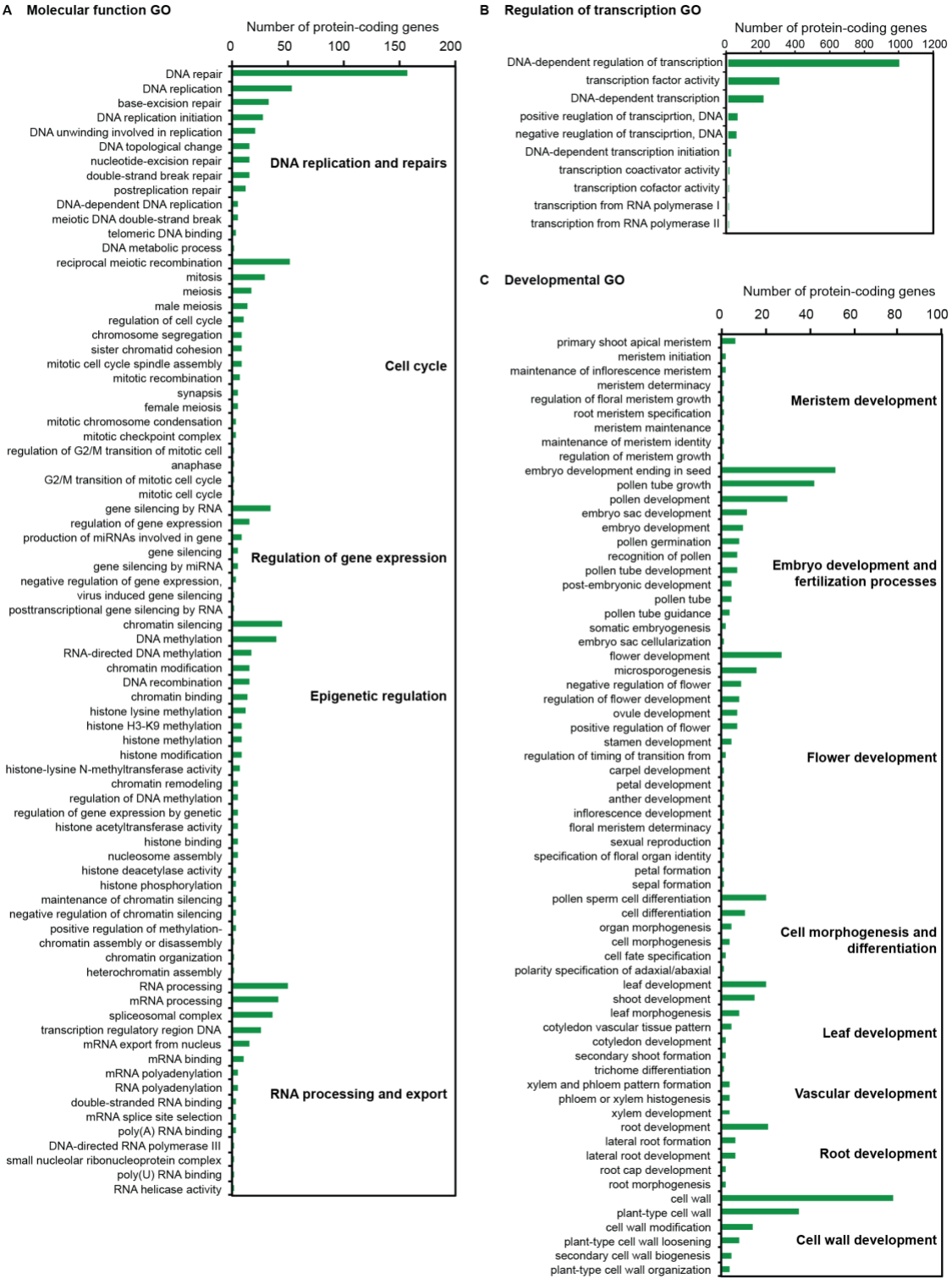

Figure S8. GO analysis of protein-coding genes that overlap with intra-genic PEs (extended from Fig. 3C). The distribution of GO terms in each category, based on molecular functions (A), regulation of transcription (B), and developmental processes (C). Fisher's exact test was utilized for assessing the GO term significance. The *p*-value and odds ratio are included in **Dataset S10**.

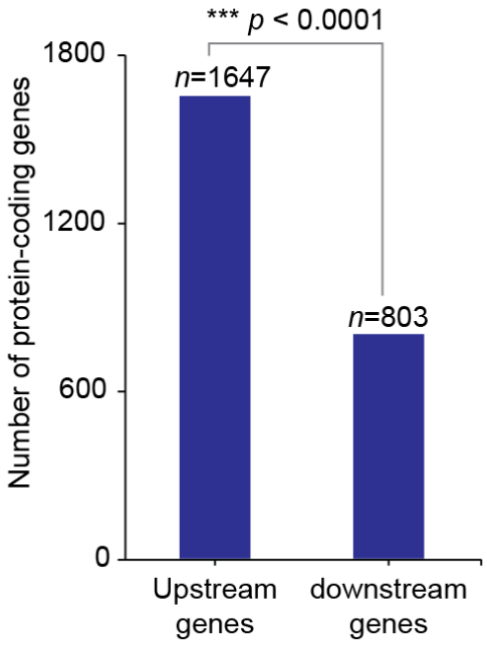

Figure S9. Distribution of TSSs relative to the PE midpoints.
This Figure is supporting **Fig. 3D**, which shows the analysis of distances between PEs midpoints (*n*=1348) and TSSs of all 2450 protein-coding genes overlapping with PEs, mapped on both DNA stranded in a strand-specific manner. Regions beyond the PE borders extended to ±10 kb from the PE midpoints were analyzed. As shown on **Fig. 3D**, the protein-coding genes were sub-divided into two groups based on where their TSSs are located relative to the PE midpoints: either upstream (i) or downstream (ii); ****p* < 0.0001. The bars represent the total number of TSSs in 1-kb intervals from PE midpoints.

