Epigenetic regulation of spurious transcription initiation in *Arabidopsis*

Ngoc Tu Le¹, Yoshiko Harukawa¹, Saori Miura¹, Damian Boer², Akira Kawabe³ & Hidetoshi Saze¹

In plants, epigenetic regulation is critical for silencing transposons and maintaining proper gene expression. However, its impact on the genome-wide transcription initiation landscape remains elusive. By conducting a genome-wide analysis of transcription start sites (TSSs) using cap analysis of gene expression (CAGE) sequencing, we show that thousands of TSSs are exclusively activated in various epigenetic mutants of *Arabidopsis thaliana* and referred to as cryptic TSSs. Many have not been identified in previous studies, of which up to 65% are contributed by transposons. They possess similar genetic features to regular TSSs and their activation is strongly associated with the ectopic recruitment of RNAPII machinery. The activation of cryptic TSSs significantly alters transcription of nearby TSSs, including those of genes important for development and stress responses. Our study, therefore, sheds light on the role of epigenetic regulation in maintaining proper gene functions in plants by suppressing transcription from cryptic TSSs.
Eukaryotic genomes are comprised a large part of mobile genetic sequences, so-called transposable elements (TEs). Due to their mobility, TEs induce various alterations to the host genome, ranging from genetic mutations to large-scale genomic rearrangements, such as inversions and translocations.

Genetic variations caused by TEs can introduce novel regulatory elements and therefore be a major driving force underlying genome evolution. On the other hand, uncontrolled activities of TEs can severely damage gene expression and the integrity of the host genomes.

To suppress negative impacts without losing potential benefits brought in by TEs, both plants and animals have evolved numerous epigenetic mechanisms involving DNA methylation, histone modifications, and small non-coding RNAs, allowing TEs remain silenced in their genomes. Compared to mammals, plants are equipped with a different set of epigenetic mechanisms for greater adaptability to dynamic environmental changes, partly due to their sessile nature. For example, in mammalian genomes, DNA sequences are mainly methylated at the cytosine in the CG dinucleotides, while in plants cytosine methylation exists in both CG and non-CG contexts, which has different functional impacts on gene and TE regulation.

In the plant model Arabidopsis thaliana, DNA methylation is established de novo by the RNA-directed DNA methylation pathway, which requires the functional activity of PolIV and PolV, two plant-specific RNA polymerases. After establishment, methylation patterns can be maintained by different factors depending on cytosine contexts. CG methylation is maintained by METHYLTRANSFERASE 1 (MET1), a plant homologue of the mammalian DNA (cytosine-5)-methyltransferase 1 (DNMT1). Maintenance of DNA methylation in CHG context, on the other hand, is facilitated by CHROMOMETHYLASE 3 (CMT3) in a positive feedback loop with the histone H3K9 methylase KRYPTONITE (KYP) (or SUPPRESSOR OF VARIEGATION 3-9 HOMOLOGUE 4 (SUHV4))7. Together with two of its paralogues, SUVH5 and SUVH6, KYP regulates the genome-wide accumulation of H3K9me2 and consequently, CHG methylation. CHH methylation can be maintained by either CMT2 or DOMAINS REARRANGED METHYLASE 2 (DRM2) depending on the features of their targets, in which DRM2 often methylates short, euchromatic TEs, while CMT2 targets long TEs located in histone H1-containing heterochromatic regions with the help of chromatin remodeler DECREASED DNA METHYLATION 1 (DDM1)8. These epigenetic pathways in plants, however, are highly interwoven. For example, MET1 and CMT3 are involved in maintaining asymmetric methylation, while DRM2 and CMT2 may also affect DNA methylation in other contexts.

Epigenetic silencing of TEs inevitably confers regulatory impacts on gene expression, especially when TEs are located close to transcription units. In plants, repressive modifications triggered by TE insertions within introns or promoter regions can attenuate or even turn off the expression of the associated genes. At a global scale, genes harboring, or located close to, silenced TEs exhibit lower expression than their counterparts. Due to such unfavorable impacts, plants have evolved specific pathways to keep transcription units clear of repressive modifications, or to tolerate the presence of such modifications when necessary. For example, in A. thaliana the Junomii C (jinJC) domain-containing histone demethylase INCREASE IN BONSAI METHYLATION 1 (IBM1) prevents repressive H3K9 methylation and consequently, CHG methylation, from accumulating at actively transcribed genes. On the other hand, host factors, such as INCREASE IN BONSAI METHYLATION 2 (IBM2) and Enhanced Downy Mildew 2 (EDM2) are required for proper transcription of genes containing heterochromatic domains, likely due to the functional importance of these domains.

The development of high resolution 5′ end-centered expression profiling techniques, such as oligo-capping methods or cap analysis of gene expression (CAGE), has greatly advanced our understanding of gene regulation at a transcription initiation level. Studies employing these techniques have revealed both common and distinct features of the core promoters and their origin and regulation, in many organisms. In mammals, for example, CAGE sequencing (CAGE-seq) analyses revealed that a large fraction of cell-type specific transcripts in stem and cancer cells originate from long terminal repeats (LTRs) of retroelements. The loss of DNA methylation also causes spurious transcription within thousands of genes in mouse embryonic stem cells. In addition, modulating DNA methylation and histone deacetylation pathways pervasively activates cryptic transcription start sites (TSSs) normally silenced in human cells. These examples demonstrate the importance of epigenetic mechanisms in regulating transcription initiation in mammalian genomes.

In plants, large-scale analyses have determined thousands of TSSs, providing fundamental information about genetic structure and regulatory elements important for transcription in plant genomes. Previous studies have also revealed core promoter structures and sequence elements associated with plant TSSs. However, these studies mainly focus on active TSSs in the wild-type background. The contribution of epigenetic regulation to shaping the genome-wide transcription initiation landscape and its functional significance in plants, therefore, remains largely unexplored.

