communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-09454-1

Nanopore sequencing and multiomics reveal predictable non-coding RNA activation in DNA methylation deficient *Arabidopsis thaliana*

Wanghong Shi[1], Luyao Wang[2], Na Zhou[3], Mengke Zhang[1], Yupeng Hao[1], Ke Nie[1], Xueying Guan[1] & Ting Zhao[1]

Long intergenic non-coding RNAs (lincRNAs) are regulatory transcripts from intergenic regions with diverse expression patterns, but whether the activation of lincRNA is widespread and predictable remains unclear. Here, we applied Oxford Nanopore Technology Direct RNA and DNA sequencing (ONT DRS and DDS) to *Arabidopsis* DNA methylation-deficient mutants (*ddm1* and *met1*) and wild type. Differential expression analysis identified 340 upregulated lincRNAs and 209 lincRNAs with consistent expression whose expression was negatively correlated with DNA methylation. Similar activation patterns were also detected in natural populations. To further characterize these lincRNAs, fifty multi-omics features were compiled to train six machine learning models for classifying *ddm1*-activated lincRNAs and Random Forest achieved the highest average precision of 0.96. Feature importance analysis highlighted population-level DNA methylation, ONT-derived RNA modification and transposable elements as key predictors. These results indicate that epigenetic variation shapes predictable lincRNA activation, establishing a framework for systematic discovery of expressible non-coding RNAs.

Noncoding regions of genomes are reported to have high transcriptional potential and therefore play important roles in genomic evolution and regulation[1,2]. Long intergenic noncoding RNAs (lincRNAs), which are transcribed from intergenic regions and encode no recognizable proteins, are widespread in eukaryotic transcriptomes and exert a wide spectrum of influences within an organism[3–5]. Indeed, mounting evidence suggests that in plants, lincRNAs play regulatory roles in developmental and physiological processes, including vernalization[6], flowering development[7], stress response[8], and crop domestication[9]. LincRNAs as an unclassified pool are far more diverse than protein-coding genes (PCGs) in terms of expression activity, increased tissue specificity, and variability in expression across individuals[10,11]. Transcription of lincRNAs can result in novel genes and provide a route for expanding genome functionality without changing the number of genes[2,12]. From an evolutionary perspective, the flexibility of lincRNAs is important for generating genomic and phenotypic plasticity, such as constructive neutral evolution.

A relationship between lincRNAs and DNA methylation has been widely reported in animals and plants at both individual and population levels[13]. For example, population studies in cancer cells[14] and have been conducted in the model plant *Arabidopsis*[15] reported that lincRNA activation is associated with hypomethylation. In plants, reprogramming of DNA methylation in the course of hybridization and polyploidization is reported to lead to the accumulation of noncoding RNAs[16–18]. Studies of DNA methylation-related mutants have also identified a large number of activated lincRNAs across multiple plant and animal species, including *Oryza sativa*, *Arabidopsis thaliana*, *Zea mays*, *Brassica rapa*, and *Mus musculus*[19,20]. Taken together, these studies indicate that genomes have a capability to integrate epigenetic modifications and transcription machinery on noncoding regions to create novel transcripts as a candidate pool for adaptive evolution within a short time. It is thus of great interest to identify and validate at a genomic scale the flexibility of those noncoding regions.

[1]Zhejiang Provincial Key Laboratory of Crop Genetic Resources, Institute of Crop Science, Plant Precision Breeding Academy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China. [2]Hainan Institute of Zhejiang University, Sanya, Hainan, China. [3]Dry Land Farming Institute, Hebei Academy of Agricultural and Forestry Sciences, Hengshui, China. e-mail: tingzhao@zju.edu.cn

It has been widely proposed that epigenetic variations can affect gene expression and cause visible phenotypic differences, allowing rapid adaptation to dynamic environmental changes[11,21–23]. Epigenetic recombinant inbred lines (epiRILs) are created by crossing a mutant line defective in DNA methylation (for example, *Arabidopsis ddm1* mutant) with a wild-type line[24]. The progeny were subsequently selfed for multiple generations, during which the methylation-deficiency mutation was removed through segregation. As a result, the final lines share the same genetic sequence as the wild type but retain heritable differences in DNA methylation derived from the original mutant parent[25]. Epigenome-wide association studies in natural populations, along with studies using epiRILs, have identified a large number of functional epigenetic variations[11,22,26]. More importantly, populational genetic analysis and comparative genome analysis have shown that a proportion of lincRNAs are under positive selection pressure[9,15,27]. That lincRNAs serve as a source of variation underlines the high potential of epigenome exploitation for breeding purposes.

In *Arabidopsis*, DNA methylation mutants exhibiting different levels of DNA methylation deficiency have been well characterized[28]. Nucleosome remodelers of the *DNA methylation 1* (*DDM1*)/Lsh family, which encode a protein similar to the chromatin-remodeling factor SWI2/SNF2, are required for maintenance of CG methylation (mCG) that is conserved in plants and animals[29]. Meanwhile, the protein encoded by *methyltransferase 1* (*MET1*) recognizes hemimethylated CG following DNA replication and methylates the naked cytosine in the daughter[30]. Subsequent maintenance of DNA methylation in heterochromatin requires *DDM1*[31]. Accordingly, *ddm1* and *met1* mutations have been shown to result in drastically decreased DNA methylation spreading on transposon elements (TEs)[32–34]. In addition, the hypomethylated status characteristic of *ddm1* can be inherited even after eight generations[35]. These features make *Arabidopsis ddm1* and *met1* mutants excellent models for the investigation of lincRNA behaviors under a precondition of DNA hypomethylation[36].

About 80% of lincRNAs have 5′ caps and 3′ poly(A) tails, making them amenable for discovery by oligo (dT)-based RNA sequencing[37]. However, lincRNA sequences are also enriched with transposable element (TE) fragments[38,39], which are highly repetitive sequences that cannot be accurately assembled by Illumina Next-Generation Sequencing (NGS) technology on which current studies rely; such technologies lack the necessary long-range information[40,41]. Direct RNA sequencing (DRS) offered by Oxford Nanopore Technologies (ONT) can sequence polyadenylated RNAs directly, without the recoding and amplification biases inherent to other sequencing methodologies[42]. Moreover, DRS can produce long reads (up to 30 kb, vs. up to 300 bp at present for Illumina RNA-seq), which can provide for accurate transcript assemblies[43]. In addition, this method can detect the RNA modifications of m^6A and m^5C simultaneously, which introduce further epi-modification features in situ at single-molecule resolution[42,44]. Therefore, DRS is an advanced technology suitable for the detailed study of lincRNAs.

To investigate the underrepresented complexity of noncoding elements in the transcriptome, we leveraged the power of DRS to characterize lincRNAs in *Arabidopsis* lines having null mutations of the DNA methylation regulators *ddm1* and *met1*. This approach was supplemented with short reads produced by Illumina RNA-seq. We identified a large set of *ddm1*-activated lincRNAs and defined their ONT-derived molecular features while examining their occurrence in natural populations. Fifty multi-omics features were further extracted, including sequence properties, ONT-derived features and population-level methylation and variation, for each lincRNA and applied six machine learning algorithms to classify *ddm1*-activated versus non-activated lincRNAs. Feature importance analysis revealed key predictive features underlying lincRNA activation and provided further insights into the interplay between epigenome dynamics and noncoding transcript regulation.

Results

Study design

DNA methylation shows substantial variation across natural populations and can trigger the activation of novel lincRNAs[15]. To examine the characteristics of these lincRNAs, including expression variation, RNA modifications, and alternative poly(A) tails, ONT DRS and direct DNA sequencing (DDS) were performed on leaf samples of *Arabidopsis thaliana* Col-0 wild type (WT) and the *ddm1* and *met1* mutants (Fig.1A). For comparison, Illumina RNA-seq (150 bp short reads) was conducted in parallel on the same samples.

Full-length lincRNAs were assembled from ONT long-read RNA-seq data, enabling the identification of lincRNAs activated under DNA methylation-deficient conditions. To further characterize these activated lincRNAs, ONT-specific features derived from the current signals were integrated with publicly available repetitive sequence annotations, population genetic metrics, and epigenetic profiles. Machine learning models were then trained by using these multi-omics features, and interpretability analyses were carried out to elucidate the molecular mechanisms underlying the activation of novel lincRNAs (Fig. 1B).

Assessment of DRS profiles and lincRNA annotation

DRS generated a total of 40 million long reads (LRs) across six samples. The median read length of passed reads was 893 bp, with a maximum length of 14,248 bp. The read length distribution and quality metrics were comparable to those reported in a previous *Arabidopsis thaliana* study[45]. However, each library yielded an average of over 6 million reads, representing a 7.4-fold increase in total base coverage compared with the ~0.8 million reads per library reported previously[45].

Quality-controlled LRs were mapped to the *Arabidopsis thaliana* TAIR10 reference genome and transcriptome, achieving average mapping rates of 99.70% and 96.85%, respectively (Supplementary Fig. 2 and Supplementary Data 2). Although long-read sequencing detected slightly fewer genes than short read (SR) sequencing (~15,471 genes), it still identified no fewer than 15,000 genes per sample (Fig.2A and Supplementary Data 3). Saturation analysis indicated that gene detection plateaued at approximately 4million reads (Fig.2B and Supplementary Data 4). Principal component analysis showed that replicates clustered within each genotype (Supplementary Fig. 2E), confirming the high reproducibility of the DRS data. Gene expression values from LR and SR sequencing were strongly correlated, with Spearman's correlation coefficients of 0.86 (Fig.2D and Supplementary Fig. 3), demonstrating that LRs can provide accurate quantitative transcriptome measurements.

As for noncoding transcripts, fewer than 1% of LRs in WT aligned to noncoding regions. In contrast, noncoding read proportions increased six-fold in *ddm1* and two-fold in *met1* (Fig.2E and Supplementary Data 2), consistent with previous observations of noncoding RNA activation in DNA methylation mutants[20]. These results indicate that LR libraries can detect biologically relevant transcripts from noncoding regions.

Using DRS data, we annotated 1663 lincRNA transcripts derived from 549 transcriptional models. Of these, 538 (98.0%) were also supported by SR sequencing (Fig.2G). Comparative analysis with the NONCODE v6[46] database revealed that 248 lincRNAs (45.2%) had been previously reported in the *Arabidopsis thaliana* noncoding RNA repertoire, while the remaining 301 (54.8%) represent novel lincRNA candidates.