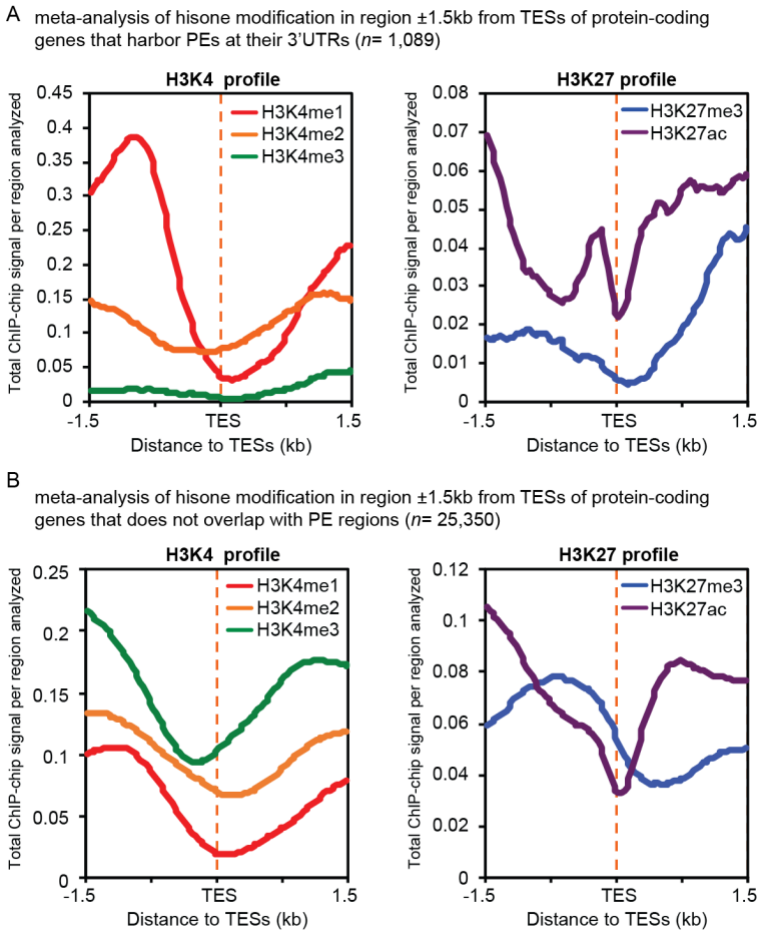

Figure S10. Chromatin signatures differ between protein-coding genes that do and do not harbor PEs in their 3′ UTRs .

Meta-analysis of histone modifications in regions flanking the TESs of protein-coding genes. Two distinct groups of protein-coding genes were analyzed: (A) the groups of protein-coding genes that harbor PEs in their 3′ UTRs; (B) the groups of protein-coding genes that do not harbor PEs. The orange dotted lines mark the TESs of protein-coding genes. The distribution of H3K4 methylation (mono-, di-, and tri-) and acetylation and trimethylation of H3K27 are shown on the left and right, respectively, in each panel. The y-axis shows the signal intensity from ChIP-chip data (normalized by number of regions analyzed); the x-axis depicts the distance ±1.5 kb from TESs of protein-coding genes.