To dissect the functional impacts of epigenetic regulation in shaping the plant transcription initiation landscape, we employ CAGE-seq to generate the genome-wide profiles of TSSs at a high resolution for various mutants of A. thaliana that compromise epigenetic control. Our analysis identifies thousands of TSSs exclusively activated in the mutant backgrounds, demonstrating that epigenetic regulation profoundly affects transcription initiation in Arabidopsis. These so-called cryptic TSSs are mainly located at heterochromatic regions, which hinder their accessibility to RNA Polymerase II (RNAPII) transcription machinery. The alteration of DNA methylation maintenance in met1 activates the largest number of cryptic TSSs, which significantly overlap with the targets regulated by other epigenetic pathways. A large fraction of cryptic TSSs originate from TEs of both retro and DNA-transposon families, suggesting that TEs are reservoirs of putative TSSs in the A. thaliana genome. Strikingly, the activation of cryptic TSSs significantly alters the regular transcription of nearby TSSs, which includes those of genes important for development and stress responses in Arabidopsis. This study, therefore, sheds light on the role of epigenetic regulation in maintaining proper gene functions in plants by suppressing transcription initiated from cryptic TSSs. In addition, the accompanying data are a valuable resource for studying the epigenetic control of the transcription of genes and TEs in plants.
mapping efficiency of 97.53%. Of which, 402,814,394 reads were mapped uniquely, compiling a large collection of CAGE-seq data for this model plant (Supplementary Data 1).

The expression of individual CAGE-based TSSs (CTSSs) was highly correlated between replicates (the median of Pearson correlation coefficients was 0.95) (Supplementary Fig. 1a, b), confirming the reproducibility of our data. In total, 37,726 consensus tag clusters representing single TSSs were identified across all samples (hereafter TSSs is used to refer to consensus tag clusters identified in this study, to distinguish from the TAIR-annotated TSSs), of which about 30% were exclusively expressed in the mutant backgrounds (Supplementary Data 2).

To confirm the relevance of our data, we analyzed the genome distribution of 26,561 TSSs identified in wild-type sample. A majority of them (18,634 or ~70%) were located in promoters and 5' UTRs of 17,722 (~64%) annotated genes (Fig. 1a), and about one-fourth (~24%) were located in intragenic regions, of which exonic TSSs were more prevalent than the intronic counterparts. Although the mechanisms leading to the prevalence of exonic TSSs in the plant genomes have yet been clear31, a part of them may represent 5’-end capped products of post-transcriptional processing of mature mRNAs, as described in human and vertebrate genomes2,3,33. Alternatively, some may correspond to cryptic promoters that trigger spurious transcription from gene bodies5,34, or to mis-annotated TSSs22. Nevertheless, consistent with a previous study21, the expression of intragenic TSSs was significantly lower than that of their counterparts located in promoters and 5’ UTRs (Fig. 1b). Moreover, the TSSs in promoters and 5’ UTRs were found in close proximity to the TAIR10-annotated TSSs (Supplementary Fig. 2a, b). A similar result was obtained using the Araport11 genome annotations, (Supplementary Fig. 3a–c), with a shift in the numbers of TSSs assigned to each genome feature (Fig. 1a, Supplementary Fig. 3a). Because of the higher consistency with the TSSs identified by our CAGE-seq (Supplementary Figs. 2a, b, 3b, c), TAIR10 annotations were used in further downstream analysis. On the other hand, active genes supported by CAGE and mRNA-seq were largely overlapped (Supplementary Fig. 2c), suggesting that active transcription events in A. thaliana can be efficiently captured by our CAGE-seq data.

We then compared wild-type TSSs identified by CAGE-seq with those reported by the paired end analysis of transcription start site (PEAT) method31. They were indeed consistent even though the samples were prepared from different tissues (Supplementary Fig. 3d–f). At a local scale, the promoter architecture of two well-studied genes, ALMT1 (AT1G08430) and sAPX (AT4G08390), was also reexamined. The former has three functional TSSs within its promoter and the latter has one upstream and one intragenic TSS21. Our data recapitulated these structures (Supplementary Fig. 4a, b), confirming its consistency with previous studies21,31.

It has been found that the loss of CG methylation at a SINE-related repeat in the promoter region triggered the ectopic expression of the homeobox gene FLOWERING WAGENINGEN (FWA), causing a late flowering phenotype of Arabidopsis11,35,36. CAGE-seq analysis identified a TSS encoded within the SINE repeat, which was highly activated in met1 and ddm1 backgrounds (Fig. 1c). In addition, the ectopic activation of the TSS of the F-box gene SUPPRESSOR OF DRM1 DRM2 cm3t (SDC), whose promoter contains a tandem repeat co-regulated by H3K9 methylation and the RdDM pathway8, was also detected by our data (Fig. 1c).

Taken together, these results demonstrate that our CAGE-seq data can be effectively exploited for the detection and analysis of both regular and cryptic TSSs under epigenetic control. Modulating epigenetic regulation activates many cryptic TSSs.

Next, we investigated the impact of epigenetic regulation on the transcription initiation landscape in the A. thaliana genome in greater details. Compromising epigenetic controls significantly affected the transcription initiated from hundreds to thousands of TSSs, in which the defect of the maintenance DNA methylation pathway in met1 induced changes at the largest number of targets (Fig. 2a), followed by ibm1, ddm1, suv-h456, and pol4. To our surprise, ibm2 and edm2, which cause the transcriptional defect of IBM116,17, had a lower number of affected TSSs than ibm1, suggesting that the IBM1 function is partially maintained in these mutants.

Of the altered TSSs, many were activated de novo in the mutant backgrounds and were not associated with any TAIR10-annotated TSSs (Fig. 2a, Supplementary Fig. 2b, Supplementary Data 3). They were also largely distinct from the TSSs reported by PEAT-seq31 and the TSSs identified in multiple tissues and light stress conditions in A. thaliana31 (Supplementary Fig. 5a, b), suggesting that they are cryptic TSSs suppressed by epigenetic mechanisms (referred herein as EPICATs, for EPigenetically Induced Consensus Transcript clusters). Our data showed that the EPICATs activated in met1 largely overlapped with the EPICATs regulated by other mutants, confirming the profound regulatory impact of MET1 on the genome-wide transcription initiation in A. thaliana (Fig. 2b). On the other hand, ddm1 and RdDM-associated mutants (pol4 and pol5) induced stronger activation of the EPICATs than met1 (Fig. 2c). Due to the minor numbers of instances, targets of ibm2 and edm2 were excluded from further analysis. Similar results were obtained using the Araport11 annotations (Supplementary Figs. 3c, 5c), confirming the robustness of our analysis.