Improving mapping accuracy in lincRNA annotation using DRS

LincRNAs often contain repetitive sequences[38,39], which can complicate read mapping. LR sequencing produces RNA reads with an average N50 of ~1055 bp (Supplementary Data 1), thereby reducing the likelihood of incorrect mapping positions[47]. In BAM format, the mapping quality value MQ0 indicates reads with multiple mapping locations or reads for which the aligner cannot confidently assign a unique position. We compared MQ0 values between LR and SR datasets in lincRNA regions. The proportion of MQ0 reads ranged from 3.14 to 4.80% in LR data, compared to 8.42 to 9.96%

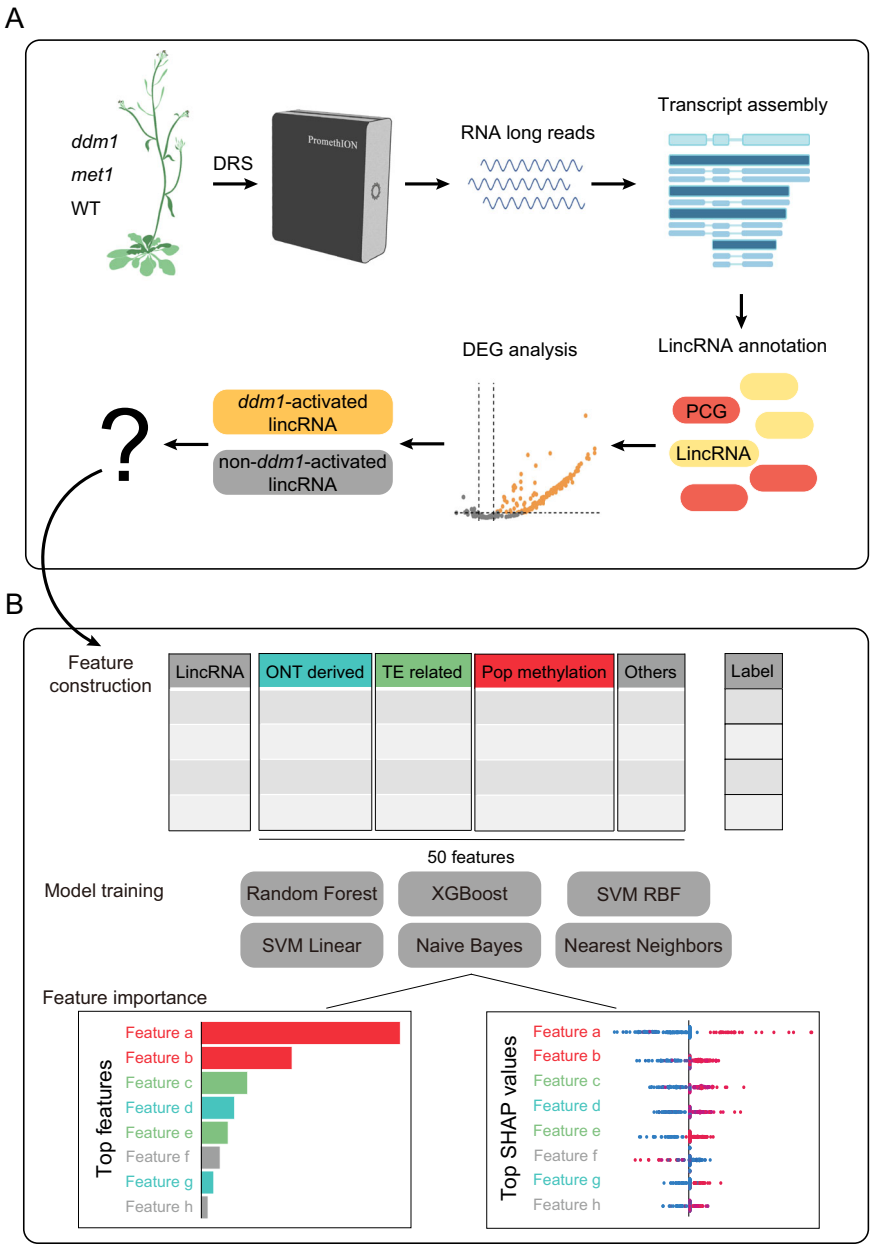

Fig. 1 | Workflow of this study. A Schematic graph showing the study workflow. Leaves of *Arabidopsis thaliana* Col-0 wild type (WT), *ddm1* and *met1* mutants were harvested for Oxford Nanopore Technology Direct RNA and DNA sequencing (ONT DRS and DDS). RNA reads were assembled for annotation. Transcriptional activation of lincRNAs was identified. **B** Feature-based machine learning model training and interpretability analysis.

in SR data, indicating that SR exhibited approximately 2.5-fold higher multi-mapping rates than LR (Supplementary Fig. 4).

LR sequencing has been reported to exhibit higher base-level error rates than SR sequencing[48]. To assess this in our data, we evaluated LR error rates across biological replicates and sequencing platforms (ONT DRS versus Illumina NGS) by comparing detected single-nucleotide polymorphisms (SNPs) and insertions/deletions (InDels). Across all replicates of WT, *ddm1*, and *met1*, ONT DRS yielded substantially higher variant frequencies than NGS. SNP variant frequencies in LR were 6.1–12.8 fold higher than those in SR (Supplementary Fig. 5), with an even stronger trend for InDels (Supplementary Fig. 5), consistent with previous reports[49] that DRS is particularly susceptible to InDel errors. Notably, ONT DRS produced highly consistent high-frequency variant sites across biological replicates within the same platform (Supplementary Fig. 5).

In LR data, the proportion of variants located within lincRNA bodies ranged from 0.94 to 6.17% for SNPs and from 0.99 to 3.99% for InDels across genotypes. Specifically, in the WT genotype, only 0.94% of SNPs (55/5866) and 0.99% of InDels (65/6570) were found within lincRNA bodies (Supplementary Fig. 5A and Supplementary Fig. 5B). These results indicate that base error has a limited impact on mapping accuracy in lincRNA analyses.

Active expression of lincRNAs in ddm1 and met1

Due to differences in sequencing platforms, lincRNA quantification was performed using counts per million (CPM) for LR libraries and fragments per kilobase of transcript per million mapped reads (FPKM) for SR libraries. In LR libraries, lincRNA expression was markedly elevated in *ddm1* (median CPM = 6.77) and *met1* (median CPM = 1.49) mutants compared with WT samples (median CPM = 0) (*ddm1* vs *met1*, $P < 0.01$; *ddm1* vs WT, $P < 0.01$; *met1* vs. WT, $P < 0.05$, Student's *t* tests). By contrast, the loss of DNA methylation in the *ddm1* and *met1* mutants exerted little impact on the transcription of PCGs (median CPM: *ddm1* = 11.1, *met1* = 12.5, WT = 13.7; *ddm1* vs *met1*, $P = 0.37$; *ddm1* vs. WT, $P = 0.26$; *met1* vs. WT, $P = 0.86$, Student's *t* tests) (Fig. 3A, C). Consistent patterns were observed in SR libraries, where lincRNA expression levels were significantly higher in *ddm1* (median FPKM = 2.31) and *met1* (median FPKM = 0.48) than in WT

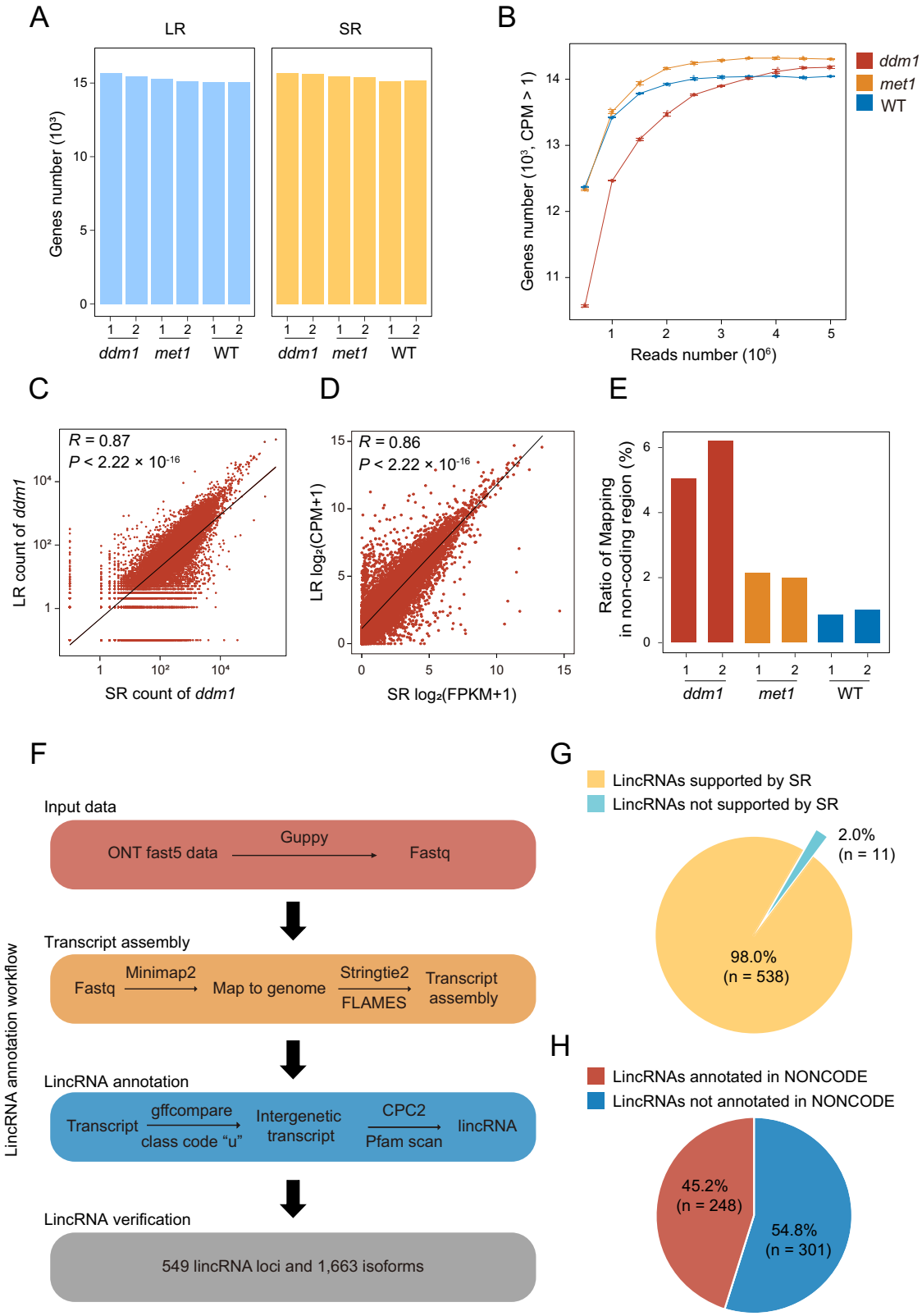

Fig. 2 | Sequencing of *Arabidopsis ddm1*, *met1*, and WT genotypes using DRS technology. A Number of genes detected by long reads (LRs) of DRS and short reads (SRs) of Illumina. **B** Sequencing saturation analysis of DRS mapping in each genotype. Y-axis represents the number of genes (Counts per million, CPM > 1) covered at a given number of reads (x-axis). Error bars represent mean ± standard deviation (SD). **C**, **D** Correlation of gene quantifications based on LRs and SRs. Each dot represents a gene. **E** Bar chart showing the proportion of LRs in the noncoding regions of the sequenced samples. **F** Workflow for lincRNA annotation based on DRS. **G** Pie chart illustrating lincRNA annotations based on DRS LR that were supported by 5 SR reads. **H** Pie chart illustrating the number and proportion of novel lincRNAs versus those previously annotated in the NONCODE database, with identity >90% and $P < 4.24 \times 10^{-11}$.