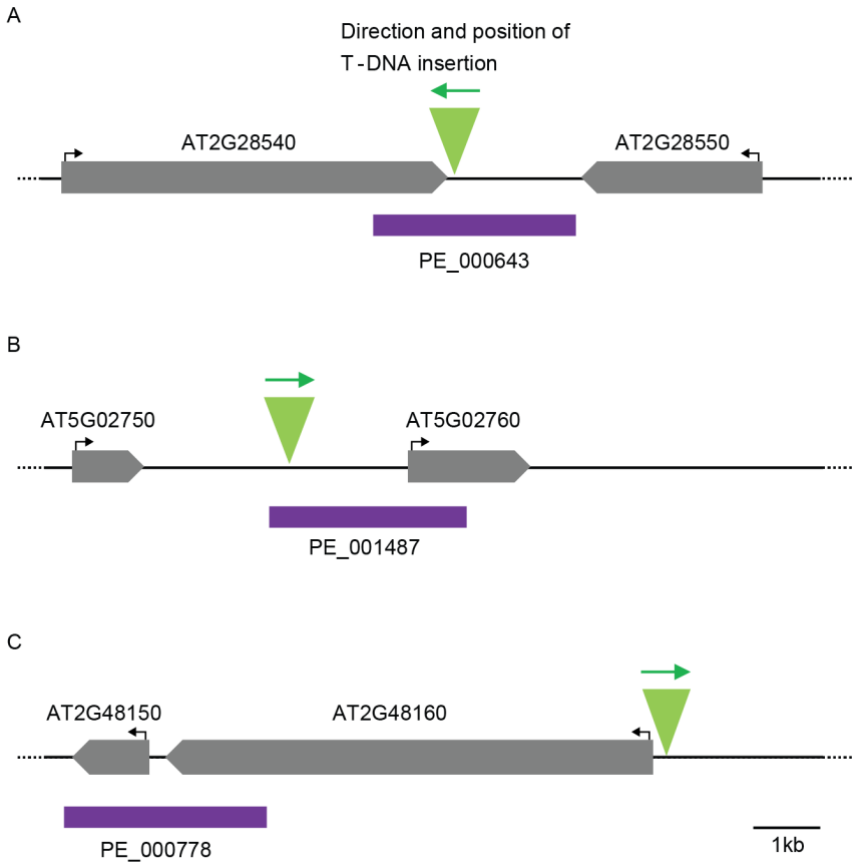

Figure S11. Some PEs are located in close proximity to known enhancer trap reporter gene insertions. The genomic location of PEs and sites of enhancer trap reporter gene insertions are shown. (A–C) Three examples of PEs (PE_000643, PE_001487, and PE_000778) and independent enhancer trap lines are shown. Among the 10 independent enhancer trap lines that were experimentally characterized by Liu et al. 2005a at the seedling germination stage, two reporter gene insertions are located in the same genomic locations as the sites of PEs (A and B, PE_000643 and PE_001487, respectively), and one insertion is located in close proximity (< 6 kb) to PE_000778 (C). The genomic DNA region (10-kb window) of the enhancer trap insertion is shown; the enhancer trap insertion sites are indicated as green inverted triangles (the green arrow refers to the direction of the T-DNA insertion). The genomic regions of PEs (purple bars) and protein-coding genes (grey bars), and the direction of the protein-coding genes (black arrows) are shown.

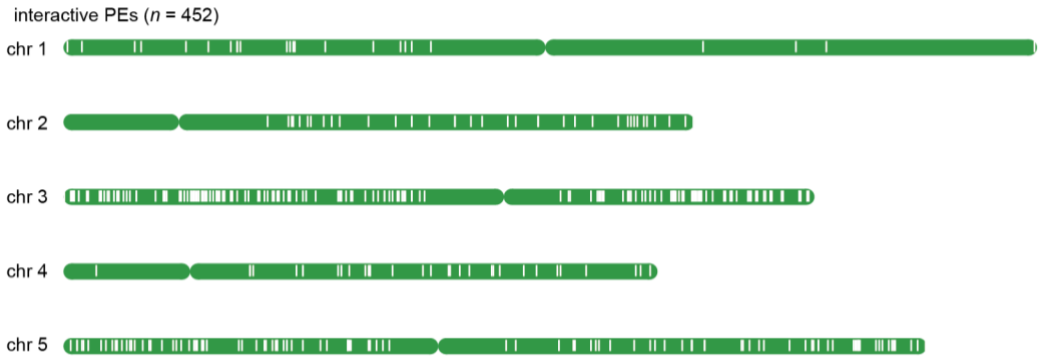

Figure S12. The genomic locations of the interactive PEs.
The genomic location of the interactive PEs (*n*= 452) is shown on the five *Arabidopsis* chromosomes. The location of each interactive PE on each chromosome is marked using a white vertical line.