As the transcription orientation at regulatory regions of eukaryotes can be either unidirectional20 or bidirectional37, we examined the directionality of transcription initiated at EPICATs. Our data showed that transcription at the EPICATs in met1 was mainly uni-directional, similar to that of the TAIR10-annotated TSSs in A. thaliana (Supplementary Fig. 6a20). Moreover, the expression levels of EPICATs were not significantly different from those of the annotated TSSs activated de novo in epigenetic mutants (Supplementary Fig. 6b). We also found that, tag clusters corresponding to the EPICATs mainly had narrow peaks (NPs), especially those activated in ddm1, met1, and pol5 (Supplementary Fig. 6c), suggesting that they may have a well-defined underlying genetic architecture13,38.

To elucidate putative mechanisms regulating the activity of EPICATs, we first examined the genomic regions where they reside. EPICATs were mainly located at intergenic regions, except the EPICATs in ibm1, of which a majority were intragenic (Fig. 3a, Supplementary Fig. 6d). These intragenic EPICATs, however, may not be directly regulated by the activity of IBM1, because they were not associated with increased CHG methylation in the ibm1 background (Fig. 3b). In contrast, the EPICATs in other mutants were located in genomic regions decorated with repressive chromatin modifications, such as DNA methylation, H3K9me2, and H3K27me1 (Supplementary Fig. 7a, b). Compared to the EPICATs in other mutants, those activated in pol4 and pol5 were also associated with a higher level of CHH methylation and 24 nt siRNAs, the hallmarks of the RdDM pathway (Supplementary Fig. 7a, c). Moreover, DNA methylation at the EPICATs in all mutants, except in ibm1, was significantly reduced, in concomitant with their activation (Fig. 3b, Supplementary Fig. 7d), suggesting that in wild-type plants transcription initiation at EPICATs is directly suppressed by repressive epigenetic modifications.

Since heterochromatic modifications, such as DNA methylation and H3K9me2, are often associated with closed chromatin in
plant genomes\textsuperscript{39}, their loss may alter the access to genomic regions harboring EPICATs. We therefore examined how the accessibility of these loci changes in the mutant backgrounds. For this purpose, the EPICATs activated in \textit{ddm1} were used as a proxy due to the large number of instances and the availability of public data characterizing chromatin openness in \textit{ddm1} \textsuperscript{40}.

Indeed, chromatin around the EPICATs became highly accessible in \textit{ddm1}, compared to wild-type plants, as measured by the sensitivity to DNaseI (Fig. 3c). Furthermore, ChIP-seq analysis showed that RNAPII phosphorylated at Ser5 (Ser5P) and Ser2 (Ser2P) in the C-terminal domain (CTD), the hallmarks of transcription initiation and elongation\textsuperscript{41} respectively, were also highly accumulated at the EPICATs in most mutant backgrounds (Fig. 3d, Supplementary Fig. 7e). These data demonstrate that repressive chromatin suppresses the activity of EPICATs by preventing the access of transcription machinery to genomic regions encompassing potential TSSs.

Ectopic transcription initiation in mutants and the convergence of various epigenetic pathways on a large number of EPICATs (Fig. 2b), together with the narrow shapes of tag...
clusters corresponding to most of the EPICATs (Supplementary Fig. 6c), suggest that these loci harbor functional genetic features, such as promoter structure and/or regulatory sequences, in addition to repressive chromatin modifications. Therefore, genetic sequences surrounding EPICATs were analyzed. Interestingly, DNA elements and motifs enriched around EPICATs exhibited spatial architecture similar to that of regular plant promoters, with a sharp accumulation of TATA-box at 36 nt upstream and CA-rich/CT-rich (Y-patch) motifs around the TSSs (Fig. 3e, Supplementary Fig. 8). TATA-box, a core promoter motif conserved in both plants and animals, was especially enriched at the EPICATs in *met1* and *ddm1*. The enrichment of the Telobox motif (AAACCCTA), which is known to recruit development-associated repressive modification H3K27me3 in *A. thaliana*, was also found at the EPICATs in *met1*, *ddm1*, and *suvh456*. The presence of the Telobox sequence around EPICATs may partially explain the accumulation of H3K27me3 at the heterochromatic regions upon the loss of DNA methylation and H3K9 methylation.

Taken together, we conclude that the *A. thaliana* genome harbors hundreds of potential TSSs equipped with functional core promoter architecture similar to that of regular TSSs. Their activities, however, are suppressed by repressive chromatin restricting their accessibility to transcription machinery.

**Gene body methylation and the suppression of intragenic TSSs.** In *A. thaliana*, about 20% of protein coding genes accumulate CG methylation in their bodies. Moreover, gene body methylation (gbM) is largely conserved across plant species, especially in angiosperms, suggesting its functional importance. Although many hypotheses have been proposed regarding the biological functions of gbM, such as suppressing spurious intragenic transcription, impeding transcriptional elongation, or reducing transcription noise, so far its role in plants has been largely elusive. By exploiting the high resolution CAGE-seq data of genome-wide TSSs, we reexamined the relationship between gbM and intragenic transcription initiation in *A. thaliana*. Our data showed that, in wild-type plants, a similar fraction of both body methylated (BM) and non body methylated (non-BM) genes harbored intragenic TSSs, suggesting that the methylation state of gene body is not significantly associated with the occurrence of...
Fig. 3 Features of the EPICATs activated in epigenetic mutants. a Genome-wide distributions of EPICATs. Genomic features were ordered as described in Fig. 1a. The number of the EPICATs activated in each mutant is given in the parentheses. b Change of CHG methylation, a hallmark of repressive chromatin, at EPICATs in epigenetic mutants. Methylation levels were calculated for 101 bp regions centering around the dominant CTSSs of the EPICATs. ***p < 2.2e-16; N.S.: Not Significant (p > 0.01), given by paired two-sided Student t-test. c Increase of chromatin accessibility at the EPICATs in the ddm1 mutant compared to wild-type (wt) plants, measured by DNase-seq data in flower (F) and leaf (L) tissues. Signals were calculated for non-overlapping 50 bp windows in the regions of ±2 Kb centering around the EPICATs, aligned by their dominant CTSSs (indicated by TSS), and then sorted by the average value of each row. d Ectopic recruitment of transcription machinery, represented by RNAPII phosphorylated at Serine 2 (Ser2P) and Serine 5 (Ser5P), to the EPICATs in ddm1, at the genome-wide scale (heatmap, left) and a representative locus (indicated by orange windows in the browser tracks, right). ChIP-seq signals in the heatmap were calculated for non-overlapping 50 bp windows in the regions of ±2 Kb centering around the EPICATs, aligned by their dominant CTSSs (indicated by TSS), and then sorted by the average value of each row. e Top 5 enriched motifs (left) and their spatial arrangement (right) at the EPICATs in met1. Significant E-value and the motif logos were given by de novo motif analysis. Distances were calculated between the midpoints of motif instances (p-value ≤ 1e-04) and the dominant CTSSs of the EPICATs, and the average profile of all motif instances is shown.
intragenic TSSs (Fig. 4a). Moreover, only a few BM genes activated intragenic EPICATs when gbM was strongly lost in met1 background (Fig. 4b, Supplementary Fig. 9a), meanwhile intragenic EPICATs could be activated at some loci without gbM (Fig. 4d). These evidences, which are consistent with the conclusions of a previous study, suggested that gbM alone is dispensable for suppressing intragenic transcription at a global scale in A. thaliana (Supplementary Fig. 9b). Although some BM genes harbored intragenic EPICATs in met1 (Fig. 4c, d), at this time, we do not know if this is a direct or indirect effect of met1 mutant. Future testing using targeted demethylation could help resolve if BM is causal at these loci.

