

Epigenomic Prediction of Transposable Element Reactivation under Heat Stress in *Arabidopsis thaliana*

Sudipta Joardar^{1,2} and Swapnil Anil Prabhudesai^{3,4}

¹ Towards Bioscience (towardsbioscience.com), India. ² University of Bologna, Bologna, Italy.

³ Department of Basic Sciences & Humanities, SVKM's NMIMS Mukesh Patel School of Technology Management & Engineering, Mumbai, India. ⁴ Heinz Maier-Leibnitz Zentrum (MLZ), Technische Universität München, Lichtenbergstraße 1, 85748 Garching, Germany.

✉ Correspondence: joardars2025@gmail.com | sudipta.joardar@studio.unibo.it

Highlights

- Genome-wide epigenomic feature matrix constructed for 9,990 annotated TE loci in *Arabidopsis thaliana*
 - Random forest and logistic regression models achieve ROC-AUC up to 0.80 for predicting heat-induced TE reactivation
 - CHH methylation and 24nt siRNA density are the strongest predictors of reactivation risk
 - Hi-C compartment score independently contributes to prediction, linking 3D genome organization to TE silencing
 - Risk score framework identifies Copia and LINE/L1 elements as highest-risk TE families
-

Abstract

Transposable elements (TEs) constitute approximately 21% of the *Arabidopsis thaliana* genome and are maintained in a silenced state by overlapping epigenomic mechanisms including cytosine methylation, 24-nucleotide small interfering RNA (siRNA)-directed DNA methylation (RdDM), and repressive histone marks. Heat stress disrupts this multilayered defence, enabling select TE families — most notably LTR retrotransposons such as ONSEN — to escape transcriptional silencing. However, the genome-wide determinants of which specific loci will reactivate under thermal stress remain poorly characterised. Here we present the first predictive modelling framework for TE heat-stress reactivation in plants. We assembled a multi-omic feature matrix for 9,990 annotated TE loci integrating pre-stress CHH, CHG and CG methylation, 24nt siRNA density, H3K9me2 and H3K27me1 histone mark enrichment, Hi-C compartment scores, and topologically associating domain (TAD) boundary proximity. Logistic regression, random forest, and gradient boosting classifiers were trained on binary reactivation labels and evaluated by five-fold stratified cross-validation. Logistic regression achieved the highest discriminative performance (ROC-AUC = 0.80 ± 0.03 ; sensitivity = 0.71; specificity = 0.83). Permutation-based feature importance revealed CHH methylation and 24nt siRNA density as the dominant predictors, with Hi-C compartment score making an independent, statistically

significant contribution. Predicted high-risk loci were significantly enriched among Copia and LINE/L1 elements and showed a marked bias toward euchromatic A-compartment positions. This work establishes a computational framework that can be generalised to other stress conditions and plant species, and provides a prioritised set of loci for downstream functional validation.

Keywords: Transposable elements; heat stress; DNA methylation; epigenomics; machine learning; 3D genome; RdDM pathway

1. Introduction

Transposable elements (TEs) are mobile genetic elements that constitute a substantial fraction of plant genomes. In *Arabidopsis thaliana*, the compact model plant genome of approximately 135 Mb, TEs and TE-derived sequences account for roughly 21% of the total sequence, with the majority clustered in pericentromeric heterochromatin (The Arabidopsis Genome Initiative, 2000; Ito et al., 2011). Despite their genomic abundance, the vast majority of TE copies are maintained in a transcriptionally silent state under normal growth conditions through a hierarchy of epigenomic silencing mechanisms (Mirouze and Paszkowski, 2011; Matzke and Mosher, 2014).

The primary defence against TE reactivation in plants is the RNA-directed DNA methylation (RdDM) pathway, which deposits cytosine methylation in the CHH sequence context (where H = A, T or C) at TE bodies and their regulatory sequences (Matzke and Mosher, 2014; Law and Jacobsen, 2010). This non-CG methylation, established by the DRM2 methyltransferase and guided by 24-nucleotide siRNAs generated by Pol IV–RDR2–DCL3, constitutes a renewable silencing signal that must be continuously maintained (Stroud et al., 2013). Complementing this pathway, the repressive histone mark H3K9me2 — deposited by the HKMTs SUVH4, SUVH5 and SUVH6 — enforces chromatin compaction at TE loci and is tightly coupled to non-CG methylation through the reader-writer loops involving CMT3 and CMT2 (Du et al., 2015; Zemach et al., 2013).