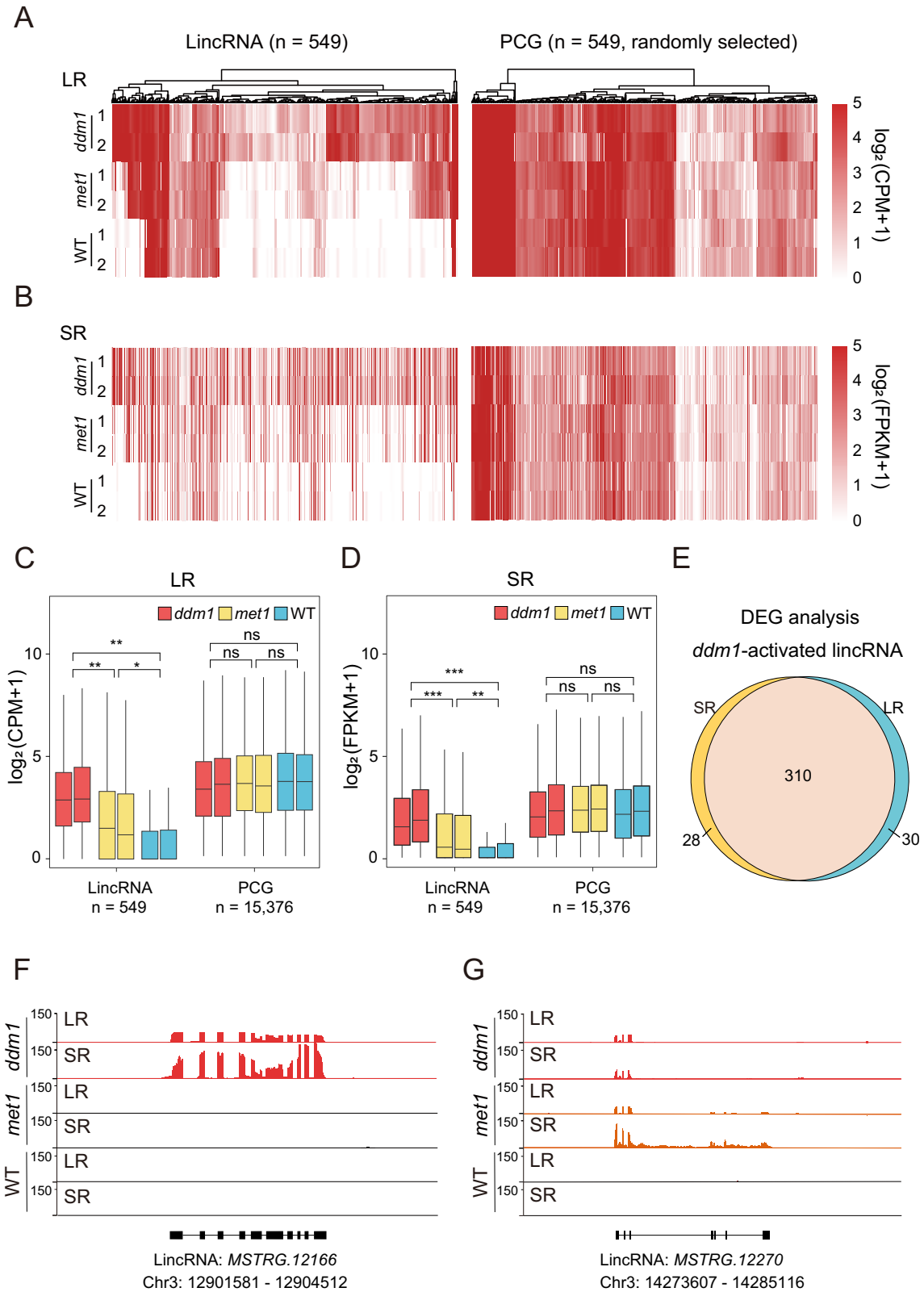

Fig. 3 | Both LR and SR sequencing reveal substantial lincRNA activation in methylation-deficient mutants. A Heatmap of 549 lincRNAs and randomly selected 549 protein coding genes (PCGs) expression profiles obtained from LRs, quantified by $\log_2(\text{CPM} + 1)$. **B** Heatmap of 549 lincRNAs and randomly selected 549 protein coding genes (PCGs) expression profiles obtained from SRs, quantified by $\log_2(\text{FPKM} + 1)$. **C** Box plot comparing lincRNA ($n = 549$) and PCG ($n = 15{,}376$) expression levels in the different datasets, quantified by $\log_2(\text{CPM} + 1)$. **D** Box plot comparing lincRNA and PCG expression levels in the different datasets, quantified by $\log_2(\text{FPKM} + 1)$. **E** Venn diagram showing the coincidence of differentially expressed lincRNAs as determined using LR and SR data. **F** IGV visualization of a *ddm1*-activated lincRNA locus, *MSTRG.12166*, using LR and SR libraries. **G** IGV visualization of a lincRNA locus activated in both *ddm1* and *met1*, using LR and SR libraries.

(median FPKM = 0) (*ddm1* vs. *met1*, $P < 0.001$; *ddm1* vs. WT, $P < 0.001$; *met1* vs. WT, $P < 0.01$, Student's *t* tests) whereas PCGs expression showed no substantial differences among samples (median FPKM: *ddm1* = 4.39, *met1* = 4.35, WT = 3.86; *ddm1* vs. *met1*, $P = 0.10$; *ddm1* vs. WT, $P = 0.52$; *met1* vs. WT, $P = 0.29$, Student's *t* tests) (Fig. 3B, D).

The trends in active lincRNAs observed in the hypomethylated mutants were further validated by differentially expressed gene (DEG) analysis. Among all annotated lincRNAs, 61.93% ($n = 340$) in *ddm1* and 22.22% ($n = 122$) in *met1* were differentially upregulated (adjusted $P < 0.05$) (Supplementary Fig. 6A and Supplementary Fig. 6D). In contrast, only 4.08% ($n = 627$, total PCG = 15,377) and 2.77% ($n = 426$) of PCGs were upregulated in *ddm1* and *met1*, respectively (Supplementary Fig. 6C and Supplementary Fig. 6F). The same DEG analysis applied to the SR libraries revealed that most lincRNA DEGs were shared between LR and SR libraries (Fig. 3E). Specifically, among the up-regulated lincRNAs detected in LR libraries, 310 (91.18%, out of 340) and 116 (95.08%, out of 122) were also differentially expressed in SR libraries in *ddm1* and *met1* (Supplementary Fig. 6G).

A representative lincRNA, *MSTRG.12166*, was visualized in both LR and SR libraries using the Integrative Genomics Viewer (IGV) (Fig. 3F). This transcript was specifically expressed in *ddm1* but was not expressed in the corresponding WT or *met1* samples. Figure 3G illustrates a second representative lincRNA, *MSTRG.12270*, that was actively expressed in *ddm1* and *met1* but not expressed in WT.

All told, the DRS profiling confirmed our previous conclusions regarding the impact of DNA hypomethylation on activation of lincRNAs in multiple organisms, which might alter transcriptome plasticity[19]. Thus, the lincRNAs were classified according to expression specificity, yielding 340 *ddm1*-activated lincRNAs and 209 non-*ddm1*-activated lincRNAs for further analysis (Supplementary Data 5).

Patterns of RNA modifications and poly(A) tail lengths in lincRNAs

ONT sequencing generates signals derived from electrical currents, which allow detection of RNA modifications (m^6A, m^5C) and measurement of poly(A) tail length[50]. Using DRS data, we identified a total of 53,359 ± 7088 sites with m^6A modifications and 1,954,238 ± 370,741 sites with m^5C modifications, as well as 1,881,542 ± 429,669 poly(A) tail length signals.

We first compared the signal detection rate in proportion of lincRNAs and PCGs. In the *ddm1* mutant, the ratio of *ddm1*-activated lincRNA with m^6A modifications (average = 65.7%) was similar to PCGs (average = 67.0%), with comparable m^5C modification frequencies (average of PCGs = 66.7%, average of *ddm1*-activated lincRNAs = 63.0%). In contrast, non-*ddm1*-activated lincRNAs exhibited lower detection ratios (m^6A = 34.0%, m^5C = 41.9%) (Fig.4A).

Poly(A) tails were detected in the majority of *ddm1*-activated lincRNAs (98.5%, $n = 335$). Among PCGs, 13,565 (88.2%) were consistently detected with poly(A) tails in both *ddm1* and WT replicates, and 15,320 (99.6%) had poly(A) tails in at least one replicate. Detection rates were lower for non-*ddm1*-activated lincRNAs (84.7% on average, $n = 147$) (Fig.4A).

Among lincRNAs and PCGs with detected signals, no significant differences in m^6A or m^5C modification levels were observed ($P > 0.05$, Wilcoxon rank-sum test). In contrast, poly(A) tail lengths differed markedly: *ddm1*-activated lincRNAs exhibited the longest tails (median = 106.1 nt), followed by non-*ddm1*-activated lincRNAs (median = 83.7 nt), whereas PCGs exhibited the shortest tail lengths (median = 72.8 nt). Pairwise comparisons between all groups were statistically significant ($P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test). Similar patterns in RNA modification detection rates and poly(A) tail length distributions were observed in the *met1* mutant (Supplementary Fig. 7). Moreover, poly(A) tail length was negatively correlated with transcriptional activity (Supplementary Fig. 8, Kruskal–Wallis rank sum test).

DNA methylation according to DDS is negatively correlated with transcriptional activity at lincRNA loci

To investigate the quantitative relationships between DNA methylation and lincRNA transcriptional activity, DDS technology was employed to quantify DNA methylation levels without chemical treatment as well as structural variations (SVs). DDS yielded 3.1 million reads with an average length of 14 kb (Supplementary Data 1). The longest read obtained was 209 kb, and the mean coverage of each sample was 61×. Mapping the reads to the *Arabidopsis thaliana* TAIR10 reference genome revealed 189 SVs in *ddm1* and 118 in *met1* mutants compared with WT (Supplementary Fig. 9). SVs were observed in only small fractions of *ddm1*-activated lincRNAs (3.82%, $n = 13$) and non-*ddm1*-activated lincRNAs (2.87%, $n = 6$). Accordingly, no statistical enrichment of structural variation was detected in *ddm1*-activated lincRNAs ($P = 0.36$, Fisher's Exact Test) (Supplementary Data 6).

To examine the association of lincRNA activity with DNA methylation level, genome-wide DNA methylation was first confirmed to be significantly reduced in *ddm1* and *met1* relative to WT ($P < 2.22 \times 10^{-16}$, Student's *t* test) (Fig. 5A), based on methylated cytosine (mC) frequencies in reads called by Guppy (v5.0.11)[51]. In addition, BS-seq data for *ddm1* and WT from a previous study[52] were compared with our DDS data, derived genome-wide methylation profiles. A sliding-window analysis was used to compute similarity across the genome, revealing a strong correlation (PCC = 0.94, $P < 2.22 \times 10^{-16}$, Student's *t* test) (Supplementary Fig. 10).