The intragenic EPICATs in met1 may correspond to 5′-end capped products of post-transcriptional processing of mature mRNAs generated at the associated gene loci, a mechanism well-described in mammals. Although we did not rule out this possibility, our data provided evidences supporting that some of these EPICATs are genuine TSSs. First, these loci exhibited a
stronger accumulation of RNAPII in met1 (Supplementary Fig. 9c). Second, only 1/124 genes harboring intragenic EPICATs also had upstream EPICATs (Supplementary Fig. 9d), suggesting that these intragenic EPICATs correspond to independent, de novo transcribed mRNAs. Third, promoter-associated DNA sequences were also present at some of these intragenic loci (Fig. 4d).

Besides met1, ibm1 also activated a comparable number of intragenic EPICATs (Supplementary Fig. 6d, Supplementary Data 4). However, it is unlikely that they are directly regulated by the activity of IBM1 (Fig. 3b, Supplementary Fig. 10a). On the other hand, although the expression of IBM1 is significantly reduced in met1 background49, the intragenic EPICATs activated in ibm1 and met1 were largely un-overlapped (Supplementary Fig. 10b). Moreover, the accumulation of RNAPII at these loci was not significantly affected in ibm1 background (Supplementary Fig. 10c), suggesting that intragenic EPICATs in ibm1 and met1 are regulated differently. Given that none of the associated genes simultaneously harbored upstream EPICATs, and that promoter-associated DNA sequences were present at some of these intragenic targets (Fig. 4e), we speculate that some of them are genuine TSSs, while some others could be derived from post-transcriptionally processed mRNAs.

RNAPII and PolIV exclusively bind to RdDM-regulated EPICATs. It has been reported that, although PolIV-dependent RNAs (P4RNAs) feature PolII-like TSSs, PolII and PolIV target distinct genomic territories49. Our data, however, showed that 24 nt siRNAs were highly enriched at genomic loci harboring the EPICATs activated in the mutants of the RdDM pathway’s components, such as pol4 and pol5 (Supplementary Fig. 7c). The biogenesis of these 24 nt siRNAs was indeed dependent on PolIV, which is responsible for the transcription of P4RNAs initiated from the corresponding EPICATs (Supplementary Fig. 11a–c). Moreover, in pol4 and pol5 backgrounds, RNAPII was highly recruited to these loci (Supplementary Fig. 7e). These evidences suggest that, genomic regions harboring the EPICATs regulated by the RdDM pathway likely possess distinct features compared to those of its regular targets, which allow PolII and PolIV exclusively function at these loci (Supplementary Fig. 11d).

TEs are a major supplier of cryptic TSSs in Arabidopsis. The existence of a large number of cryptic TSSs within a small and compact genome, like that of A. thaliana, has raised important questions regarding their origin. Investigations involving mammalian genomes have shown that TEs are a major genetic element that can be exapted as TSSs in the host genomes51,52. Although less prevalent, several lines of study have demonstrated a similar function of TEs in plant genomes53,54. Together with the evidence that EPICATs are mainly located at intergenic regions decorated with repressive chromatin modifications (Fig. 3a, Supplementary Fig. 7a, b), we speculated that many cryptic TSSs in the A. thaliana genome may have originated from TEs. The data indicated that TEs contribute to up to 65% of the EPICATs activated in the mutant backgrounds (Supplementary Fig. 12a). Additionally, hundreds of TEs harboring active TSSs were identified in wild-type background (Fig. 5a, Supplementary Data 5). TEs, therefore, may serve as a reservoir of potential functional TSSs in A. thaliana, similar to their role in animal genomes.

There are numerous types of TEs with different origins and mobility strategies1,2, which greatly affect their abilities to induce genetic variations to the host genomes. Therefore, the TSS-encoding potential of each TE family in the A. thaliana genome was examined. Although EPICATs were associated with various TE families (Fig. 5a), compared to the genome-wide average, LTR/Gypsy members were enriched among TEs harboring the EPICATs in ddm1 and met1 (p = 2.0e-52 and 6.0e-49, respectively, Hypergeometric test), while members of the LTR/Copia family were highly represented among the TE targets of suvh456 (p = 8.0e-10 and 2.0e-31, respectively, Hypergeometric test). In addition, the DNA/En-Spm family was highly associated with the EPICATs in met1, ddm1, and suvh546 (Fig. 5a, p < 1.6e-16 for all, Hypergeometric test). Due to the minor numbers of TE instances associated with the EPICATs in ibm1, pol4, and pol5, they were skipped from enrichment analysis. The data suggest that both retro- and DNA transposons are genetic suppliers of cryptic TSSs in the A. thaliana genome.