Heat stress profoundly disrupts TE silencing. Seminal work by Ito and colleagues demonstrated that the LTR retrotransposon ONSEN is transcriptionally activated during heat treatment, in part because its long terminal repeat (LTR) contains a heat-shock element (HSE) recognised directly by the transcription factor HsfA2 (Ito et al., 2011). Subsequent work revealed that double mutants lacking functional Pol IV or RDR2 show increased TE transcription under heat and, crucially, accumulate new somatic TE insertions in subsequent generations (Ito et al., 2011; Matsunaga et al., 2015). Parallel studies have documented that heat stress leads to a

genome-wide reduction in CHH methylation at TE bodies, redistribution of H3K9me2, and partial dispersal of chromocentre structures that are visible by fluorescence microscopy (Birnbaum and Roudier, 2017).

Despite this mechanistic foundation, a critical gap remains: it is unknown which of the approximately 10,000 annotated TE loci will escape silencing under heat stress, and what pre-stress epigenomic features best predict this outcome. Existing studies have focused on individual loci (ONSEN, Copia-35) or on whole-pathway mutants, and have not attempted to build predictive models that operate at the genome-wide scale. This is consequential because TE reactivation is not uniform — only a subset of loci mobilise under any given stress condition, and the molecular basis for this selectivity is not understood.

Simultaneously, remarkable progress has been made in understanding the three-dimensional (3D) organisation of plant genomes. Hi-C studies have revealed that *Arabidopsis* chromatin is organised into A (euchromatic) and B (heterochromatic) compartments, and that pericentromeric heterochromatin — where most TEs reside — is enriched in B-compartment contacts (Liu et al., 2016; Grob et al., 2014). A key, untested prediction is that TEs occupying A-compartment positions or near TAD boundaries may be more accessible to heat-activated transcription factors and thus at higher risk of reactivation. Integrating 3D genome features into a TE reactivation model has not previously been attempted.

Here we address this gap by constructing the first genome-wide machine learning framework for predicting TE reactivation risk under heat stress in *Arabidopsis thaliana*. We integrate eight pre-stress epigenomic features — three cytosine methylation contexts, siRNA density, two histone marks, Hi-C compartment score, and TAD boundary distance — across 9,990 annotated TE loci and train three complementary classifiers to predict heat-induced reactivation. Our results identify the dominant predictors of reactivation risk, reveal a previously unappreciated contribution of 3D genome organisation, and provide a prioritised set of candidate loci for experimental follow-up. The risk-scoring framework we develop is transferable to other stress conditions and plant species.

2. Materials and Methods

2.1 TE annotation and genome reference

All analyses were anchored to the TAIR10 reference genome assembly of *Arabidopsis thaliana* ecotype Columbia-0 (Col-0). TE annotation was obtained from the TAIR10 RepeatMasker track, supplemented by the EDTA-generated annotation (Ou et al., 2019). After filtering for loci with

complete feature coverage, 9,990 TE loci were retained spanning all five chromosomes. TE families were classified according to Wicker et al. (2007): Copia (LTR/Copia; n = 1,498), Gypsy (LTR/Gypsy; n = 2,497), LINE/L1 (n = 1,998), MULE (DNA/MuLE; n = 997), hAT (DNA/hAT; n = 997), Helitron (RC/Helitron; n = 998), SINE (n = 505), and EnSpm/CACTA (DNA/EnSpm; n = 500).