Comparison of gene body DNA methylation in *ddm1*-activated lincRNAs between *ddm1* (mean = 11.98%) and WT (mean = 79.17%) revealed a 67.20% decrease in the mutant (Fig. 5B). Meanwhile, non-*ddm1*-activated lincRNAs and PCGs exhibited respective decreases of only 16.97% (Fig. 5C) and 5.58% (Fig. 5D). These results confirm that lincRNA activation in *ddm1* is significantly associated with alterations in DNA methylation. This is further illustrated by representative *ddm1*-activated lincRNAs such as *MSTRG.2956*, *MSTRG.6558*, *MSTRG.12164*, and *MSTRG.12166*, at which loci the genomic sequences in WT and *ddm1* were identical, but RNA activation was specific to the *ddm1* mutant and corresponded with reduced DNA methylation (Fig. 5E). Hence, CG methylation is most likely a major obstacle to the expression of lincRNAs.

ddm1-activated lincRNAs can be detected in Arabidopsis natural populations

The above findings prompted us to investigate the expression and DNA methylation of *ddm1*-activated lincRNAs in natural populations of *Arabidopsis*, to validate the inferred lincRNA activity patterns on a larger scale. To expand the biological scope of our analysis, we retrieved leaf transcriptome for 728 *Arabidopsis* natural accessions from the 1001 Genomes Project[21].

We assessed the expression of both *ddm1*-activated and non-*ddm1*-activated lincRNAs, considering Reads Per Kilobase per Million mapped reads (RPKM) ≥ 1 as expressed. Notably, 68.31% ($n = 375$) of annotated lincRNAs were expressed in at least one accession (Fig. 6A and Supplementary Data 7). Furthermore, lincRNAs previously activated under hypomethylated conditions in the *ddm1* mutant were also expressed in natural populations, although at low frequencies (Fig. 6B and Supplementary Data 7).

Expression variability of *ddm1*-activated lincRNAs was significantly greater than that of non-*ddm1*-activated lincRNAs or PCGs, based on average coefficient of variation (CV, $P < 2.22 \times 10^{-16}$, Wilcoxon test) (Fig. 6C). For instance, the multi-exon *ddm1*-activated lincRNA *MSTRG.19777* was undetectable (RPKM < 1) in most accessions, yet showed high and consistent expression in three accessions (IP-San-10, Gd-1, and Hue-3) with a stable RNA splicing pattern (Fig. 6D). In these accessions, the corresponding genomic region exhibited reduced DNA methylation across CG, CHG, and CHH contexts (Fig. 6E).

Collectively, these results indicate that in natural *Arabidopsis* populations, certain accessions display lincRNA activation comparable to that observed in the *ddm1* mutant, likely as a consequence of reduced genome-wide DNA methylation.

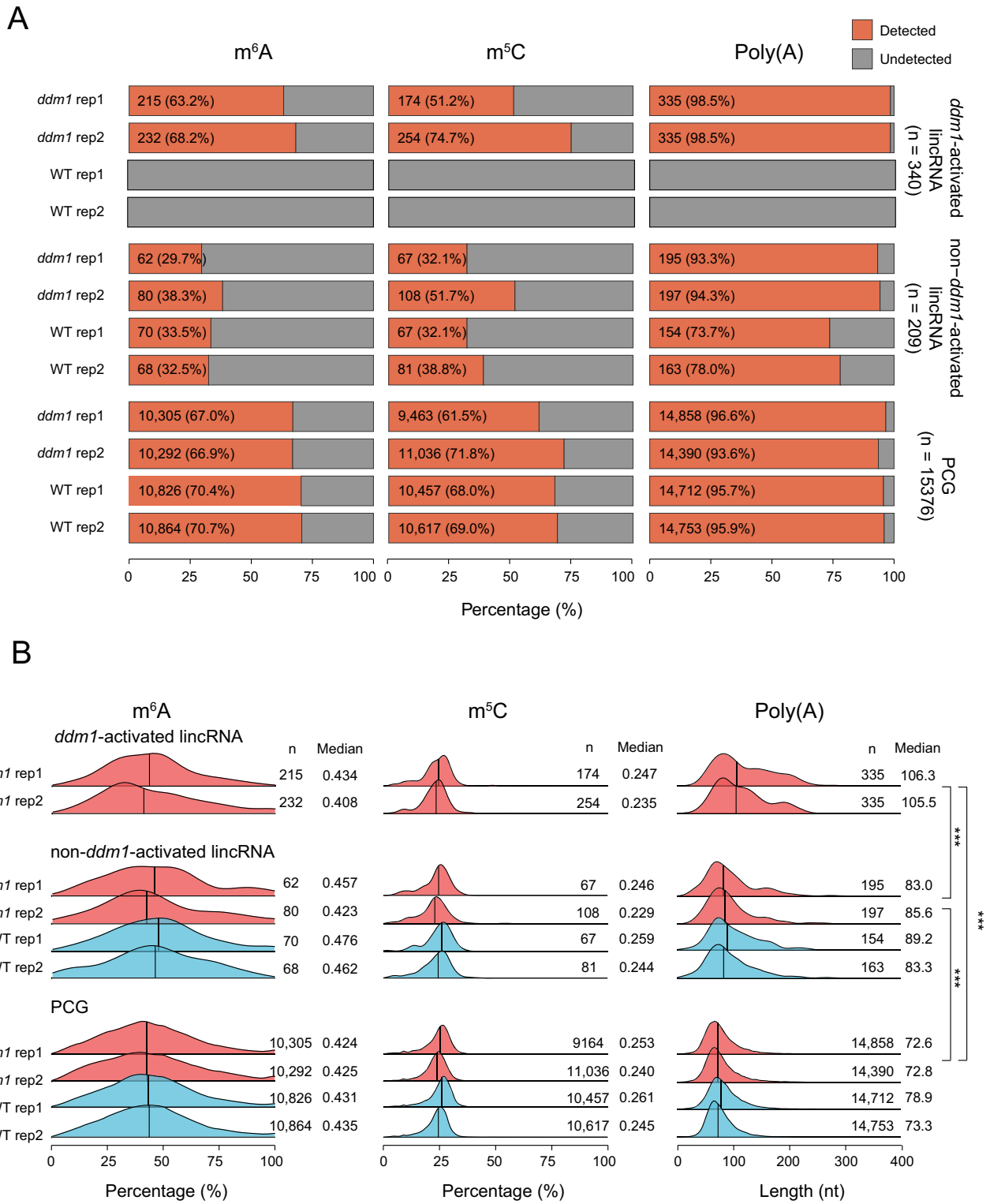

Fig. 4 | RNA epigenetic modifications and poly(A) tail lengths derived from ONT sequencing. A Histogram showing the ratio of lincRNAs and PCGs detected with m^6A, m^5C and poly(A) signals in different samples. **B** Distribution of m^6A, m^5C and poly(A) length based on ONT sequencing among *ddm1*-activated lincRNAs, non-*ddm1*-activated lincRNAs, and PCGs. Black vertical line indicates the median level.

Prediction of ddm1-activated lincRNAs by machine learning

To further characterize these activated lincRNAs, we hypothesized that lincRNAs could be classified into two categories based on their activation status under DNA methylation deficiency. In total, 50 multi-omics features were compiled for each lincRNA (Supplementary Data 8 and 9), including population-level DNA methylation features, ONT-derived features, TE-related features. The remaining features were categorized as "other features". Model construction was performed separately using six different algorithms.

Random Forest achieved the highest predictive performance (ACC = 0.82, AP = 0.96, AUC-ROC = 0.93), followed closely by XGBoost (ACC =

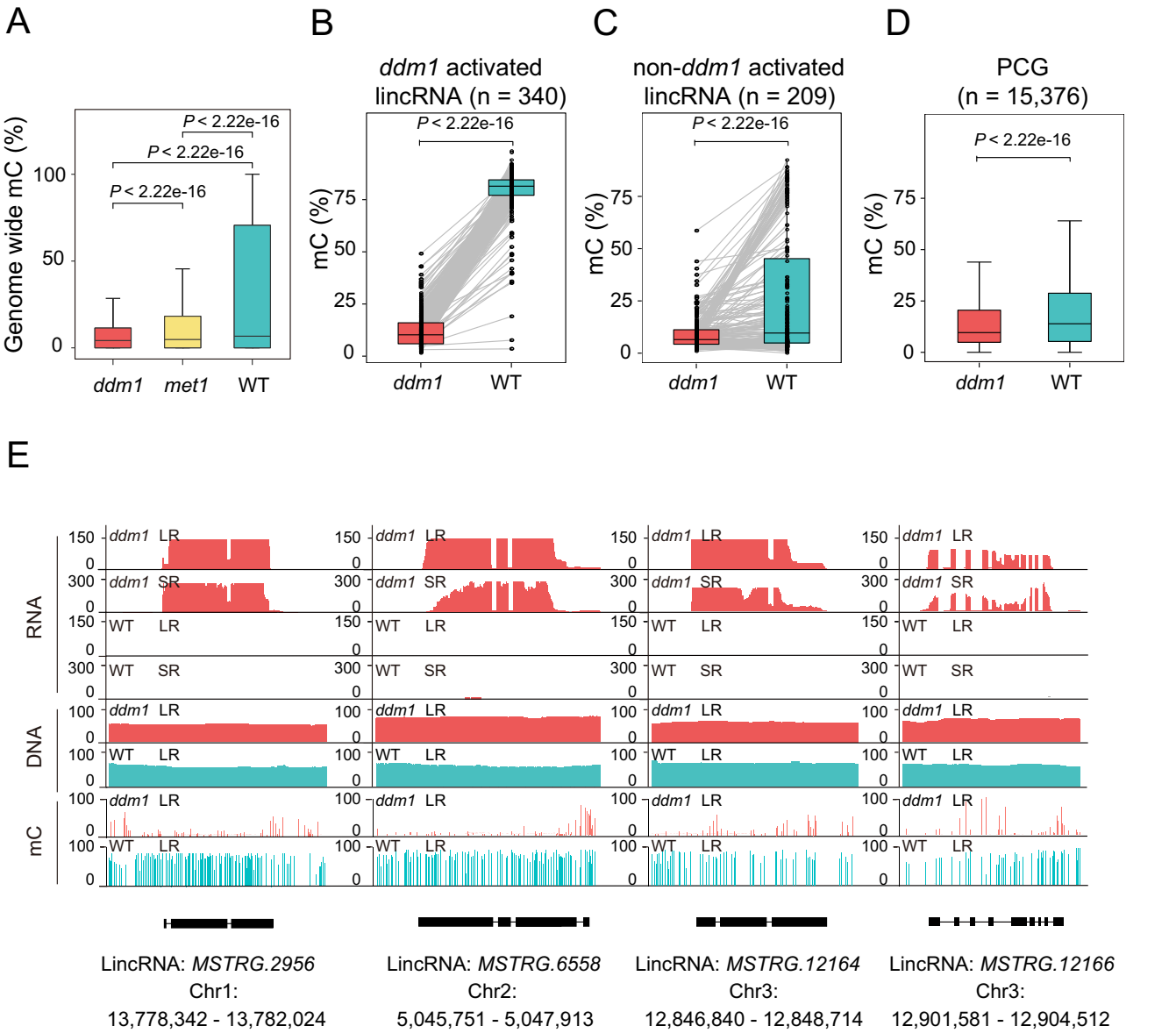

Fig. 5 | Decreased DNA methylation activates novel lincRNA transcription. **A** Boxplot showing DNA methylation levels at CpG sites based on ONT DDS. **B** Boxplot showing DNA methylation average levels on *ddm1*-activated lincRNA bodies ($n = 549$) in *ddm1* mutants and wild-type as indicated at the bottom. **C** Boxplot showing DNA methylation average levels on non-*ddm1*-activated lincRNA bodies ($n = 549$) in *ddm1* mutants and wild-type. **D** Boxplot showing DNA methylation average levels of PCG gene bodies in *ddm1* mutants and wild-type. **E** IGV visualization of the *ddm1*-activated lincRNA *MSTRG.2956*, *MSTRG.6558*, *MSTRG.12164* and *MSTRG.12166*. From top to bottom: DRS data showing the transcription coverage of lincRNAs in *ddm1* mutant; NGS data showing the transcription coverage of lincRNAs in *ddm1* mutant; DRS data showing the transcription coverage of lincRNAs in WT; NGS data showing the transcription coverage of lincRNAs in WT; DDS data showing the DNA reads coverage; DDS data illustrating the decrease of DNA methylation in *ddm1* mutant compared with WT.