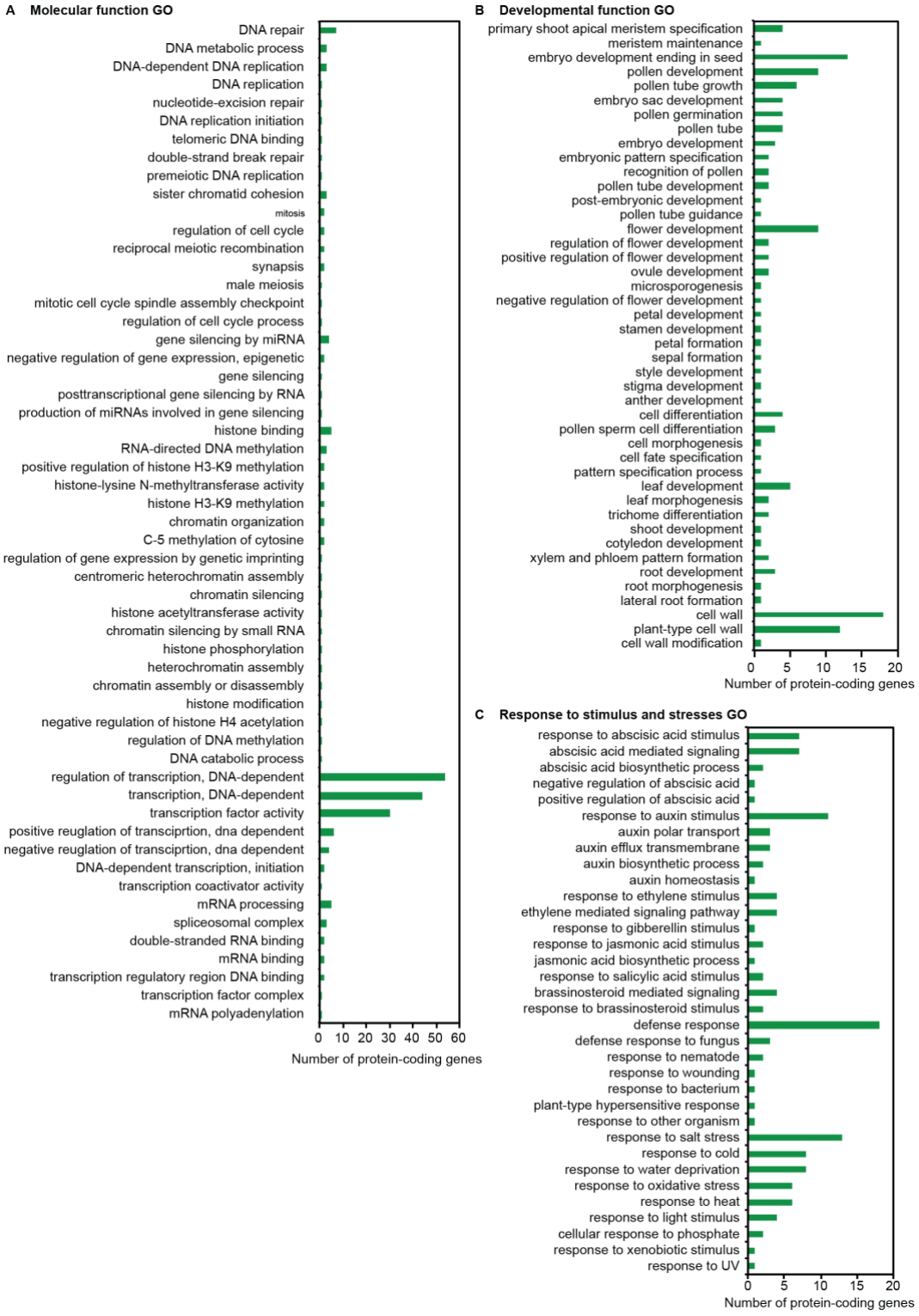

Figure S13. GO analysis of protein-coding genes that interact with PEs.
The distribution of GO terms in each category is presented based on molecular functions (A), developmental processes (B), and responses to stimulus and stresses (C). Fisher's exact test was utilized for assessing the GO term significance. The *p*-value and odd ratio are included in **Dataset S13**.

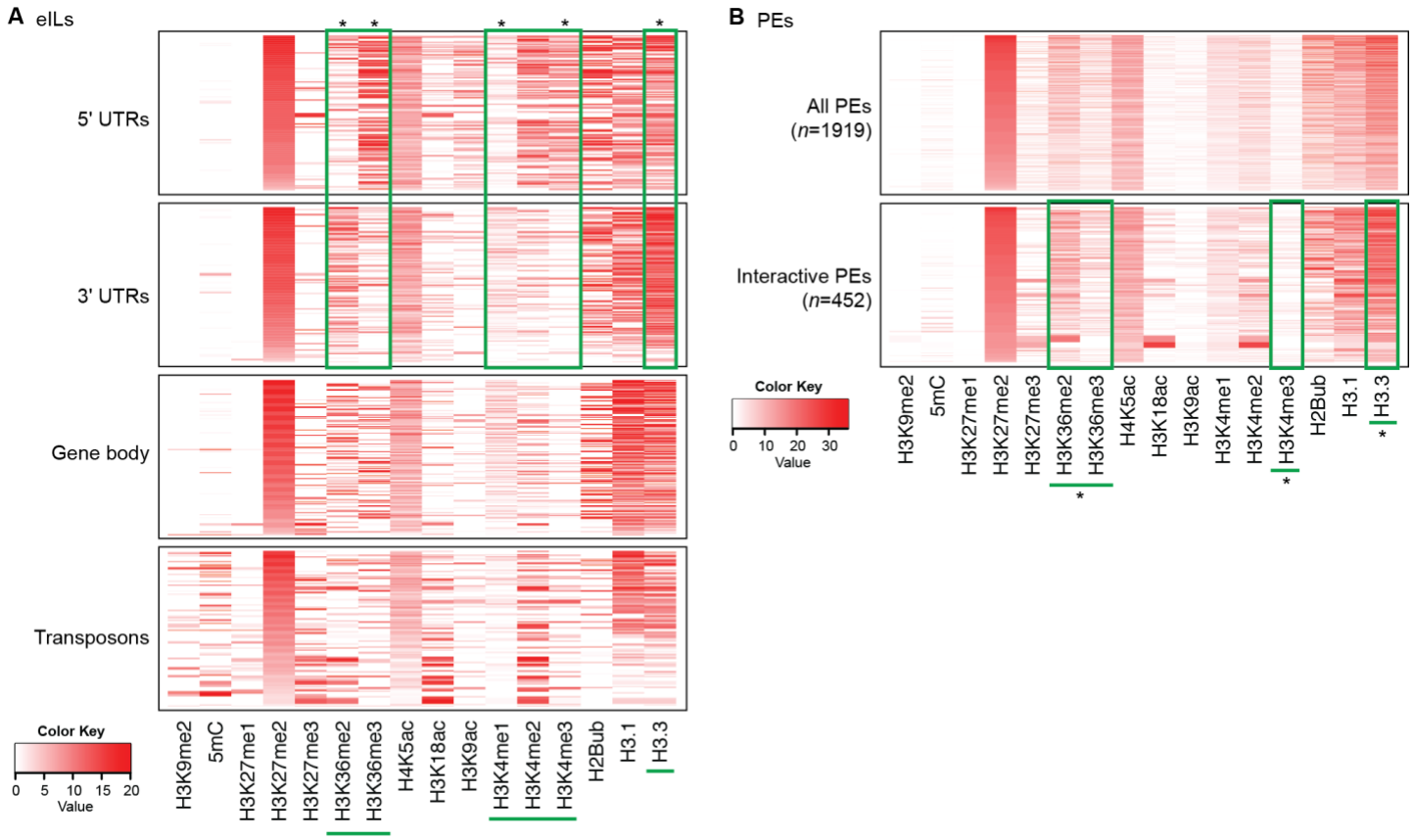

Figure S14. Epigenomic signatures associated with eILs and PEs.
Heatmaps representing chromatin signatures of eILs (A) and PEs (B). Sixteen genome-wide epigenome datasets were used, including 13 on histone modifications, two on histone variants, and one on 5mC DNA methylation. The color scale represents the enrichment (red) or depletion (from white to pale pink) of specific chromatin modifications. All identified PEs and eILs were analyzed independently and each row represents an epigenetic profile of one single region. Brackets and asterisks highlight the different epigenetic signatures between different groups of eILs and PEs. (A) Chromatin signatures of different categories of eILs associated with particular genomic features, such as 5′UTRs, 3′UTRs, the gene body of protein-coding genes, and TEs. (B) Chromatin signatures of all 1919 identified PEs (top) and, specifically, interactive PEs forming chromatin loops with eILs (bottom).