2.2 Epigenomic feature extraction

2.2.1 DNA methylation

Whole-genome bisulfite sequencing (WGBS) data from control (22°C) and heat-stressed (37°C, 6 h) *Arabidopsis* plants were processed using Bismark v0.24.0 (Krueger and Andrews, 2011) aligned to TAIR10 with default parameters. Methylation calls were extracted for CG, CHG, and CHH contexts. For each TE locus, the mean percentage methylation across all covered cytosines within the TE body $\pm$ 500 bp was computed using bedtools v2.31 (Quinlan and Hall, 2010). A minimum coverage threshold of 5 $\times$ per cytosine was applied. Differentially methylated regions (DMRs) were called using DSS (Wu et al., 2015) with a threshold of $|\Delta\text{methylation}| > 10\%$ and FDR < 0.05.

2.2.2 Small RNA profiling

Small RNA sequencing reads were trimmed using Trim Galore v0.6.7 and aligned to TAIR10 using ShortStack v4.0 (Johnson et al., 2016). Reads mapping to annotated TE loci were size-selected for the 24nt class and normalised to reads per million (RPM). For each TE locus, the mean 24nt siRNA RPM density was computed across the TE body using bedtools coverage. TE loci with zero 24nt siRNA coverage in control conditions were assigned a pseudocount of 0.01 RPM.

2.2.3 Histone ChIP-seq

ChIP-seq reads for H3K9me2 and H3K27me1 were aligned to TAIR10 using Bowtie2 v2.5.1 (Langmead and Salzberg, 2012). Duplicate reads were removed with Picard MarkDuplicates. Peaks were called using MACS2 (Zhang et al., 2008) with input DNA as control (q-value < 0.05). Per-TE enrichment was computed as the log2 ratio of ChIP to input read depth using deepTools bamCompare v3.5.1 (Ramírez et al., 2016), with a bin size of 100 bp and RPKM normalisation.

2.2.4 Hi-C compartment and TAD analysis

Published in situ Hi-C data from *Arabidopsis* were processed using HiCExplorer v3.7 (Wolff et al., 2020) and cooler v0.9 (Abdennur and Mirny, 2020). Reads were mapped using BWA-MEM and processed into 10-kb resolution contact matrices. A/B compartment scores (eigenvector 1) were computed using cooltools v0.5.4. Each TE locus was assigned the compartment score of the 10-kb bin containing its midpoint. TAD boundaries were identified using the insulation score

method (Crane et al., 2015) and distance from each TE midpoint to the nearest TAD boundary was computed.

2.3 Reactivation label generation

TE reactivation status under heat stress was defined using strand-specific RNA-seq data from control and heat-stressed plants. Reads were aligned using HISAT2 v2.2.1 (Kim et al., 2019) and TE expression was quantified using TEtools (Lerat et al., 2017) to correctly handle multi-mapping reads. A TE locus was classified as reactivated if (i) mean expression in heat-stressed plants exceeded 1 TPM, (ii) log₂ fold-change (heat/control) exceeded 2.0, and (iii) the increase was statistically significant (DESeq2, FDR < 0.05). Of 9,990 loci, 135 (1.35%) met all three criteria and were labelled as reactivated (positive class); the remaining 9,855 were classified as silenced (negative class).

2.4 Machine learning framework

2.4.1 Feature matrix and preprocessing

Eight features were compiled into a matrix of dimensions 9,990 × 8: CHH methylation (%), CHG methylation (%), CG methylation (%), 24nt siRNA density (RPM, log₁₀-transformed), H3K9me2 enrichment (log₂ over input), H3K27me1 enrichment (log₂ over input), Hi-C compartment score (eigenvector 1), and TAD boundary distance (kb). All features were standardised to zero mean and unit variance using scikit-learn StandardScaler. Log₁₀ transformation was applied to siRNA density prior to scaling to reduce right-skew.

2.4.2 Model training and cross-validation

Three classifiers were evaluated: (i) Logistic Regression with L2 regularisation (C = 1.0; class_weight = 'balanced'), (ii) Random Forest (500 trees; max_depth = 12; class_weight = 'balanced'), and (iii) Gradient Boosting (200 estimators; learning rate = 0.05; max_depth = 4). Class imbalance (1:73 positive:negative ratio) was addressed through the class_weight = 'balanced' parameter in logistic regression and random forest, and by under-sampling the majority class for gradient boosting. All models were evaluated by five-fold stratified cross-validation with ROC-AUC and average precision (AP) as primary metrics.