0.81, AP = 0.94, AUC-ROC = 0.91). Even the least effective classifier, Nearest Neighbors, attained an accuracy of 0.75 (AP = 0.91, AUC-ROC = 0.88) (Fig. 7A–C and Supplementary Data 10), demonstrating that diverse machine-learning methods can reliably distinguish *ddm1*-activated lincRNAs from non-activated counterparts.

To assess the contribution of different feature types, we performed classification under multiple feature-combination settings. The model integrating all four feature categories (50 features in total) achieved the highest predictive performance (AP = 0.96, AUC-ROC = 0.93) (Fig. 7B, C). Using genomic features alone as the baseline, the random forest model reached an AP of 0.75 and an AUC-ROC of 0.69 (Fig. S13). Adding population-level methylation features increased AP to 0.92 (+22.7%) and AUC-ROC to 0.90 (+30.4%). Incorporating ONT-derived features improved AP by 9.3% and AUC-ROC by 5.8%, while inclusion of TE features raised AP by 10.7% and AUC-ROC by 11.6% (Supplementary Fig. 13A).

Interpretability of classification models reveals biological drivers of lincRNA activation

Integrating feature importance across six models revealed that both the average level and CV of population-level DNA methylation were strongly associated with activation status (Supplementary Fig. 12), contributing 19.3%, 25.2%, and 14.2% of total feature importance in the random forest, XGBoost, and naive Bayes models, respectively (Fig. 7D–F). Overlap between TEs and lincRNAs also showed a notable link to activation (Supplementary Fig. 12). Specifically, the support vector machine (SVM) (radial basis function (RBF)), SVM (Linear), and naive Bayes models consistently highlighted regions containing Class II TIR/CACTA, Gypsy, and Mutator elements as highly important (Fig. 7F and Supplementary Fig. 11A). ONT-derived RNA modification features, such as m^5C and m^6A, likewise ranked among the top predictors in multiple models (Supplementary Fig. 12).

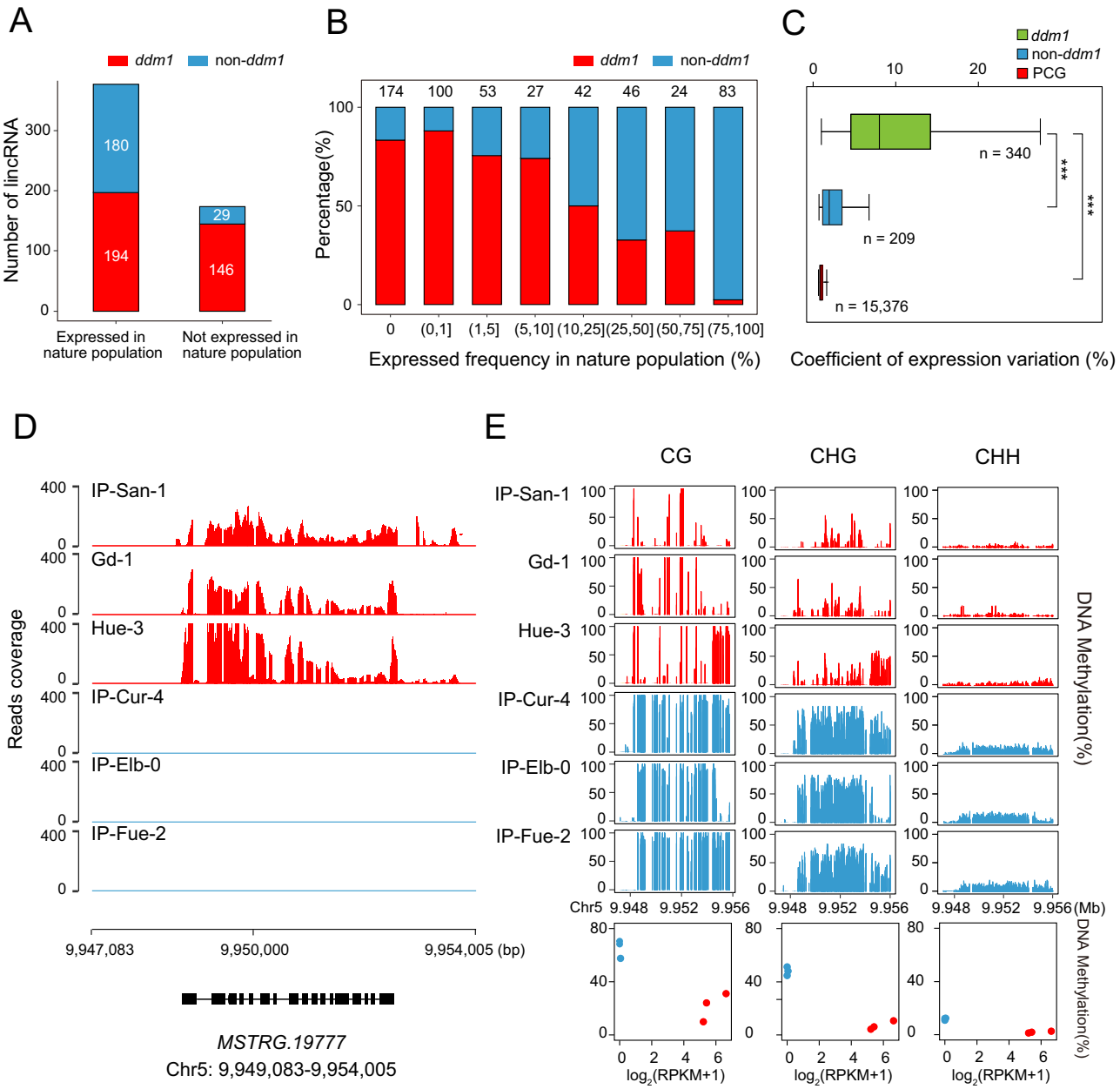

Fig. 6 | *ddm1*-activated lincRNAs are widespread in natural populations at low frequency. **A** Bar plot showing the number of *ddm1*-activated and non-*ddm1*-activated lincRNAs expressed in natural populations. **B** Frequency of expression (FPKM > 1) for *ddm1*-activated and non-*ddm1*-activated lincRNAs in natural populations. The *x*-axis represents frequency. **C** Distribution of the coefficient of variation (CV, ratio of standard deviation to mean) among *ddm1*-activated lincRNAs ($n = 340$), non-*ddm1*-activated lincRNAs (n = 209), and PCGs expressed in natural populations ($n = 15,376$). **D** IGV visualization of a *ddm1*-activated lincRNA, *MSTRG.19777*, detected in six accessions. **E** DNA methylation levels of *MSTRG.19777* in selected *Arabidopsis* accessions. The scatter plot below illustrating the distribution of six accessions, with the x-axis representing expression levels using $\log_2(\text{RPKM} + 1)$ and the *y*-axis showing the average methylation level within the gene body.

SHapley Additive exPlanations (SHAP) analysis across all six algorithms confirmed the feature importance rankings, again placing population-level DNA methylation at the forefront. High methylation levels in CG, CHG, and CHH contexts positively influenced classification outcomes, biasing predictions toward the positive class (*ddm1*-activated) (Fig. 7G–I and Supplementary Fig. 11D–F). In naive Bayes, nearest neighbors, and SVM models, overlap with TEs, particularly CACTA and Gypsy, also increased the likelihood of positive-class predictions (Fig. 7I and Supplementary Fig. 11D–F).

Discussion

Long-read sequencing technologies, exemplified by ONT, are now widely applied, particularly in genome assembly. At the transcriptome level, DRS has primarily been used to investigate PCGs, enabling the detection of SVs and epigenetic modifications by virtue of its long-read capability and inherent ability to profile native RNA modifications[50]. Here, we demonstrate that DRS can accurately identify and quantify lincRNAs. Annotation against the NONCODE database v6[46] confirmed that DRS not only enables comprehensive characterization of native lincRNAs but also discovers novel noncoding RNA.

The core advantage of DRS lies in its ability to generate exceptionally long reads, which is critical for resolving noncoding RNAs originating from complex genomic regions, offer distinct advantages in differentiating homologous transcripts within complex genomes containing redundant sequences, particularly in polyploid species and in repetitive regions