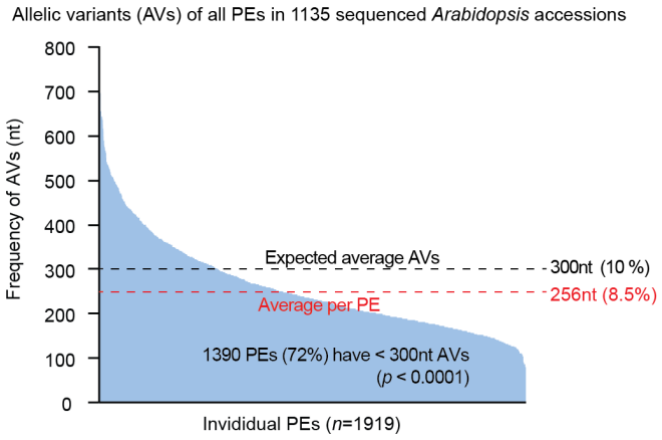

Figure S15. PEs are highly conserved in 1135 natural *Arabidopsis* variants.
The DNA sequence conservation of identified PEs measured as the frequency of AVs per PE region in 1135 sequenced *Arabidopsis* accessions. The expected average AV across the entire genome in a genomic context-independent manner is 300 nt per 3-kb region for all the accessions (black dotted line). The average AV per total PE region is 256 nt ($p < 0.0001$, red dotted line), and more than 72% of the total PEs have higher conservation, with AV less than 300 nt.

CAPTIONS FOR DATASET S1 TO S14

Dataset S1. The list of identified PEs (*n*=1919)

Dataset S2. Summary of RNA-seq data

Dataset S3. The list of PEs regions affected by the exosome RRP41 depletion (*n*=207)

Dataset S4. Summary of TFs with binding peaks in PE regions

Dataset S5. Summary of the groups of TF families with binding peaks in PE regions

Dataset S6. The total number of PE midpoint-proximal regions bound by individual TFs.

Dataset S7. Summary of the three most enriched binding motifs of the five TFs that bind to a large number of PE midpoint-proximal regions

Dataset S8. Summary of the genomic features overlapping PEs

Dataset S9. Genomic features overlapping PEs

Dataset S10. GO analysis of protein-coding genes that overlap with intra-genic PEs

Dataset S11. The 1076 PE-eIL intrachromosomal interactions

Dataset S12. Summary of the analysis of PEs forming multiple chromatin loops

Dataset S13. GO analysis of protein-coding genes that interact with PEs

Dataset S14. A summary of the high-diversity panel, which includes 10 selected accessions and the reference accession Col-0, and the average allelic variants (AVs) at all PEs for each accession

SI METHODS

Mapping and assembly of transcriptomes

The Tuxedo protocol was used to map and align the RNA-seq reads in each dataset (a total of four data series, including both poly(A)$^+$ and ribo-minus) and to assemble the transcriptome (Trapnell et al. 2012). The same protocol was applied on all datasets in the four data series, including poly(A)$^+$ and ribo-minus. RNA-seq reads of each biological replicate were mapped independently to the *Arabidopsis* TAIR10 genome assembly using the splice junction mapper TopHat (v2.0.9) (Kim et al. 2013) with default parameters for stranded RNA-seq (--coverage-search --microexon-search --library-type=fr-firststrand) (Dataset S1). The mapped reads of each biological replicate were independently assembled into transcripts using Cufflinks (v2.1.1) (Trapnell et al. 2010).

The assembled transcripts were further processed using associated functions in the Cufflinks package (as described below). All transcript assemblies from the same data series were processed through the Cuffmerge program. The Cuffmerge function was used to merge novel and known isoforms, and eliminate transcribed fragments (transfrags), which might be artifacts. The merged assemblies were quantified in each condition by Cuffdiff and tested for the statistical significance of each observed change in expression between induced *rrp41-i* and uninduced control RRP41. The differentially expressed transcripts were considered significant if the uncorrected *p*-value was greater than the false discovery rate after Benjamini-Hochberg correction for multiple-testing.

Meta-analysis of transcriptional activity at PE regions

To profile the strand-specific RNA expression at PE regions, the .bam file (accepted_hits.bam) containing all qualified read alignments generated by the read aligner, TopHat, for each dataset was used. To generate the strand-specific read alignment files, the .bam files associated with each dataset were analyzed separately, and the reads from the positive strand (forward) and those from the negative strand (reverse) were separated into two files using the genomecov function in BEDTools version 2.24 (Quinlan and Hal 2010). To eliminate library size factors, reads per ten million mapped reads (RPTM) were calculated by normalizing read coverage with the total number

of properly mapped reads in each dataset.

The map function in BEDTools version 2.24 (Quinlan and Hal 2010) was used to determine the RNA expression at PE regions. The map function maps the strand-specific read alignment file of each dataset to the identified PE regions (total of 1919 PE regions). The entire 3-kb region of all the PEs was divided into 100 intervals (30-nt per bin), and the accumulation of normalized read density (RPTM, as a measurement of RNA expression) for each RNA-seq dataset (described above) was tabulated per 30-nt bin. The strand-specific read density in each of 100 intervals in the entire 3-kb PE region was graphed in metagene-like plot relative to the PE midpoint. The data for each biological replica derived from the same condition and library type were merged to represent the transcriptional activity per data series, including poly(A)$^+$ WT, ribo-minus WT, poly(A)$^+$ *rrp41-i*, and ribo-minus *rrp41-i*. The RNA expression at the top (forward) strand and bottom were analyzed and depicted independently.