Since ddm1 affected the largest number of TEs harboring EPICATs, and these elements largely overlapped with TEs activated in other mutants (Fig. 5a, Supplementary Fig. 12b), we examined if they possess any specific features that facilitate their ectopic activation in ddm1 background. Compared to their counterparts, which either contain active TSSs in wild-type plants or do not harbor any EPICATs, TEs harboring EPICATs were more highly methylated in both CG and non-CG contexts (Fig. 5b). They were also substantially longer (Fig. 5b), suggesting that these TEs are likely younger insertions that still maintain intact structures with transcription and transposition capacities, that may be a trigger for greater accumulation of DNA methylation and other repressive modifications at the associated loci. Analysis of the core promoter motifs identified at the ddm1-activated EPICATs (Supplementary Fig. 8) showed that they were more prevalent among EPICAT-harboring TEs (Fig. 5c). However, there were still hundreds to thousands of inactive TEs associated with these motifs (Supplementary Fig. 12c). As a case study, the genetic structure associated with the EPICATs located in the LTR regions of the Gypsy TEs was investigated in a more detail. This was because the LTR/Gypsy family contributed a large number of elements harboring the EPICATs in ddm1 and met1...
(Fig. 5a), and its members still maintain transcription/transposition potential in the *Arabidopsis* genome. Although LTR sequences surrounding the CAGE-seq peaks were largely diverged between and within Gypsy sub-families, they commonly shared putative TATA-box and TSS-associated YR motifs (Fig. 5d, Supplementary Fig. 12d). However, the conservation of sequences/motifs surrounding the LTR-encoded TSSs could not fully explain their activation in the mutant backgrounds. Moreover, although a significant loss of repressive modifications (e.g., DNA methylation) was observed at many TEs regardless of their association with the EPICATs in *ddm1* (Fig. 5e), only EPICAT-harboring elements became highly accessible in the mutant, especially at their two ends (Fig. 5f). Concomitantly, RNAPII was highly recruited to these loci, together with an
Fig. 5 TEs are a major genetic supplier of cryptic TSSs in the A. thaliana genome. a Transposon families harboring the EPICATs. Shown are total numbers of TEs. All, wild-type: TEs in the A. thaliana genome and TEs harboring active TSSs in wild-type plants, respectively. *TE families enriched compared to the genome-wide average (p < 3e-04, Hypergeometric test). b Violin and/or boxplots showing the length (left), DNA methylation (middle), and Distance to the nearest gene (right), of TEs harboring (EPICAT+), not harboring (EPICAT–) the EPICATs in ddml, and harboring active TSSs in wild-type plants (wtACTIVE). **p < 2.2e-16, two-sided Mann–Whitney test. c Fractions of TEs harboring consensus DNA motifs found at the ddml-activated EPICATs. d Sequence alignment of representative LTRs of TEs belonging to the Gypsy ATGPI sub-family. LTR sequences were aligned based on positive positions of cryptic TSSs detected by CAGE-seq. A, C, G, T nucleotides were colored in green, yellow, red, and blue. Putative TATA-box and YR-motifs are indicated by black box and red line, respectively. Relative positions and strand information of each LTRs in the genome are indicated (bottom left). Heatmaps of the normalized CAGE-seq data mapped to the LTRs using one replicate of each mutants are shown on the right side of the bottom panel. e DNA methylation of TEs in the wild-type (wt) and ddml backgrounds. TE lengths were normalized to 1 Kb and aligned by their two ends (indicated by TSS and TES). Methylation levels were calculated for non-overlapping 100 bp windows within TE body and surrounding region of ±2 Kb, and sorted by the average value of each row. f Chromatin accessibility at TEs in ddml and wild-type (wt) plants, given by DNase-seq data in flower (F) and leaf (L) tissues. Signals were calculated and presented as described in e, with the window size of 50 bp. g Ecopic recruitment of RNAPII Ser2P and Ser5P to TEs in the wild-type (wt) and ddml plants. Signals were calculated and presented as described in e, with the window size of 50 bp.

Discussion

To understand how transcription initiation in plants is epigenetically regulated, we have generated a comprehensive maps of TSSs in various epigenetic mutants of A. thaliana using CAGE-seq. Compared to mammals, epigenetic mechanisms regulating transcription initiation in plants are much less clear, mainly due to a lack of suitable resources which allow the investigation of the alteration of transcription initiation under different conditions. This study, therefore, provides valuable reference data for research communities to enlighten the impact of epigenetic regulation on transcription initiation landscapes in plants.

Our study showed that, in epigenetic mutant backgrounds, thousands of cryptic TSSs are activated, in which the mutant of maintenance DNA methylation met1 regulates the largest number of targets (Fig. 2a). A large number of cryptic TSSs reside in TE sequences, which are dominantly contributed by members of the LTR/Gypsy, LTR/Copia, and DNA/En-Spm families (Fig. 5a). Interestingly, there is a clear difference in DNA methylation between TEs with and without EPICATs, where the former accumulate higher DNA methylation (Fig. 5b, c). This suggests that the DNA methylation of TEs could be largely influenced by their potential to initiate transcription. On the other hand, the analysis of LTR sequences indicated that the conservation of core promoter elements alone is not sufficient for transcription.
initiation (Fig. 5d, Supplementary Fig. 12d) as their transcription levels are largely varied, even among LTRs with nearly identical sequences. The ability of TE-encoded TSSs to initiate transcription may, therefore, also be dependent on their relative positions within TEs (e.g., whether they are located at the 5' or 3'-end of the TEs), and/or local chromatin environments, such as higher-order chromatin conformation and long-range enhancer interactions.

In mammals, the loss of gene-body DNA methylation caused by DNMT3b knockout triggers spurious RNAPII recruitment and cryptic transcription initiation from intragenic regions. The analysis of intragenic TSSs in the present study showed that a
complete loss of gbM in the \textit{met1} mutant does not profoundly activate intragenic transcription in the \textit{Arabidopsis} genome (Fig. 3a, 4, Supplementary Fig. 9a, b). Recruitment of DNMT3b to genic regions in mammals is dependent on histone H3K36 methylation\textsuperscript{62}. In yeast, H3K36 methylation (H3K36me) mediated by SET2 suppresses cryptic intragenic transcription initiation\textsuperscript{63}. In plants, however, concurrent loss of both gbM and H3K36me3 does not show significant difference in transcription between (BM) and unmethylated (UM) loci\textsuperscript{46}. On the other hand, regulation of cryptic transcription from intronic heterochromatin by the RdDM pathway\textsuperscript{64}, and the suppression of intragenic antisense transcripts by histone H1 and DNA methylation\textsuperscript{65} have also recently been reported. These results suggest that plants may employ additional layers of epigenetic regulation to prevent spurious transcription initiation, especially in intragenic regions.