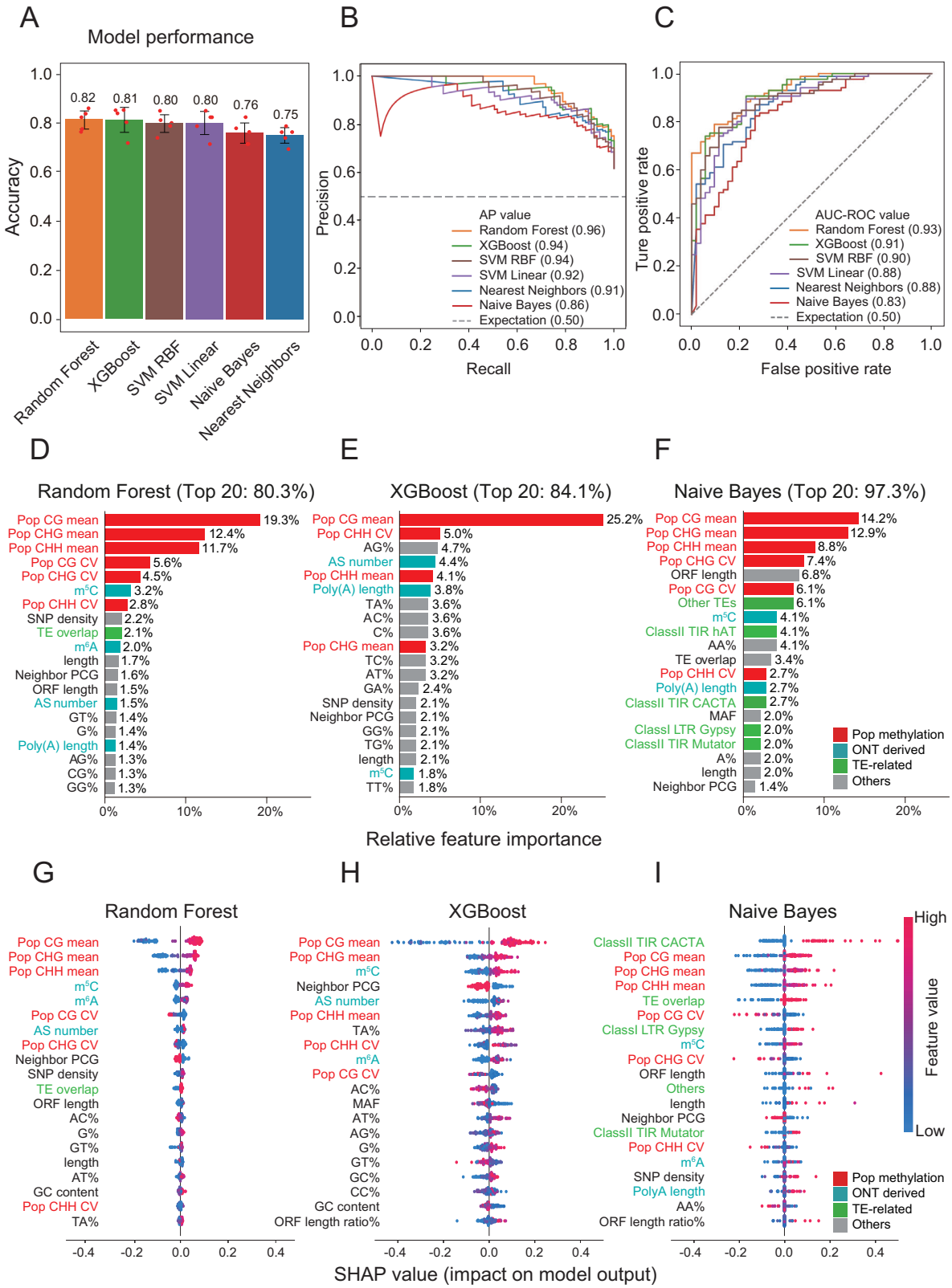

Fig. 7 | Machine learning predicts *ddm1*-activated lincRNAs. A Prediction accuracy of six machine learning models evaluated by five-fold cross-validation. The histogram shows mean accuracy and the dots represent the five-fold cross-validation results. Error bars represent mean ± SD. **B** Precision-recall curve for prediction of *ddm1*-activated lincRNAs, with annotated average precision (AP) values. **C** Receiver operating characteristic (ROC) curve for prediction of *ddm1*-activated lincRNAs, with annotated area under curve (AUC) values. **D**–**F** Top 20 feature importance rankings of the Random Forest, XGBoost, and Naive Bayes models derived from in silico perturbation. **G**–**I** The beeswarm plot of SHAP analysis with top 20 features. Higher SHAP values indicate a stronger contribution of the feature toward the positive outcome (*ddm1*-activated). The red part in the feature value represents a higher value.

enriched for TEs. Although our current dataset is preliminary, it reveals a clear and systematic distinction between *ddm1*-activated and non-activated lincRNAs with high confidence. Given that *Arabidopsis thaliana* is a species with relatively low TE content, the advantages of DRS are expected to be even more pronounced in species with high levels of repetitive sequences, such as maize, cotton, and barley.

Another unmatched advantage of DRS is its ability to directly detect chemical modifications on native RNA molecules. Such molecules frequently carry diverse base modifications, including m^6A and m^5C, which are known to influence RNA stability, transcription, and translation. In this study, we observed that most *Arabidopsis* lincRNAs also harbor m^6A and m^5C modifications, strongly suggesting that lincRNAs may be extensively involved in RNA modification-mediated regulatory networks. These findings open new perspectives for understanding lincRNA functions.

De novo activation of lincRNAs, previously observed in *ddm1* and *met1* mutants, was also detected in natural populations. This phenomenon indicates that genomes without global hypomethylation retain the intrinsic ability to utilize noncoding regions and initiate transcription under specific environmental or developmental contexts. Such capacity represents a fundamental aspect of genome plasticity.

Likewise, despite substantial efforts over the past decade to annotate lincRNAs across multiple species, determining the genetic and functional consequences of lincRNA variants remains a significant challenge. Johannes et al. employed two parental lines with nearly identical DNA sequences but contrasting DNA methylation profiles, one wild type (WT) and the other a mutant in the *DDM1* gene, to generate a set of epiRILs in *Arabidopsis thaliana*[25]. These epiRILs exhibited substantial heritable variation in flowering time and plant height (~30%), and several parental DNA methylation variants (epialleles) were stably inherited for at least eight generations[25]. epiRILs display pleiotropic phenotypic variation[24,25,53]. Furthermore, large-scale methylation mapping efforts have been conducted across diverse germplasm resources, resulting in published DNA methylation atlases for *Arabidopsis*[21], rice[54], cotton[13], tomato[55], soybean[56], maize[57], and pear[58]. These studies have revealed extensive DNA methylation polymorphisms within populations and have identified numerous epigenetic loci associated with phenotypic traits through epigenome-wide association analyses. Our findings demonstrate that DNA methylation exerts a far greater impact on the lincRNA transcriptome than on that of PCGs, thereby offering greater opportunities for selection and innovation in genome plasticity.

Molecular breeding has traditionally relied on the stable and heritable genetic variation found in DNA sequences. However, long-term artificial selection has narrowed genetic diversity, limiting the potential of conventional hybridization in many elite crops. Epigenetic variation, in which lincRNAs may serve as key regulators, represents a promising yet often underexplored resource. Epigenetics offers a paradigm to generate novel phenotypic diversity without altering the underlying DNA sequence. By inducing or selecting for epigenetic modifications, it is possible to create new epialleles that reshape transcriptional networks, activating or silencing key genes[59,60]. LincRNAs were not only involved in regulating crucial agronomic traits like development[61,62] and flowering time[15] but also highly responsive to environmental cues, acting as a bridge between the environment and the genome[17]. For example, the lincRNA *IPS1*, for instance, functions as a critical molecular signal in the plant response to phosphate starvation[63].

LincRNAs may be less practical than DNA-level markers for routine breeding, but their greatest value lies in enabling the generation of novel germplasm. Inspired by the success of epiRILs in *Arabidopsis*, loss-of-function mutants of DNA methyltransferases could be developed in major crops such as rice, maize, and wheat. By applying systematic crossing and backcrossing schemes, these populations would harbor extensive, stably heritable variation in both DNA methylation and lincRNA transcription. Such an approach could systematically activate silent genomic elements, thereby unlocking novel combinations of agronomic traits and driving genuine genomic innovation for the future of crop breeding.

Conclusion

We investigated the transcriptional flexibility of lincRNAs under DNA methylation-deficient conditions using *Arabidopsis thaliana* mutants (*ddm1* and *met1*). By leveraging DRS and DDS, we achieved accurate lincRNA assembly, comprehensive profiling of epi-transcriptomic modifications, precise measurement of poly(A) tail lengths, and direct quantification of DNA methylation without chemical treatment.

Our results demonstrate that DNA hypomethylation induces widespread activation of lincRNAs while exerting minimal influence on protein-coding gene transcription. Notably, a subset of *ddm1*-activated lincRNAs was detected in natural *Arabidopsis* populations, indicating that such epigenetic states can arise in nature and may contribute to transcriptomic diversity at the population level.

By integrating ONT-derived features with multi-omics annotations, we developed robust machine learning models capable of classifying *ddm1*-activated lincRNAs with high accuracy. Population-level DNA methylation consistently emerged as the most influential predictive feature, with transposable element overlap and RNA chemical modifications providing additional predictive power. These findings highlight the interplay between genome repeats and epigenetic variation in shaping lincRNA activation potential.

Collectively, our work establishes a multi-omics analytical framework for identifying and characterizing methylation-sensitive lincRNAs, elucidates the regulatory connection between epigenome dynamics and non-coding RNA activation. By confirming that hypomethylation-driven lincRNA activation can be recapitulated in natural populations, lincRNAs can serve as a reservoir of transcriptional innovation with potential adaptive and evolutionary significance in plants.

Methods

Study design

This study analyzed leaf tissue from three *Arabidopsis thaliana* genotypes: the wild-type (WT) Columbia (Col-0) accession, the *ddm1-2* mutant (locus: *AT5G66750*), and the *met1-3* mutant (TAIR stock: CS16394, locus: *AT5G49160*). The samples were subjected to three sequencing methodologies: ONT DRS, ONT DDS, and Illumina RNA sequencing (RNA-seq).

Plant materials and growth conditions

Seeds of *Arabidopsis thaliana* WT (Col-0), *ddm1* and *met1* mutants were surface-sterilized by sequential immersion in 20% sodium hypochlorite (NaClO) for 10 min and 70% ethanol for 5 min. Following sterilization, the seeds were washed four times with sterile distilled water and plated on half-strength Murashige and Skoog (MS) agar. The plates were then incubated in a growth chamber for 30 days under a 16 h light/8 h dark photoperiod. For each genotype, four biological replicates were generated, each consisting of approximately 2 g of leaf tissue harvested from 30-day-old plants. The samples were immediately flash-frozen in liquid nitrogen.

Nanopore DRS library preparation

Total RNA of two randomly selected biological replicates was extracted from *Arabidopsis thaliana* leaf tissue using TRIzol reagent (Invitrogen) following the manufacturer's protocol. The concentration and purity of the extracted RNA were determined using a Qubit RNA HS Assay Kit and a NanoDrop 2000 spectrophotometer (both from Thermo Fisher Scientific), respectively. Subsequently, RNA integrity was validated on an Agilent 2100 Bioanalyzer, with only samples exhibiting an RNA Integrity Number (RIN) of ≥8.0 proceeding to downstream steps. Polyadenylated [poly(A)+] RNA was subsequently enriched from 2–5 μg of total RNA using the Dynabeads™ mRNA Purification Kit (Invitrogen). The concentration of the resulting poly(A) + RNA was then re-quantified using a Qubit assay. For each sample, approximately 200–500 ng of the purified poly(A) + RNA served as the input for library preparation using the DRS Kit (SQK-RNA002, ONT), strictly following the official protocol. The prepared libraries were loaded onto individual PromethION R9.4.1 flow cells (FLO-PRO002, ONT) and sequenced for a duration of 48 to 72 h. To minimize

batch effects, each biological replicate for each condition was sequenced independently on a separate flow cell. All sequencing services were performed by Wuhan Benagen Tech Solutions Company Limited. This process generated raw signal data in multi-read FAST5 files, which were used for subsequent downstream analyses.

Alignment of DRS

Raw multi-FAST5 files generated were base-called using Guppy (v 5.0.11)[51] in high-accuracy mode, which can achieve a calling accuracy of more than 99%. Reads were separated into "pass" and "fail" folders according to the default Qscore threshold (Phred ≥ 7), only pass reads were retained for downstream analysis.

For each sample, ONT long reads were aligned to the *Arabidopsis thaliana* TAIR10 reference genome using Minimap2 (v2.18)[64], a splice-aware long-read aligner, with the parameters -ax splice -k14 -uf --secondary=no to allow spliced alignment of direct RNA reads while suppressing secondary mappings. The resulting SAM files were converted to BAM format, sorted by genomic coordinates, and indexed using SAMtools (v1.6)[65]. Alignments flagged as secondary or supplementary were removed, and only primary alignments on the forward (flag 0) or reverse (flag 16) strand were retained for further lincRNA for subsequent lincRNA assembly and quantification.