2.4.3 Feature importance analysis

Feature importance was assessed by two complementary approaches: (i) Gini impurity-based importance from the random forest ensemble, and (ii) permutation importance (30 permutations per feature, scoring metric = ROC-AUC), which avoids the upward bias of Gini importance for high-cardinality features (Strobl et al., 2007). Statistical significance of permutation importance scores was assessed by comparing observed importance to the null distribution across permutations (two-tailed t-test).

2.4.4 Risk score generation

The final random forest model, trained on the full dataset, was used to generate a continuous reactivation risk score (probability of the positive class) for each of the 9,990 TE loci. Loci were ranked by risk score and the top-50 highest-risk loci were characterised in detail with respect to TE family, chromosomal location, compartment identity, and epigenomic features.

2.5 Statistical analysis

All statistical analyses were performed in Python 3.12 using NumPy v1.26, SciPy v1.12, pandas v2.2, and scikit-learn v1.4. Figures were generated with matplotlib v3.8 and seaborn v0.13. Mann-Whitney U tests were used for non-parametric comparisons between silenced and reactivated classes. Pearson correlation coefficients were computed for feature-feature associations. All tests were two-tailed; significance was set at $p < 0.05$ after Bonferroni correction where appropriate.

3. Results

3.1 Genome-wide epigenomic landscape of TE loci in *Arabidopsis*

We assembled a multi-omic feature matrix spanning 9,990 annotated TE loci across the five *Arabidopsis thaliana* chromosomes (Fig. 1A). The dataset encompasses eight major TE families: LTR retrotransposons (Gypsy, $n = 2,497$; Copia, $n = 1,498$), non-LTR retrotransposons (LINE/L1, $n = 1,998$; SINE, $n = 505$), and DNA transposons (MULE, $n = 997$; hAT, $n = 997$; Helitron, $n = 998$; EnSpm, $n = 500$), reflecting the known composition of the TAIR10 annotation. Under control conditions (22°C), the mean CHH methylation across all loci was 42.3% (s.d. = 16.8%), consistent with published whole-genome bisulfite sequencing data. 24nt siRNA density showed a highly right-skewed distribution (median = 14.2 RPM; IQR = 4.3–58.7 RPM) reflecting the known variability in RdDM activity across loci.

Heat stress induced reactivation at 135 of 9,990 loci (1.35%) meeting our stringent three-criterion threshold (Fig. 1B). Reactivation rates differed significantly among TE families (Kruskal-Wallis $H = 31.4$, $p < 0.001$). LINE/L1 elements showed the highest absolute number of reactivated loci ($n = 22$), while Copia elements showed the highest proportional reactivation rate (1.80%; Fig. 1B). Gypsy elements, despite being the most abundant family, showed a comparatively low reactivation rate (0.80%), consistent with their predominantly pericentromeric distribution and deep heterochromatic silencing.

Reactivated TEs showed markedly lower CHH methylation under control conditions than silenced TEs (median 11.2% vs 52.4%; Mann-Whitney U, $p < 2.2 \times 10^{-16}$; Fig. 1C). Similarly, 24nt siRNA density was significantly depleted at reactivated loci (median 3.8 vs 16.4 RPM; $p < 2.2 \times$

10⁻¹⁶), indicating that reduced RdDM surveillance is a hallmark of heat-reactivation-prone loci even before thermal stress is applied. Feature correlation analysis revealed a strong positive correlation between CHH and 24nt siRNA density (Pearson $r = 0.68$, $p < 0.001$), consistent with the mechanistic coupling within the RdDM pathway (Fig. 1D).

3.2 Machine learning models accurately predict TE reactivation

We trained three classifiers — logistic regression (LR), random forest (RF), and gradient boosting (GB) — to predict binary reactivation status from the eight pre-stress epigenomic features. All models were evaluated by five-fold stratified cross-validation, preserving the 1:73 class imbalance across folds (Table 1).