Microarray data analysis

The analysis of microarray data (ChIP-chip) was adapted from Charron *et al.* (Charron et al. 2009). Briefly, quantile normalization and a Hidden Markov model (HMM) were used. A minimal run of 100 nt was required, while allowing a maximal gap of 100 nt, and the cutoff of ≥0.5 posterior probabilities with peak containing 4 probes on average. Default settings were used for the remaining parameters. By analyzing raw ChIP-chip data, the HMM posterior probability scores per specific ChIP were tabulated genome-wide. These scores represent the signal intensities per specific ChIP, and the signal intensities were used in metagene-like plots in the subsequent analysis. Sets of genome regions (peaks) representing significant enrichment of specific histone modifications and Pol II occupancy were also identified.

Analysis of transcriptionally active PEs

To determine if the identified PEs are actively transcribed, RNA expression of both the poly(A)$^+$ and ribo-minus RNA data series in both wild type and mutant conditions were tabulated in each PE region. The PEs were considered actively transcribed if the region (3-kb) had at least 1 RPTM (reads per ten million, >1 RPTM) in

either the forward or reverse direction. BEDTools version 2.24 (Quinlan and Hal 2010) was used to analyze the transcriptional activity of the PE regions.

The differential RNA expression in the poly(A)$^+$ datasets of each PE region was detected by DESeq2 version 1.8.2 (Love et al. 2014). The RNA expression originating from the top (forward) and bottom (reverse) DNA strands was analyzed separately. The differentially expressed PE regions in the exosome-deficient dataset (*rrp41-i*) versus the corresponding control (RRP41) were determined using the following selection criteria: adjusted P value (Padj) < 0.1 and $\log_2$(fold change) ≥ 0.5849 for regions with more than 1.5 fold up-regulation or $\log_2$ (fold change) ≤ 0.5849 for regions with more than 1.5 fold down-regulation.

Analyzing the TF binding motifs in the PE regions (MEME-FIMO)

To quantify the TF binding motifs that occurred in PE regions, we extracted 499 experimentally characterized TF binding motifs curated from more than 1700 TFs and 50 TF families from the AGRIS dataset. These TF binding motifs ranged from 5 to 49 nt in length with an average of 12.5 nt. FIMO (Find Individual Motif Occurrences) (Grant et al. 2011) was used to scan for known TF binding motifs within the entire length of PE regions that were statistically significant. The forward and reverse sequences of all PE regions were scanned for TF binding motifs and this identified 699,448 statistically significant motif occurrences (statistical threshold, $p < 0.0001$). The *p*-values were also determined based on multiple testing correction by the FIMO package.

Next, all the TF binding motifs were subjected to metagene analysis to determine their relative positions in PE regions, specifically, relative to the PE midpoints (orange dotted lines). Detailed metagene analysis at the PE regions was done using BEDTools version 2.24 (Quinlan and Hal 2010). Briefly, the entire 3-kb region of all the PEs was divided into 100 intervals (30-nt per bin), and the frequency of occurrences of the binding motif for each TF was tabulated per 30-nt bin. The TF binding motif occurrences in each interval in the entire 3-kb PE region were graphed in a metagene-like plot relative to the PE midpoint (Fig. 2A).

The top 20 most frequent TF binding motifs were analyzed further, and subjected to metagene analysis individually. The motif sequences that resemble known *cis*-regulatory elements, including the PAS-like and TATA-box like sequences, were also identified. All the TFs that had binding motifs in the PE regions were grouped into their respective TF families, which were determined based on the presence of conserved domains.

Additionally, a previous study (Zhang et al. 2012) showed that two MADS domain TFs could bind to DHSs, suggesting that *cis*-elements can be identified in DHSs. However, this study focused only on a subset of open chromatin regions located in promoter regions.

Analysis of peaks of TF binding in the PE regions (DAP-seq)

To determine TF footprints at PE regions, all peaks of TF binding in PE regions were extracted from the DAP-seq peak calling dataset for all 522 TFs ($p < 0.0001$, Bedtools Fisher's exact test). Fisher's exact test was used to test the statistical significance of data associations with a set of randomized controls; the genomic interval based BEDtools Fisher version 2.24 (Quinlan and Hal 2010), which compared a simulated random set of similar genomic regions (control, TAIR10 genome-based) versus the observed enrichment, was used. Next, all the peaks of TF binding in PE regions were subjected to metagene analysis to determine the spatial relationship at PE regions, specifically, relative to the PE midpoints. Detailed metagene analysis at the PE regions was done using BEDTools version 2.24 (Quinlan and Hal 2010). The peaks of TF binding specifically in close proximity to PE midpoints (midpoint-proximal region) were analyzed further. This analysis also resulted in identification of 5 TFs (represented by the red bar in Fig. 2E, $p < 0.0001$, chi-square test) that bind to >600 PEs in the midpoint-proximal regions (±500 nt from the midpoints).

The binding profile of each of these 5 TFs was subjected to the metagene analysis individually to determine the overall distribution of the sites bound by each TF in the PE region. The binding sites of each TF in the entire 3-kb PE region and their binding sites specifically at the midpoint-proximal regions of the PE (±500 nt from the midpoints) were analyzed separately. The sequences of these binding sites were subjected to motif discovery analysis using MEME-ChIP (Machanick and Bailey 2011) with default parameters (E-value ≤ 0.05) to identify the three most enriched binding motifs, which represent consensus TF binding sequences in the PE regions. For each motif MEME-ChIP finds, an E-value is reported. E-value represents an estimate of the number of motifs expected to be found by chance if the letters of the input sequences were shuffled. Small E-value (e.g. less than 0.05) is assigned to motifs that are statistically significant and thus very unlikely to be random sequence artifacts.