Generation of lincRNA annotations using DRS

For each sample, transcript models were assembled from the DRS reads using StringTie2 (v2.0)[66] in its long-read mode (--long). This assembly was guided by the *Arabidopsis thaliana* TAIR10 reference genome annotation. To enhance the accuracy of the isoform annotations, reads with similar splice-junction coordinates were clustered using FLAMES (v1.0)[67]. This tool refines transcript models by correcting potential errors in splice sites and boundaries through comparison with the reference annotation. To reduce redundancy and merge potentially incomplete transcripts, an assembled isoform was collapsed into an existing reference model if its splice junctions were within 5bp and its start or end coordinates were within 100 bp of that reference. Following this, only isoforms supported by a minimum of five uniquely mapped reads were retained, defining our set of high-confidence transcripts. These high-confidence isoforms were then compiled into a single, non-redundant GTF annotation file. Novel intergenic transcripts were then identified using gffcompare (v0.12.2)[68] by comparing this new annotation to the TAIR10 reference. Based on gffcompare (v0.12.2)[68] classification codes, transcripts classified as "u", "x", and "i" were labeled as lincRNAs, natural antisense transcripts, and intronic RNAs, respectively and only transcripts labeled "u" were considered as candidate lincRNAs. The sequences of these candidate lincRNAs were extracted, and their protein-coding potential was assessed using two independent methods: CPC2[69] (v2.0b, default parameters) and Pfam-scan via HMMER v3.2.1 (https://github.com/EddyRivasLab/hmmer) against the Pfam-A database[70]. Candidate lincRNAs that were predicted as noncoding by CPC2(v2.0b) and showed no significant matches to any Pfam protein domain were designated as high-confidence lincRNAs. This final, long-read-derived transcript annotation containing PCGs as well as lincRNAs, served as the basis for all subsequent downstream analyses.

Gene and transcript abundance estimation from DRS data

Gene-level read counts were quantified from the mapped reads using FeatureCount (v1.6.2)[71]. The quantification was guided by our long-read-derived transcript annotation (in GTF format), and the parameters --longRead and --primaryOnly were specified to accommodate Nanopore reads and to count only primary alignments, respectively. The resulting raw count matrix served as the input for differentially expressed genes (DEGs) analysis, which was performed using the DESeq2 package (v1.44.0)[72] with the negative binomial statistic approach. The cutoff criteria for the selection of DEGs were: $|\log_2\text{-fold change}| > 1$ and false discovery rate (FDR) < 0.05. For visualization purposes, gene expression levels were normalized to CPM. To evaluate sequencing saturation, long reads obtained from DRS were randomly subsampled to datasets containing between 0.5 and 5 million reads, at increments of 0.5 million. For each subsampling level, five independent random replicates were generated. Gene expression was quantified using CPM, and genes with CPM > 1 were classified as expressed. The total number of expressed genes was recorded for each replicate, and saturation curves were generated to assess the relationship between sequencing depth and transcriptome coverage.

Poly(A) tail length estimation and alternative polyadenylation using DRS

The DRS library preparation protocol involves ligating an adapter directly to the 3' poly(A) tail. This distinctive feature allows estimation of poly(A) tail length from the raw ionic current signal generated as an RNA molecule passes through the nanopore. Poly(A) tail lengths were estimated on a per-read basis using the "polya" module of Nanopolish (v0.11.0)[73] with "polya-merged". Only length estimates that passed the internal quality control filter (QC tag: "Pass") were retained for downstream analysis. These filtered estimates were then aggregated at both the gene and isoform levels using a custom in-house Perl script.

Detection of methylated adenosine and cytosine in DRS

As a prerequisite for modification analysis, the base-called multi-read FAST5 files were converted into a single-read FAST5 format using the ont_fast5_api Python library (https://github.com/nanoporetech/ont_fast5_api). RNA modification analysis was performed using Tombo (v1.5.1)[74] with default paramters. The raw ionic current signals were first re-aligned to the reference transcriptome using the resquiggle command, followed by modification detection with the Tombo detect_modifications tool[74]. To identify potential modification sites, this detection step was executed twice: once using the canonical 5-methylcytosine (5mC) model and once using the de novo model. In parallel, the specific detection of N^6-methyladenosine (m^6A) was conducted using MINES with its default parameters (https://github.com/YeoLab/MINES). These filtered estimates were then aggregated at both the gene and isoform levels using a custom in-house Perl script.

DDS library preparation

High-quality genomic DNA (gDNA) was isolated from fresh leaf tissue using a standard CTAB protocol. The extracted gDNA was then subjected to quality control: its integrity was verified on a 0.75% agarose gel, its purity was assessed with a NanoDrop One spectrophotometer (Thermo Fisher Scientific), and its concentration was precisely quantified using a Qubit 3.0 Fluorometer (Life Technologies). The final libraries were sequenced on a PromethION instrument (ONT) using R9.4.1 Cells. Sequencing was conducted for 48 h by the service provider, Wuhan Benagen Tech Solutions Company Limited. Basecalling of the raw sequencing data was performed using Guppy software (v5.0.11)[51].

Structure variation (SV) calling based on DDS

The resulting long reads that passed quality filtering were aligned to the *Arabidopsis thaliana* TAIR10 reference genome using Minimap2 (v2.18)[64], and SVs were called with cuteSV (v1.0.12)[75], a sensitive SV detection implementation for long reads sequencing, from processed bam files using the suggested parameters for ONT: --max_cluster_bias_INS 100, --diff_ratio_merging_INS 0.3, -max_cluster_bias_DEL 100, --diff_ratio_merging_DEL 0.3. SV information was stored in VCF format[76].

Illumina RNA sequencing library preparation

RNA-seq libraries were constructed using a modified TruSeq RNA Sample Preparation protocol (Illumina, San Diego, CA, USA)[16]. Briefly, total RNA was first treated with DNase I to remove potential genomic DNA contamination. Subsequently, ribosomal RNA (rRNA) was depleted from the total RNA using custom-designed complementary DNA oligos (Integrated DNA Technologies, Coralville, IA, USA) and RNase H-mediated digestion (Invitrogen, Waltham, MA, USA). The resulting rRNA-depleted RNA was used for library construction. The final libraries were sequenced on an

Illumina HiSeq 2500 platform to generate 150-bp paired-end (PE150) reads. All library preparation and sequencing services were provided by Benagen (Wuhan, China).

Gene and transcript abundance estimation based on short reads

The raw paired-end 150-bp short reads were subjected to quality trimming and adapter removal using fastp (v0.20.1)[77], The resulting high-quality reads were then aligned to the *Arabidopsis thaliana* TAIR10 reference genome using HISAT2 (v2.2.1)[78] with the --dta parameter. Gene expression quantification was performed with StringTie2 (v2.0)[66] with default parameters, using a long-read-derived transcript annotation that included lincRNAs. Expression levels were normalized and reported as FPKM. Differential expression analysis was carried out with DESeq2 (v1.44.0) using the negative binomial statistical model[72]. Cutoff criteria for the selection of DEGs were: $|\log_2\text{-fold change}| > 1$ and FDR < 0.05.

Detection of single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) based on DRS and Illumina RNA-seq

Variant calling was performed for all replicates of WT, ddm1, and met1 genotypes, including sequencing data from both ONT and Illumina platforms. To enable a fair comparison of variants between the two technologies, identical alignment and variant detection workflows were applied to all samples. Reads were aligned to the TAIR10 reference genome with minimap2 (v2.18)[64] and PCR duplicates were marked and removed by using SAMtools (v1.6)[65]. SNPs and InDels were identified using BCFtools (v1.15.1)[79]. For each dataset, variant calling began with BCFtools (v1.15.1)[79] "mpileup" to compute genotype likelihoods against the reference genome, followed by "call -mv" to identify raw variant sites. Raw variant calls were filtered to retain high-confidence variants, defined as having a quality score (QUAL) ≥ 30, read depth (DP) ≥ 10, and mapping quality (MQ) ≥ 30. Sites with missing genotypes were removed. Filtered VCF files for WT, *ddm1*, and *met1* were subsequently compared to assess variant differences across replicates and sequencing platforms.

Public data integration for gene expression and DNA methylation profiling

Publicly available transcriptomes from 728 *Arabidopsis* natural accessions from the 1001 Genomes Project[21] were downloaded from Gene Expression Omnibus (GEO, accession id: GSE80744). This dataset consisted of single-end sequencing reads. The reads were quality trimmed using fastp (v0.20.1)[77], then aligned to the *Arabidopsis thaliana* TAIR10 reference genome using HISAT2 (v2.2.1)[78] with the following parameter: --dta, guided by long-read-derived transcript annotation. As the dataset used single-end sequencing, gene expression levels were quantified using RPKM with StringTie (v2.0)[66] with the long-read-derived transcript annotation.

Corresponding methylomes were retrieved from GEO (accession id: GSE231564)[52]. The Illumina bisulfite sequencing (BS-Seq) data[52] was quality controlled by fastp (v0.20.1)[77]. The clean reads were then aligned to the corresponding reference, and DNA methylation levels were calculated using Bismark (v0.22.3)[80] with default parameters. Single-base methylation levels were calculated as the number of methylated reads divided by the total number of reads covering the site. Only sites with a sequencing depth of ≥5 were retained for further analysis. To compare methylation between ONT DDS and BS-seq, the average methylation ratio (defined as the number of methylated cytosines divided by the total number of callable cytosines) was calculated within non-overlapping 5-kb genomic windows. Correlation analyses were performed in R software (v4.4.1) by Pearson correlation, and statistical significance was evaluated with a two-tailed Student's *t* test.

Prediction of ddm1-activated lincRNAs using a multi-omics features

Machine learning analyses were performed in Python (v3.10.9). The XGBoost model was implemented using the xgboost library (v2.1.4), and five additional models, random forest, SVM with linear and RBF kernels, k-Nearest Neighbors, and Gaussian Naive Bayes, were implemented using the scikit-learn library (v1.6.1). Model hyperparameters were optimized via grid search, with the optimal configuration determined by the highest mean accuracy in five-fold cross-validation.

For feature importance analysis, Random Forest utilized the built-in RandomForestClassifier scores, while linear-kernel SVM employed the absolute values of model weights. For RBF-kernel SVM, k-nearest neighbors, Gaussian naive Bayes, and XGBoost, permutation importance was calculated using scikit-learn, measuring the change in model performance when individual features were shuffled. Feature importance scores from each model were normalized to percentages. Additionally, SHAP analysis was conducted using the shap package (v0.48.0)[81].

A total of fifty multi-omics features were generated for model training, comprising population-level epigenetic features, ONT-derived features, TE-related features, and other sequence/population variation features. Details are as follows:

Population-level DNA methylation features ($n = 6$): Mean level and CV across lincRNA bodies in CG, CHG, and CHH contexts. (Data from 728 *Arabidopsis thaliana* accessions[21]).

Population-level genetic variation features ($n = 2$): SNP density and mean minor allele frequency across lincRNA bodies (Data from *Arabidopsis thaliana* 1001 Genomes Project, https://1001genomes.org/data/GMI-MPI/releases/v3.1/).

ONT-derived features ($n = 5$): Exon number (number of exons per lincRNA, based on ONT DRS annotation); alternative splicing number (from ONT DRS transcript assemblies); RNA m^6A and m^5C modification status (binary indicator, 1 = modified, 0 = not modified, derived from DRS data); poly(A) tail length (from DRS poly(A) analysis).