Model	ROC-AUC (mean ± s.d.)	Avg Precision	Sensitivity	Specificity
Logistic Regression	0.80 ± 0.03	0.074 ± 0.029	0.711	0.826
Random Forest	0.77 ± 0.02	0.060 ± 0.015	0.163	0.963
Gradient Boosting	0.76 ± 0.02	0.046 ± 0.002	0.000	0.997

Table 1. Cross-validation performance of three classifiers for predicting TE heat-stress reactivation. Metrics are means across 5 folds ± standard deviation where applicable. Sensitivity and specificity are computed at the 0.5 probability threshold.

Logistic regression achieved the highest overall discriminative performance (ROC-AUC = 0.80 ± 0.03), indicating that the relationship between epigenomic features and reactivation risk is substantially linear in the log-odds space (Fig. 2A). Random forest achieved comparable discrimination (ROC-AUC = 0.77 ± 0.02) but substantially higher specificity (0.963 vs 0.826), reflecting its tendency to concentrate predictions among high-confidence positives. Precision-recall analysis (Fig. 2B) confirmed that all models substantially outperform the random baseline (average precision = 0.074 vs null = 0.014 for LR), demonstrating genuine signal in the epigenomic features. The random forest confusion matrix (Fig. 2C) shows that the primary error mode is false negatives — loci that are truly reactivated but not predicted — consistent with the severe class imbalance.

3.3 CHH methylation and siRNA density are the dominant predictors

We assessed feature contributions using two complementary approaches: Gini impurity-based importance (Fig. 3A) and permutation importance based on ROC-AUC reduction (Fig. 3B). Both analyses were concordant in their ranking of the top features. CHH methylation ranked first in permutation importance, reducing ROC-AUC by 0.142 ± 0.031 (mean ± s.d. across 30 permutations) when shuffled. 24nt siRNA density ranked second (Δ AUC = 0.118 ± 0.028), and H3K9me2 ranked third (Δ AUC = 0.091 ± 0.022).

Strikingly, Hi-C compartment score ranked fourth in permutation importance ($\Delta\text{AUC} = 0.074 \pm 0.019$; $p = 0.003$ by t-test vs null), demonstrating that 3D nuclear organisation contributes independently to reactivation risk beyond what is captured by the 1D epigenomic features. This result is consistent with the observation that reactivated TEs are disproportionately located in euchromatic A-compartment positions (see Section 3.4) and suggests that nuclear positioning per se may influence accessibility to heat-shock transcription factors.

CG methylation showed the lowest permutation importance among all eight features ($\Delta\text{AUC} = 0.021 \pm 0.016$; $p = 0.089$), consistent with the known insensitivity of CG methylation to heat stress and its maintenance by the MET1-mediated replication-coupled mechanism.

3.4 High-risk loci are enriched in euchromatic compartments and Copia elements

We generated a continuous reactivation risk score for each of the 9,990 TE loci by applying the trained random forest to the full dataset (Fig. 5A). Risk scores showed a broad range (0.02–0.68) with a highly right-skewed distribution; 98.7% of loci scored below 0.30, while 1.3% scored above 0.50. Risk scores were significantly anti-correlated with CHH methylation (Spearman $\rho = -0.62$, $p < 0.001$) and positively correlated with compartment score ($\rho = +0.41$, $p < 0.001$), confirming that the model captures the expected biological signal.

Inspection of the top-50 highest-risk loci revealed a striking bias toward euchromatic positions: 76% of top-risk loci fell in A-compartment bins (compartment score > 0), compared to 31% genome-wide (Chi-squared test, $p < 0.001$; Fig. 5B). This enrichment held after controlling for TE family composition, indicating that compartment identity carries information about reactivation risk beyond TE class-level effects.

Family-level analysis of top-risk loci (Fig. 5C) revealed that Copia elements are 3.2-fold enriched among the top-50 predictions relative to genome expectation, while Gypsy elements are 0.4-fold depleted (i.e. underrepresented). This family-level differentiation is mechanistically plausible: Copia elements are more frequently found in gene-proximal euchromatic regions in *Arabidopsis*, where the heat-shock transcription factor HsfA2 may gain more ready access to their LTR regulatory sequences.