In comparison, the general binding motifs for each TF, which were non-specific to genomic context and determined by interrogating all peaks of binding genome-wide (determined previously by O'Malley et al. 2016) were separately analyzed and depicted.

Analysis of the genomic locations of the identified PEs and the landscape of TSSs and TESs of protein-coding genes relative to intra-genic PEs

The protein-coding genes that overlap with intra-genic PEs were subjected to GO analysis. The libraries of extensive GO terms for *Arabidopsis* TAIR10 genome were extracted from TAIR (The Arabidopsis Genome Initiative 2000) (file: ATH_GO_GOSLIM.txt). Enrichment analysis of GO category was done in R (version 3.4.0). Fisher's exact test was utilized for assessing the GO term significance. The *p*-value and odd ratio are listed in Dataset S10. The enriched GO terms were retrieved for the protein-coding genes overlapping intra-genic PE regions, and the terms were grouped based on biological processes, molecular functions, and cellular components.

The distance and the directionality of TSSs of protein-coding genes relative to PE midpoints were determined using BEDTools version 2.24 (Quinlan and Hal 2010). Only the intra-genic PEs (*n*=1348) that overlap exclusively with protein-coding genes were subjected to the analysis. To specifically analyze the landscape of TSSs and TESs relative to intra-genic PEs, a heatmap like visualization was adapted to illustrate the localization of TSSs and TESs relative to the PE midpoints. Transcription start sites (TSSs) refer to the 5′ end and transcription end sites (TESs) refer to the 3′ end of TAIR 10 protein-coding gene annotations.

First, the intra-genic PEs were classified into four groups based on the numbers of overlapping protein-coding genes. Only Groups 1 and 2 were used for the following analysis. Second, the entire protein-coding gene units were mapped relative to the PE midpoints. The colors of protein-coding gene units were represented in a red to blue color scale based on COLOR BREWER (Harrower and Brewer 2003) , with red referring to 5′ end (TSS) and blue referring to 3′ end (TES) of protein-coding genes. Different ranges of distance from PE midpoints were analyzed, including ±1500 nt, ±500 nt, and ±100 nt, to narrow the viewing window relative to the midpoint.

Analysis of intrachromosomal interactions at PE regions

The Hi-C dataset consist of 5773 pairs of intrachromosomal interactions, with each interacting region being 2-kb in length (the resolution of the experiment) (Wang et al. 2015). BEDTools version 2.24 (Quinlan and Hal 2010) was used to determine if there were any intersection between identified PE regions and the intrachromosomal interactive regions identified by Hi-C experiment. Only the PE regions, which overlapped with intrachromosomal interactive regions more than 100 nt were considered to avoid false-positive results. Therefore, a total of 452 PEs that can interact intrachromosomally with PE-interacting loci (eILs) were identified; statistical significance testing by Fisher's exact test (implemented in BEDtools Fisher function (Quinlan and Hal 2010) (24%; p = 1.7e-14, odd-ratio = 1.5, Fisher's exact test). Fisher's exact test was used to test the statistical significance of data associations with a set of randomized controls; the genomic interval-based BEDtools Fisher version 2.24 (Quinlan and Hal 2010), which compared a simulated random set of similar genomic regions (control, TAIR10 genome-based) versus the observed enrichment, was used. These PEs form a total of 1076 possible PE-eIL interactions (chromatin loops). A total of 879 PE-interacting eILs was identified, which originated from a total number of 716 non-redundant interacting loci. The interactions based on their size range (local interactions, intermediate-range interactions and long-range interactions), and the number of total interactions per PE based on the size of the loops were determined.

Analysis of eILs, the PE interacting loci

To determine if PEs can interact with specific component of protein-coding genes, the positions of eILs were mapped against protein-coding gene units genome-wide using BEDTools version 2.24 (Quinlan and Hal 2010). These components of protein-coding gene unit include 5′UTRs, 3′UTRs, and the gene body. The protein-coding genes co-localized with the identified eILs were subjected to gene ontology (GO) analysis as described above. Additionally, the structures of protein-coding genes annotated with specific enriched GO terms (plasma membrane-associated and integral component of the membrane) were analyzed.

Analysis of epigenomic features at eILs and PE regions

The collection of epigenetics scores of the previously published ChIP-seq and ChIP-chip data collectively analyzed by Wang et al. 2015 were used to tabulate the epigenetics scores at genomic region corresponding to

eILs and PEs. An integer score of 1 to 4 was given to each of 400 bins across the entire genome proportionally to the quintile enrichment in that bin as the epigenetics scores per specific chromatin modifications. The epigenetic scores at genomic regions corresponding to eILs and PEs were extracted and analyzed using BEDTools version 2.24 (Quinlan and Hal 2010) and visualized using heatmap.2 function in the gplot package version 3.0.1 for R (http://cran.r-project.org/package=gplots).