The activation of spurious transcription from cryptic TSSs would inevitably alter transcription from nearby regular TSSs (Fig. 6, Supplementary Fig. 14). The data showed that such alteration may occur in several different scenarios. First, an activated cryptic TSS located upstream may function as the major initiation site facilitating the formation of a read-through transcript, which can suppress transcription from a downstream regular TSS, as observed at AT2G16050 and SQN loci (Fig. 6b). This regulatory effect is likely facilitated by a less understood mechanism known as transcriptional interference\textsuperscript{66,67}. Secondly, the activation of a cryptic TSS located downstream may attenuate transcription initiated from an upstream regular TSS and trigger the production of spurious transcripts, as observed at AT2G14580, AT2G15042, and AT5G28442 loci (Supplementary Fig. 14). Thirdly, when cryptic and regular TSSs are situated close to each other, but in divergent directions, transcription from the regular TSS may also be suppressed (Fig. 6d). Such repressive impacts could be facilitated by competitive binding to regulatory sequences of transcription initiation complexes associated with the two TSSs\textsuperscript{66}, or by the mechanism suppressing transcription from divergent promoters\textsuperscript{68}, or by the lack of a mechanism facilitating bi-directional transcription in plants\textsuperscript{69} compared to mammals\textsuperscript{70}.

Whether the epigenetic regulation of cryptic TSSs brings any potential developmental and/or adaptive advantages or disadvantages to a plant species is of great interest in plant research. As epigenetic information is relatively flexible and can be reprogrammed according to environmental stimuli, the mechanisms described here may provide plants with a fast and efficient mean for tuning, or even inverting the polarity of regulatory inputs on, gene expression. In addition, potential activation and co-option of cryptic TSSs can provide alternative promoters to the existing transcription units, as observed at AT2G16050 and COQ3 loci (Fig. 6b, c, Supplementary Fig. 14a, b), which may help plants customize gene functions during development\textsuperscript{51,52}. Such events can also create opportunities for plants to innovate their transcriptional response to environmental changes. However, the mis-control of cryptic TSSs encoded in TEs may trigger developmental abnormality in plants\textsuperscript{11,69}. In addition, modulating 3′ and/or 5′ UTRs of a transcript without changing its coding potential can critically affect its function in response to pathogen attacks in \textit{Arabidopsis}\textsuperscript{70}. Epigenetic suppression of a cryptic TSS at the 5′ UTR of the LRR gene AT2G15042 (Supplementary Fig. 14a) may, therefore, help maintain the proper response of \textit{Arabidopsis} to viral infection\textsuperscript{71}. Importantly, activation of the cryptic TSS upstream of SQN (AT2G15790), a gene important for vegetative shoot maturation in \textit{Arabidopsis}\textsuperscript{59}, leads to ectopic production of aberrant transcripts and a decreased accumulation of the normal one (Fig. 6b). Although the impacts of such transcriptional attenuation on plant development are to be confirmed, it has been shown in \textit{A. thaliana} that, light-induced regulation of alternative promoters could generate proteins with differential localizations from the same genes, which help alleviate the impact of changing light conditions on the plant\textsuperscript{72}. Our data, therefore, demonstrate that the epigenetic regulation of cryptic TSSs would profoundly and critically affect proper responses of plant species to ever changing environmental conditions. Additionally, as many protein coding genes in \textit{A. thaliana} possess multiple active upstream as well as intragenic TSSs, it would be interesting to investigate whether cryptic TSSs are still in the process of being co-opted to become functional in the \textit{Arabidopsis} genome.

**Methods**

\textbf{Plant materials.} \textit{ddm1-1}, \textit{met1-3}, \textit{ibm1-4}, \textit{ibm2-2}, and \textit{edm2-9} mutants have been described previously\textsuperscript{16,73–75}. \textit{suvh456} and \textit{nrpe1-7} seeds were kindly provided by Dr. Kakutani and Dr. Kanno, respectively. The T-DNA insertion line of \textit{nrpda1-3} (SALK, 128428) was obtained from the Arabidopsis Biological Resource Center (https://abrc.osu.edu). All the mutants are in Columbia (Col) background. The second generation of homozygous \textit{met1}, \textit{ddm1}, \textit{ibm1}, \textit{ibm2}, and \textit{edm2} were also used for the RNA experiments described below. \textit{nrpda1}, \textit{nrp1}, and \textit{suvh456} were maintained as homozygous for at least three generations before the experiments. The seeds were germinated and grown on 1/2 Murashige and Skoog (MS) plate under long-day conditions (16-h light; 8-h dark) at 22°C.

\textbf{RNA extraction and CAGE.} For CAGE analysis, 10 to 12-day-old whole seedlings of wild-type Col and mutant plants were pooled for RNA extraction. Total RNA was extracted using RNAiso (TAKARA), and DNA was digested with TURBO DNase (Thermo Fisher Scientific), followed by purification by RNasey Plant MiniKit (QIAGEN). Four technical replicates of WT Col and \textit{met1}, and two technical replicates of other samples were prepared for CAGE. Single end 75bp CAGE libraries were prepared and sequenced in DNAFORM (Yokohama, Japan). RNA quality was assessed by Bioanalyzer (Agilent) to ensure that the RIN (RNA integrity number) was over 7.0, and A260/280 and 260/230 ratios were over 1.7.

**ARTICLE**

**NATURE COMMUNICATIONS** | (2020)11:3224 | https://doi.org/10.1038/s41467-020-16951-w | www.nature.com/naturecommunications
CAGE sequencing data analysis. The CAGE sequencing (CAGE-seq) data were processed as follows: sequencing reads were trimmed using Trimmomatic (v0.30) with the following parameters: HEAD:5CROP:1, TRAILING:20, to remove non-specific guanines and low quality bases at the read ends. These were then mapped to the Arabidopsis Col reference genome by HISAT2 (v2.0.0-beta) 77, allowing up to ten alignments for a single read. Due to low mapping coverage, met1A replicates was excluded from further analysis. met1.3 was also discarded due to its low correlation with two other replicates (met1.1 and met1.2). Then, uniquely mapped reads were used to identify TSSs at a single base resolution (CTSSs) by CAGEr (v1.20.0) 78 with the following parameters: sequencingQualityThreshold = 20, mappingQualityThreshold = 20. After being normalized and tags count per million (TPM). CTSS in each sample were grouped into tag clusters by the paracu method, with threshold = 0.1, nrPassThreshold = 2, removeSingletons = TRUE, keepSingletonAbove = 0.3, minStability = 2, maxLength = 100. Finally, tag clusters from individual samples were merged into a common set of consensus tag clusters by the aggregateTagCluster function, with threshold = 0.3, qLow = NULL, qUP = NULL, maxDist = 100, excludeSignalBelowThreshold = TRUE. Each consensus tag cluster was then considered a single reliable TSS, represented by its dominant CTSS, to distinguish from the TSSs annotated by TAIR10. Promoter width was defined by the distance between the 10th (qLow = 0.1) and 90th percentile (qUP = 0.9) quantiles of the cumulative distribution of CAGE signal along each tag cluster, as described in ref. 79. Raw tag counts were used to identify differentially expressed TSSs in the mutants compared to wild-type plants by DESeq2 (v1.0.0), allowing up to two mismatches. DESeq2 mapped to an identical position were collapsed into a single read, and only the best alignment was kept for a read mapped to multiple locations. Mapping results were given in Supplementary Data 9.