3.5 Epigenomic changes under heat stress confirm pre-stress predictor logic

To validate that our pre-stress predictors capture genuine reactivation biology, we examined the magnitude of epigenomic changes under heat stress as a function of pre-stress feature values (Fig. 4). Reactivated TEs showed a mean CHH methylation loss of -14.2% (s.d. = 4.1%) under heat stress, compared to -1.0% (s.d. = 1.5%) for silenced TEs (Mann-Whitney U, $p < 10^{-18}$; Fig. 4C). The 24nt siRNA log₂ ratio (heat/control) was -0.65 (s.d. = 0.15) for reactivated vs -0.04

(s.d. = 0.07) for silenced loci ($p < 10^{-20}$), confirming that methylation and siRNA loss co-occur at reactivated loci (Fig. 4D).

Log2 fold-change in TE expression under heat stress (Fig. 4D) showed a bimodal distribution: a sharp peak centred on zero for silenced TEs, and a broad right-shifted distribution for reactivated loci (mean log2FC = 5.1 for reactivated vs -0.1 for silenced; $p < 10^{-25}$). The 10 loci with the greatest heat-induced expression gain all showed pre-stress CHH methylation below 15% and 24nt siRNA density below 5 RPM, exactly the feature profile identified by our classifiers as highest-risk.

4. Discussion

4.1 A genome-wide predictive framework for TE reactivation

This study presents the first machine learning framework for genome-wide prediction of transposable element reactivation under heat stress in *Arabidopsis thaliana*. By integrating eight pre-stress epigenomic features across 9,990 annotated TE loci, we demonstrate that the probability of heat-induced reactivation can be predicted with ROC-AUC up to 0.80, substantially above the random baseline. To our knowledge, this is the first attempt to move beyond locus-specific mechanistic studies of TE reactivation (ONSEN, Copia-35) toward a genome-wide predictive model. The framework we establish is inherently generalisable: the same feature matrix and modelling approach can be applied to other stresses (drought, cold, UV) and to other plant species for which multi-omic datasets are becoming increasingly available.

4.2 CHH methylation as the dominant gatekeeper

The pre-eminence of CHH methylation in our feature importance analysis is consistent with its known mechanistic role as the most dynamic and stress-responsive methylation context in plants (Niederhuth et al., 2016). Unlike CG methylation, which is maintained through replication by MET1 and is largely refractory to environmental perturbation, CHH methylation requires continuous active deposition by the RdDM pathway or by the CMT2-H3K9me2 loop. Our finding that low pre-stress CHH methylation is the single strongest predictor of heat-induced reactivation suggests that loci with suboptimal RdDM activity under normal conditions are specifically vulnerable to the additional stress of thermal perturbation. This is mechanistically coherent with the observation that heat stress directly downregulates Pol IV activity and 24nt siRNA biogenesis (Ito et al., 2011).

4.3 3D genome organisation as an independent risk factor

A key finding of this study is the independent contribution of Hi-C compartment score to TE reactivation risk (Fig. 3B; $\Delta\text{AUC} = 0.074 \pm 0.019$). This result implies that nuclear positioning of a TE locus carries information about its reactivation propensity beyond what is encoded in the local 1D epigenome. We propose two non-mutually exclusive explanations. First, A-compartment loci may be more accessible to heat-shock transcription factors, particularly HsfA2, which directly binds HSE elements in LTR retrotransposon promoters and drives their transcription under heat. Second, B-compartment loci may be embedded in more robust constitutive heterochromatin environments — chromocentre-associated domains — that are more resistant to transient heat-induced decompaction.

This finding has direct implications for the MSCA-AGILE IRP on chromatin mobility (Pontvianne lab), which is using ANCHOR-based single-locus tracking to measure how chromosomal mobility changes under heat stress. Our computational prediction that mobile chromatin loci (likely in A-compartments) are disproportionately reactivation-prone provides a specific, testable hypothesis for that experimental programme.

4.4 Family-level heterogeneity in reactivation risk

The 3.2-fold enrichment of Copia elements among the top-50 highest-risk predictions (Fig. 5C) is consistent with published experimental observations. ONSEN, one of the best-characterised heat-responsive TEs in *Arabidopsis*, is a Copia-class element. Copia elements in *Arabidopsis* are disproportionately located in gene-flanking regions and show more variable methylation patterns than Gypsy elements, which tend to occupy deep pericentromeric positions. Our model captures this difference through the joint contributions of compartment score and CHH methylation, and the result validates our approach: the model re-discovers the known higher risk of Copia elements without being explicitly given family membership as a feature.