DNA sequence conservation in PE regions

A total of 1135 out of the entire collection of 1227 genome accessions were extensively sequenced and analyzed by the 1001 *Arabidopsis thaliana* genomes project (1001 Genomes Consortium 2016). The associated variant call format file (VCF), which contains all sequence variants at each genomic position present in all the sequenced 1135 genome accessions in a meta-information manner (1001genomes_snp-short-indel_only_ACGTN.vcf.gz) was published. The sequence variants, also referred to as allelic variants (AVs), were determined genome-wide when compared to the DNA sequences of the entire genome with the reference genome sequences (Col-0). The 10 selected accessions in the high-diversity panel (Kawakatsu et al. 2016) were analyzed, and their corresponding sequence variants recorded in the VCF file were extracted and analyzed. VCFtools version 0.1.14 (Danecek et al. 2011) were used to extract the sequence variants at the PEs regions, and the extracted sequence variants were parsed. The DNA sequence conservation of identified PEs was measured as frequency of AVs per PE region in all genomes. To expand the analysis to the entire collection of 1135 genome accessions, all possible AVs in all 1135 accessions at corresponding PE regions were extracted from the VCF file (1001genomes_snp-short-indel_only_ACGTN.vcf.gz), and the AVs in each PE region per genome were analyzed independently. To compare the average AV at PE regions with the expected average AV across the entire genome previously calculated in a genomic context-independent manner, the frequency of AVs in each PE region per entire collection of 1135 genomes was determined. One-sample *t*-test was used to determine the significance of the difference between average AV at PE regions and expected average AV across the entire genome in a genomic context-independent manner. The *p*-value resulted from the statistical calculation was less than 0.0001.

SI REFERENCES

1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**: 481–491.

Ahmad K, Henikoff S. 2002. Histone H3 variants specify modes of chromatin assembly. *Proc Natl Acad Sci USA* **99 Suppl 4**: 16477–16484.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077.

Calo E, Wysocka J. 2013. Modification of enhancer chromatin: what, how, and why? *Molecular Cell* **49**: 825–837.

Campisi L, Yang Y, Yi Y, Heilig E, Herman B, Cassista AJ, Allen DW, Xiang H, Jack T. 1999. Generation of enhancer trap lines in Arabidopsis and characterization of expression patterns in the inflorescence. *Plant J* **17**: 699–707.

Charron J-BF, He H, Elling AA, Deng XW. 2009. Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. *The plant Cell* **21**: 3732–3748.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen P-Y, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, et al. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**: 388–392.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

ENCODE Project Consortium, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Lajoie BR, Landt SG, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Goldberg AD, Banaszynski LA, Noh K-M, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X, et al. 2010. Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell* **140**: 678–691.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

Harrower M, Brewer CA. 2003. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. **39**: 311–318.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26**: 1293–1300.

Ji H, Vokes SA, Wong WH. 2006. A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. **34**: e146–e146.

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. 2009. H3.3/H2A.Z double variant-containing nucleosomes mark “nucleosome-free regions” of active promoters and other regulatory regions. **41**: 941–945.

Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al. 2016. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166**: 492–505.

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

Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, Rosen BD, Cheng C-Y, Moreira W, Mock SA, et al. 2015. Araport: the Arabidopsis information portal. **43**: D1003–9.

Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AMS, Strahl BD, Patel DJ, Jacobsen SE. 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**: 385–389.

Liu P-P, Koizuka N, Homrichhausen TM, Hewitt JR, Martin RC, Nonogaki H. 2005a. Large-scale screening of Arabidopsis enhancer-trap lines for seed germination-associated genes. *Plant J* **41**: 936–944.

Liu P-P, Koizuka N, Martin RC, Nonogaki H. 2005b. The BME3 (Blue Micropylar End 3) GATA zinc finger transcription factor is a positive regulator of Arabidopsis seed germination. *Plant J* **44**: 960–971.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**: 1696–1697.

O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**: 1280–1292.

Ooi SL, Henikoff JG, Henikoff S. 2010. A native chromatin purification system for epigenomic profiling in Caenorhabditis elegans. **38**: e26–e26.

Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, Federation A, Chao J, Elliott O, Liu Z-P, et al. 2015. RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell* **161**: 774–789.

Quinlan AR, Hal IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Sein H, Värv S, Kristjuhan A. 2015. Distribution and maintenance of histone H3 lysine 36 trimethylation in transcribed locus. *PLoS ONE* **10**: e0120200.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* **15**: 272–286.

Stroud H, Otero S, Desvoyes B, Ramírez-Parra E, Jacobsen SE, Gutierrez C. 2012. Genome-wide analysis of histone H3.1 and H3.3 variants in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **109**: 5370–5375.

Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, Martienssen R. 1995. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes & Development* **9**: 1797–1810.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796–815.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562–578.

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

Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, Lanz C, Weigel D. 2015. Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Research* **25**: 246–256.

Weber B, Zicola J, Oka R, Stam M. 2016. Plant Enhancers: A Call for Discovery. *Trends in Plant Science* **21**: 974–987.

Wierzbicki AT, Cocklin R, Mayampurath A, Lister R, Rowley MJ, Gregory BD, Ecker JR, Tang H, Pikaard CS. 2012. Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes & Development* **26**: 1825–1836.

Wollmann H, Holec S, Alden K, Clarke ND, Jacques P-É, Berger F. 2012. Dynamic deposition of histone variant H3.3 accompanies developmental remodeling of the Arabidopsis transcriptome. ed. T. Kakutani. *Plos Genetics* **8**: e1002658.

Yang W, Jefferson RA, Huttner E, Moore JM, Gagliano WB, Grossniklaus U. 2005. An egg apparatus-specific enhancer of Arabidopsis, identified by enhancer detection. *Plant Physiology* **139**: 1421–1432.

Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E. 2011. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. **39**: D1118–22.

Zentner GE, Henikoff S. 2013. Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol* **20**: 259–266.

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

Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biol* **10**: R62.

Zhang X, Clarenz O, Cokus S, Bernatavichute YV, Pellegrini M, Goodrich J, Jacobsen SE. 2007. Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *Plos Biology* **5**: e129.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.