ChiP-seq data of PolIV (NRPD1) and the list of NRPD1 binding loci were obtained from ref. 80. Genomic locations of NRPD1 binding loci were then converted from TAIR10 to TAIR9 coordinates using the update_coordinates.pl script provided by TAIR. ChiP-seq data of RNAPII in pol4 and corresponding wild-type plants were obtained from ref. 81. These data were processed as described above. Preprocessed RNAPII SerSP ChiP-seq data (in bigwig format) in pol5 were downloaded from ref. 84 and directly used for visualization.

Bisulfite sequencing data analysis. Whole-genome bisulfite sequencing (WGBS) MethylC-seq data of wild-type plants and epigenetic mutants were retrieved from ref. 89. High quality reads (q ≥ 28), trimmed to remove adapter effects and sequencing bias, were mapped to the Arabidopsis Col reference genome using Bismark (v0.12.1) 80 allowing up to two mismatches. Bases covered by fewer than 3 reads were excluded, and only uniquely mapped reads were used for further analysis. Methylation levels were calculated using MethylKit (v0.5.7) 82. The list of BM, intermediate methylated (IM), and unmethylated (UM) genes were obtained from ref. 44. To exclude the potential impacts of non-CG methylation on the activation of intragenic EPICATs, only met1-activated intragenic EPICATs with less (0%) CG methylation in the 100 bp regions centering around their dominant CTSSs were examined (Supplementary Data 4).

Small RNA sequencing data analysis. Sequencing data of 24 nt small interference RNAs (sRNAs) in wild-type and nrpd1 mutant plants were obtained from ref. 85 and trimmed by TrimmGalore (v0.4.5) 86 with Cutadapt (v1.8.3) 87, using the following parameters: stringency:4, quality:20, length:15, max_length:30. PolIV-dependent small RNAs (P4RNAs) longer than 27 nt in dcl2/3/4 and corresponding wild-type plants were obtained from ref. 80 and trimmed by Trimmomatic. These data were then mapped to the Arabidopsis Col reference genome by Bowtie (v1.0.0), allowing up to two mismatches. Only uniquely mapped reads were used for further analysis.

Sequence motif analysis. De novo motif analysis and search of motif instances were conducted using MEME suite (v4.11.2) with default parameters 89.

Gypsy LTR analysis. Gypsy family sequences were retrieved from the TAIR database and aligned to obtain the full-length sequence for each family.LTR regions were then determined by comparing 5’ and 3’ ends of TE sequences and also checked by LTR_FINDER (v1.0.2) 79. Several copies from each family were used to obtain consensus sequences of LTRs (Supplementary Data 10). Consensus sequences of Gypsy LTRs were used to search for LTR sequences in the Arabidopsis genome (TAIR10) using BLAST (v2.0.5) 80. BLAST hits shorter than 100 bp were discarded. LTR sequences were then aligned using ClustalW (v2.1.8) 81, and edited using Jalview (v2.11.0) 86.

Data visualization. Figures were created using deepTools (v3.3.0) 82, Integrated Genome Browser (IGB) (v9.1.2) 53 with the Araport11 version of genome annotations, Excel, and the R package ggplot2 (v2.3.1) 97. DNA methylation files were firstly converted from bedGraph into bigWig format by the bedGraphToBigWig function (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64), then used to generate heatmap and metatplot figures using deepTools. RNA-seq data were normalized to reads per million (RPM), and a single replicate was used to create IGB track. ChiP-seq signals were normalized to log2(ChIP/input), and a single replicate of RNAPII (both Ser5P and Ser2P) were used for visualization in IGB. Small RNA sequencing data and RNAPII ChiP-seq data with no input samples were normalized to counts per million (CPM).

5’-RACE and quantitative PCR. 5’-RACE was performed by SMARTer RACE kit (TAKARA) according to the manufacturer’s instructions. Quantitative PCR (qPCR) was performed following the method described in ref. 57. All primers used in this study are listed in Supplementary Data 11.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. Sequencing data have been deposited to the DDBJ Sequence Read Archive under the accession codes DRA009314 and DRA009847. Processed CAGE-seq data are also accessible via the following web link: https://plantgenetics.ostp.io/. The source data underlying Figs. 1b, 2c, 3b, 4d, 5b, and 6c, d and Supplementary Figs. 6b–c, 7d, and 14b–c are provided as a Source Data file.
In-house R codes and bash scripts customized for analyzing data are available from the authors upon request.