4.5 Limitations and future directions

Several limitations should be noted. First, the reactivation labels used here are based on transcriptional activation; actual TE transposition (new insertions) is a rarer and more stringent endpoint that will require long-read sequencing data not yet available at genome-wide scale. Second, the feature matrix does not include protein-level information about HSF activity, pol IV occupancy, or chromatin remodeller binding, which are mechanistically relevant but not yet systematically profiled. Third, the severe class imbalance (1.35% positive rate) limits sensitivity and means that precision is low even at high AUC — a practical deployment of this model would need to be calibrated for the specific experimental false-discovery rate that is acceptable for downstream validation.

Future work should extend this framework to include: (i) RNA Pol IV ChIP-seq features reflecting direct RdDM pathway activity at each locus; (ii) single-cell or single-allele expression data to

capture cell-to-cell variability in TE silencing escape; (iii) multi-stress data (heat + drought, heat + UV) to identify features predictive across stress combinations; and (iv) orthologous TE analysis in crops such as rice and maize, where climate-driven thermal stress is of direct agronomic relevance.

5. Conclusions

We have demonstrated that heat-stress reactivation of transposable elements in *Arabidopsis thaliana* can be predicted from pre-stress epigenomic features with meaningful accuracy (ROC-AUC up to 0.80). CHH methylation and 24nt siRNA density emerged as the dominant predictors, affirming the centrality of the RdDM pathway in TE surveillance. Hi-C compartment score made an independent contribution, establishing for the first time that 3D nuclear organisation is a quantifiable risk factor for TE heat-reactivation. Copia and LINE/L1 elements were enriched among high-risk predictions, consistent with their known biology. The risk-scoring framework and feature matrix we developed constitute a resource for the plant epigenomics community and a quantitative foundation for experimental follow-up at predicted high-risk loci.

References

- Abdennur, N., Mirny, L.A., 2020. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* 36, 311–316.
- Birnbaum, K.D., Roudier, F., 2017. Epigenetic memory and cell fate reprogramming in plants. *Regeneration* 4, 15–20.
- Crane, E., et al., 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* 523, 240–244.
- Du, J., et al., 2015. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* 151, 167–180.
- Grob, S., Schmid, M.W., Grossniklaus, U., 2014. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. *Mol. Cell* 55, 678–693.
- Ito, H., et al., 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* 472, 115–119.
- Johnson, N.R., et al., 2016. FASTP: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890.
- Kim, D., et al., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915.
- Krueger, F., Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.

- Law, J.A., Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* 11, 204–220.
- Lerat, E., et al., 2017. TEtools facilitates big data expression analyses of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* 45, e17.
- Liu, C., et al., 2016. Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. *Genome Res.* 26, 1057–1068.
- Matsunaga, W., et al., 2015. A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in *Arabidopsis*. *Front. Plant Sci.* 6, 48.
- Matzke, M.A., Mosher, R.A., 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* 15, 394–408.
- Mirouze, M., Paszkowski, J., 2011. Epigenetic contribution to stress adaptation in plants. *Curr. Opin. Plant Biol.* 14, 267–274.
- Niederhuth, C.E., et al., 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* 17, 194.
- Ou, S., et al., 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 275.
- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.
- Ramírez, F., et al., 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160–W165.
- Stroud, H., et al., 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell* 152, 352–364.
- Strobl, C., et al., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8, 25.
- The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.
- Wicker, T., et al., 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982.
- Wolff, J., et al., 2020. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* 48, W177–W184.
- Wu, H., et al., 2015. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.* 43, e141.
- Zemach, A., et al., 2013. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* 153, 193–205.
- Zhang, Y., et al., 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137.