TE-related features ($n = 9$): Binary overlap indicators (1 = overlap, 0 = no overlap) with the following classes: ClassII TIR/CACTA, ClassII TIR/Mutator, ClassII RC/Helitron, ClassII TIR/hAT, ClassI LTR/Copia, ClassI LTR/Gypsy, ClassI non-LTR, Other TEs and overall TE overlap (annotations from TAIR10 genome release: https://www.arabidopsis.org/download/).

Sequence composition and structural features ($n = 28$): lincRNA length; nucleotide composition: A%, T%, C%, G%; dinucleotide frequencies; GC content; presence of PCGs within 10 kb upstream or downstream; open reading frame (ORF) length and ORF length ratio (calculated using EMBOSS, v6.6.0.0, getorf[82]); fickett score (calculated using CPC2 v2.0b[69]); predicted coding probability (calculated using CPC2 v2.0b[69]); isoelectric point (pI, calculated using CPC2 v2.0b[69]).

Statistics

All statistical analyses were performed in R software (v4.4.1). Two-tailed unpaired Student's *t* tests were used for Fig. 2C, Supplementary Fig. S3, and Fig. 5A–D, whereas two-tailed paired Student's *t* tests were applied to Fig. 3C, D. For Fig. 4B and Supplementary Fig. S7, Wilcoxon rank-sum test was conducted. For Supplementary Fig. S8, a Kruskal–Wallis rank sum test was conducted, followed by Dunn's post hoc test with Bonferroni correction for multiple comparisons. Statistical significance was defined as $P < 0.05$. Levels of significance are indicated as follows: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$, ns (not significant). Data are presented as mean ± standard deviation.

Reporting summary

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

Data availability

Raw reads from ONT and Illumina RNA-seq experiments have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA780475. Genome-wide methylation data derived from ONT DDS (Fig. 5A–D) are available at Figshare (https://doi.org/10.6084/m9.figshare.30879272). All other data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

Code for machine learning model construction is available at https://github.com/sxppppp/DRS_machine_learning.

Received: 8 April 2025; Accepted: 18 December 2025;
Published online: 31 December 2025

References

1. Lee, H., Zhang, Z. & Krause, H. M. Long noncoding RNAs and repetitive elements: junk or intimate evolutionary partners? *Trends Genet.* **35**, 892–902 (2019).
2. Palazzo, A. F. & Koonin, E. V. Functional long non-coding RNAs evolve from junk transcripts. *Cell* **183**, 1151–1161 (2020).
3. Waseem, M., Liu, Y. & Xia, R. Long non-coding RNAs, the dark matter: an emerging regulatory component in plants. *Int. J. Mol. Sci.* **22**, 86 (2020).
4. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
5. Andergassen, D. & Rinn, J. L. From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo. *Nat. Rev. Genet.* **23**, 229–243 (2022).
6. Heo, J. B. & Sung, S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331**, 76–79 (2011).
7. Zhao, Y. et al. Natural temperature fluctuations promote COOLAIR regulation of FLC. *Genes Dev.* **35**, 888–898 (2021).
8. Wang, D. et al. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. *Plant J.* **90**, 133–146 (2017).
9. Zheng, X. M. et al. Genome-wide analyses reveal the role of noncoding variation in complex traits during rice domestication. *Sci. Adv.* **5**, eaax3619 (2019).
10. Mele, M. et al. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
11. de Goede, O. M. et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184**, 2633–2648.e19 (2021).
12. Vakirlis, N. et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
13. Zhao, T. et al. Population-wide DNA methylation polymorphisms at single-nucleotide resolution in 207 cotton accessions reveal epigenomic contributions to complex traits. *Cell Res.* **34**, 859–872 (2024).
14. Wang, Z. et al. lncRNA epigenetic landscape analysis identifies EPIC1 as an oncogenic lncRNA that interacts with MYC and promotes cell-cycle progression in cancer. *Cancer Cell* **33**, 706–720.e9 (2018).
15. Schmid, M. W. et al. Contribution of epigenetic variation to adaptation in Arabidopsis. *Nat. Commun.* **9**, 4446 (2018).
16. Zhao, T. et al. LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. *Genome Biol.* **19**, 195 (2018).
17. Tao, X. et al. Neofunctionalization of a polyploidization-activated cotton long intergenic non-coding RNA DAN1 during drought stress regulation. *Plant Physiol.* **186**, 2152–2168 (2021).
18. Wang, P. et al. Global profiling of lncRNAs expression responsive to allopolyploidization in cucumis. *Genes* **11**, 1500 (2020).
19. Zhao, T. et al. Absence of CG methylation alters the long noncoding transcriptome landscape in multiple species. *FEBS Lett.* **595**, 1734–1747 (2021).
20. Tan, F. et al. DDM1 represses noncoding RNA expression and RNA-directed DNA methylation in heterochromatin. *Plant Physiol.* **177**, 1187–1197 (2018).
21. Kawakatsu, T. et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**, 492–505 (2016).
22. Furci, L. et al. Identification and characterisation of hypomethylated DNA loci controlling quantitative resistance in Arabidopsis. *Elife* **8**, e40655 (2019).
23. Kooke, R. et al. Epigenetic basis of morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. *Plant Cell* **27**, 337–348 (2015).
24. Zhang, Y. Y., Latzel, V., Fischer, M. & Bossdorf, O. Understanding the evolutionary potential of epigenetic variation: a comparison of heritable phenotypic variation in epiRILs, RILs, and natural ecotypes of *Arabidopsis thaliana*. *Heredity* **121**, 257–265 (2018).
25. Johannes, F. et al. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.* **5**, e1000530 (2009).
26. He, L. et al. A naturally occurring epiallele associates with leaf senescence and local climate adaptation in Arabidopsis accessions. *Nat. Commun.* **9**, 460 (2018).
27. Glenfield, C. & McLysaght, A. Pseudogenes provide evolutionary evidence for the competitive endogenous RNA hypothesis. *Mol. Biol. Evol.* **35**, 2886–2899 (2018).
28. 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).
29. Du, J. M., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16**, 519–532 (2015).
30. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
31. Osakabe, A. et al. The chromatin remodeler DDM1 prevents transposon mobility through deposition of histone variant H2A.W. *Nat. Cell Biol.* **23**, 391–U177 (2021).
32. Vongs, A., Kakutani, T., Martienssen, R. A. & Richards, E. J. *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**, 1926–1928 (1993).
33. Kankel, M. W. et al. Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics* **163**, 1109–1122 (2003).
34. Lippman, Z. et al. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004).
35. Quadrana, L. & Colot, V. Plant transgenerational epigenetics. *Annu. Rev. Genet.* **50**, 467–491 (2016).
36. Tao, P. et al. The overexpression of lncRNA MEG3 inhibits cell viability and invasion and promotes apoptosis in ovarian cancer by sponging miR-205-5p. *Int. J. Clin. Exp. Pathol.* **13**, 869–879 (2020).
37. Ulitsky, I. & Bartel, D. avidP. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
38. Kapusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
39. Babarinde, I. A. et al. Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res.* **49**, 9132–9153 (2021).
40. Wang, X. et al. Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. *Nat. Commun.* **10**, 5009 (2019).
41. Ebbert, M. T. W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* **20**, 97 (2019).
42. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
43. Seki, M. et al. Evaluation and application of RNA-Seq by MinION. *DNA Res.* **26**, 55–65 (2019).
44. Qin, H. et al. DENA: training an authentic neural network model using Nanopore sequencing data of Arabidopsis transcripts for detection

and quantification of N(6)-methyladenosine on RNA. *Genome Biol.* **23**, 25 (2022).
45. Zhang, S. et al. New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. *Nucleic Acids Res.* **48**, 7700–7711 (2020).
46. Zhao, L. et al. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res.* **49**, D165–D171 (2021).
47. Monzo, C., Liu, T. & Conesa, A. Transcriptomics in the era of long-read sequencing. *Nat. Rev. Genet.* **26**, 681–701 (2025).
48. Liu-Wei, W. et al. Sequencing accuracy and systematic errors of nanopore direct RNA sequencing. *BMC Genom.* **25**, 528 (2024).
49. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
50. Sun, K. et al. Advances in nanopore direct RNA sequencing and its impact on biological research. *Biotechnol. Adv.* **85**, 108710 (2025).
51. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
52. Lee, S. C. et al. Chromatin remodeling of histone H3 variants by DDM1 underlies epigenetic inheritance of DNA methylation. *Cell* **186**, 4100–4116 e15 (2023).
53. Cortijo, S. et al. Mapping the epigenetic basis of complex traits. *Science* **343**, 1145–1148 (2014).
54. Cao, S. et al. Asymmetric variation in DNA methylation during domestication and de-domestication of rice. *Plant Cell* **35**, 3429–3443 (2023).
55. Niu, K.-X. et al. Suppressing ASPARTIC PROTEASE 1 prolongs photosynthesis and increases wheat grain weight. *Nat. Plants* **9**, 965–977 (2023).
56. Shen, Y. et al. DNA methylation footprints during soybean domestication and improvement. *Genome Biol.* **19**, 128 (2018).
57. Xu, G. et al. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nat. Commun.* **11**, 5539 (2020).
58. Song, B. et al. Increased DNA methylation contributes to the early ripening of pear fruits during domestication and improvement. *Genome Biol.* **25**, 87 (2024).
59. Srikant, T. & Tri Wibowo, A. The underlying nature of epigenetic variation: origin, establishment, and regulatory function of plant epialleles. *Int. J. Mol. Sci.* **22**, 8618 (2021).
60. Yan, H. et al. Autopolyploidization in switchgrass alters phenotype and flowering time via epigenetic and transcription regulation. *J. Exp. Bot.* **70**, 5673–5686 (2019).
61. Zhang, Y.-C. et al. Long noncoding RNAs as molecular architects: shaping plant functions and physiological plasticity. *Mol. Plant* **18**, 1643–1671 (2025).
62. Zhang, Y. C., Yuan, C. & Chen, Y. Q. Noncoding RNAs and their roles in regulating the agronomic traits of crops. *Fundam. Res.* **3**, 718–726 (2023).
63. Franco-Zorrilla, J. M. et al. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* **39**, 1033–1037 (2007).
64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
65. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
67. Tian, L. et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol.* **22**, 310 (2021).
68. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
69. Kang, Y. J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).
70. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
71. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
73. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
74. Stoiber, M. et al. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. Preprint at *bioRxiv*, https://doi.org/10.1101/094672 (2017).
75. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
76. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
78. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
79. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
80. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
81. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems*, 4768–4777 (Curran Associates Inc., 2017).
82. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).

Acknowledgements

This study was supported by the Biological Breeding-National Science and Technology Major Project (2023ZD04076).

Author contributions

T.Z. conceptualized the project. X.G., T.Z., L.W., Y.H., and K.N. performed the experiment. W.S., T.Z., N.Z., M.Z., and X.G. analyzed the data. W.S., T.Z., and X.G. prepared the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-09454-1.

Correspondence and requests for materials should be addressed to Ting Zhao.

Peer review information *Communications Biology* thanks Tatiana Garcia and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary handling editors: Jorge Duitama and Kaliya Georgieva. A peer review file is available.