Received: 9 December 2019; Accepted: 1 June 2020; Published online: 26 June 2020

References

1. Feil, C. D. & Van, H. V. Transposable elements, epigenetics, and genome evolution. Science 338, 758–767 (2012).
2. Lisch, D. How important are transposons for plant evolution? Nat. Rev. Genet. 14, 49 (2013).
3. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18, 71 (2017).
4. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8, 272–285 (2007).
5. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. 11, 204–220 (2010).
6. Zhang, H., Lang, Z. & Zhu, J.-K. Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. 19, 489–506 (2018).
7. Du, J. et al. Dual binding of chromomethylase domains to H3K9ME2-containing nucleosomes directs DNA methylation in plants. Cell 151, 167–180 (2012).
8. Zemach, A. et al. The Arabidopsis nucleosome remodeler DDM1 allows dna methyltransferases to access H1-containing heterochromatin. Cell 153, 193–205 (2013).
9. Stroud, H., Greenberg, M. V., Feng, S., Bernatavichute, Y. V. & Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell 152, 352–364 (2013).
10. Liu, J., He, Y., Amasino, R. & Chen, X. siRNAs targeting an intronic gain-of-function epigenetic allele of a homeodomain gene. Mol. Cell 6, 791–802 (2000).
11. Lisch, D. How important are transposons for plant evolution? Nat. Rev. Genet. 14, 49 (2013).
12. Henderson, I. R. & Jacobsen, S. E. Tandem repeats upstream of the cold-dependent promoter −45 of the Arabidopsis chalcone synthase-related direct repeats. Proc. Natl Acad. Sci. USA 106, 20430–20435 (2009).
13. Hollister, J. D. & Gaut, B. S. Epigenetic silencing of transposable elements: a regulatory role for retrotransposons in pluripotency maintenance. Nat. Rev. Genet. 12, 180 (2011).
14. Le, T. N., Miyazaki, Y., Takuno, S. & Saze, H. Epigenetic regulation of transcription start sites encoded in long terminal repeats. NAR Genet. Exp. 5, D1148–D1153 (2015).
15. Eick, D. & Geyer, M. The ma polymerase II carboxy-terminal domain (CTD) code. Chem. Rev. 113, 8456–8490 (2013).
16. Xiao, J. et al. cis and trans determinants of epigenetic silencing by polycomb repressive complex 2 in Arabidopsis. Nat. Genet. 49, 1546–1552 (2017).
17. Deleris, A. et al. Loss of the DNA methyltransferase MET1 induces H3K9 hypermethylation at PcG target genes and redistribution of H3K27 trimethylation to transposons in Arabidopsis thaliana. PLoS Genet. 8, e1003622 (2012).
18. Takuno, S. & Gaut, B. S. Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. 29, 219–227 (2011).
19. Bewick, A. J. & Schmitz, R. J. Gene body DNA methylation in plants. Curr. Opin. Plant Biol. 36, 103–110 (2017).
20. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. 39, 1570–1575 (2007).
21. Horvath, R., Laenen, B., Takuno, S. & Slote, T. Single-cell expression noise and gene-body methylation in Arabidopsis thaliana. Heredity 123, 81–91 (2019).
22. Bewick, A. J. et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl Acad. Sci. USA 113, 9111–9116 (2016).
23. Rigal, M., Kevei, Z., Pelissier, T. & Mathieu, O. DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. EMBO J. 31, 2981–2993 (2012).
24. Zhai, J. et al. A one precursor one siRNA model for Pol IV-dependent siRNA biogenesis. Cell 163, 445–455 (2015).
25. Faulkner, J. G. et al. The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. 41, 563–571 (2009).
26. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 23, 169–180 (2013).
27. Settles, A. M., Baron, A., Barkan, A. & Martienssen, R. A. Duplication and suppression of chloroplast protein translocation genes in maize. Genetics 157, 349–360 (2001).
28. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24, 1242–1255 (2012).
29. Tsukahara, S. et al. Bursts of retrotransposition occurred in Arabidopsis. Nature 461, 423–426 (2009).
30. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. Biochim. Biophys. Acta Gen. Regul. Mech. 1860, 157–165 (2017).
31. Prunet, N. et al. SQUINT promotes stem cell homeostasis and floral meristem termination in Arabidopsis through APETALA2 and CLAVATA signalling. J. Exp. Bot. 66, 6905–6916 (2015).
32. Fejes-Toth, K. et al. Post-transcriptional processing generates a diversity of 5′–flanking and 3′–flanking end-centered c promoters. Nat. Genet. 44, 461–465 (2012).
33. Settles, A. M., Baron, A., Barkan, A. & Martienssen, R. A. Duplication and suppression of chloroplast protein translocation genes in maize. Genetics 157, 349–360 (2001).
34. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24, 1242–1255 (2012).
35. Tsukahara, S. et al. Bursts of retrotransposition occurred in Arabidopsis. Nature 461, 423–426 (2009).
36. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. Biochim. Biophys. Acta Gen. Regul. Mech. 1860, 157–165 (2017).
37. Prunet, N. et al. SQUINT promotes stem cell homeostasis and floral meristem termination in Arabidopsis through APETALA2 and CLAVATA signalling. J. Exp. Bot. 66, 6905–6916 (2015).

22. Yamamoto, Y. Y. et al. Differentiation of core promoter architecture between plants and mammals revealed by LDDSS analysis. Nucleic Acids Res. 35, 5219–5226 (2007).
23. Mejia-Guerra, M. K. et al. Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. Plant Cell. 27, 3309–3320 (2015).
24. Yamamoto, Y. Y. et al. Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics 8, 67 (2007).
25. Yamamoto, Y. Y. et al. Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. Plant J. 60, 350–362 (2009).
26. Morton, T. et al. Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. Plant Cell. 26, 2746–2740 (2016).
27. Fejes-Toth, K. et al. Post-transcriptional processing generates a diversity of 5′–flanking and 3′–flanking end-centered c promoters. Nat. Genet. 44, 461–465 (2012).
28. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24, 1242–1255 (2012).
29. Tsukahara, S. et al. Bursts of retrotransposition occurred in Arabidopsis. Nature 461, 423–426 (2009).
30. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. Biochim. Biophys. Acta Gen. Regul. Mech. 1860, 157–165 (2017).
31. Prunet, N. et al. SQUINT promotes stem cell homeostasis and floral meristem termination in Arabidopsis through APETALA2 and CLAVATA signalling. J. Exp. Bot. 66, 6905–6916 (2015).
58. Avelange-Macherel, M.-H. & Joyard, J. Cloning and functional expression of ATCOQ3, the Arabidopsis homologue of the yeast COQ3 gene, encoding a methyltransferase in plant mitochondria involved in ubiquinone biosynthesis. Plant J. 14, 203–213 (1998).

59. Meinke, D. W. Genome-wide identification of EMBRYO-DESTRUCTIVE (EMB) genes required for growth and development in Arabidopsis. New Phytol. 14, 306–325 (2019).

60. Santamaria, M., Thomson, C. J., Read, N. D. & Loake, G. J. The promoter of a basic PR1-like gene, ATPR1B, from Arabidopsis establishes an organ-specific expression pattern and responsiveness to ethylene and methyl jasmonate. Plant Mol. Biol. 47, 641–652 (2001).

61. Todd, C. D., Deniz, Ö., Taylor, D. & Branco, M. R. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. Elife 8, e44344 (2019).

62. Teissandier, A. & Bourc...