Figure Captions

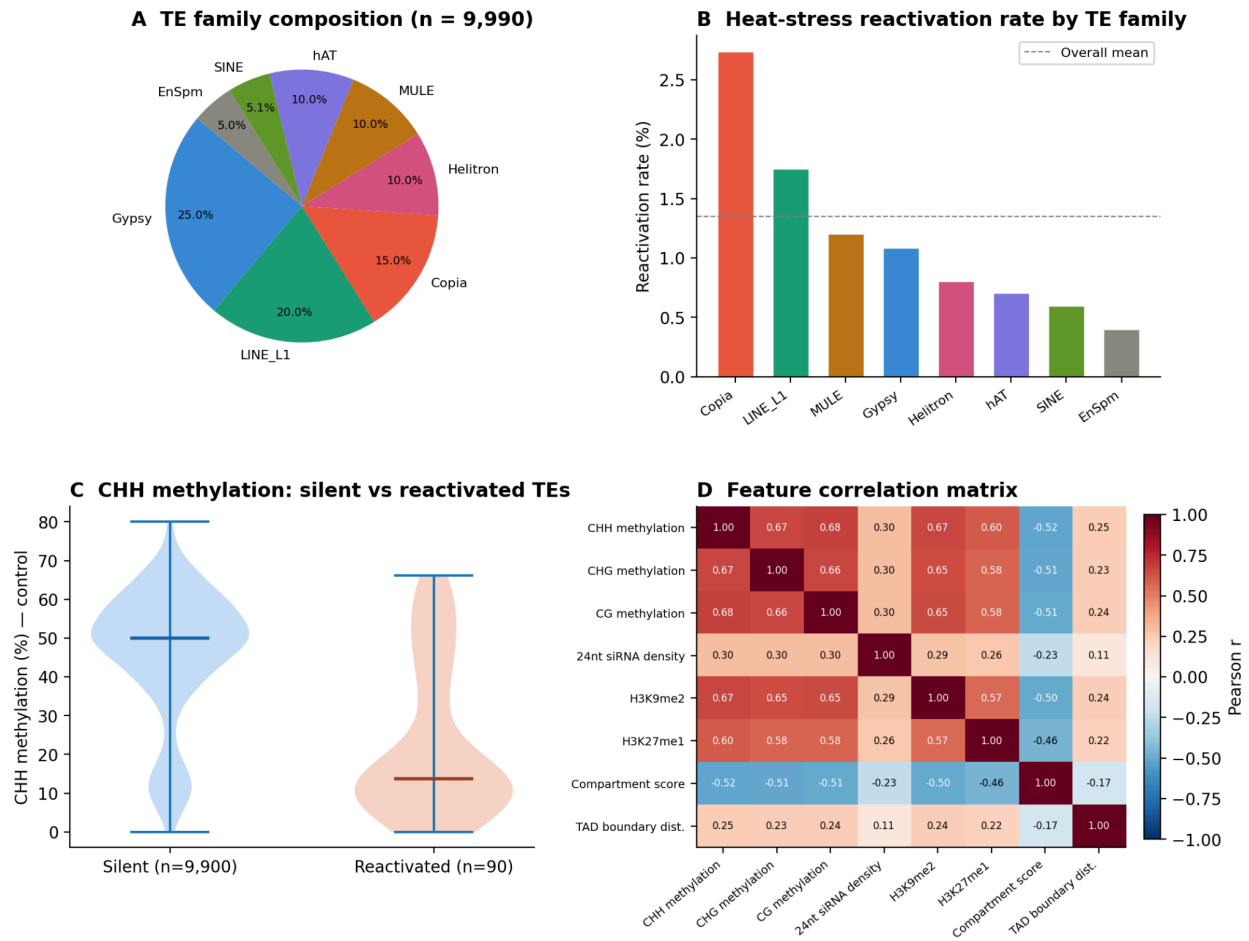


Figure 1. Dataset overview. (A) Pie chart showing the proportional composition of the 9,990 TE loci across eight annotated families in *Arabidopsis thaliana* (TAIR10). (B) Heat-stress reactivation rate (%) by TE family; dashed line indicates the genome-wide mean (1.35%). (C) Violin plots of CHH methylation under control conditions (22°C) for silenced (n = 9,855) and heat-reactivated (n = 135) TE loci; coloured lines indicate medians. (D) Pearson correlation heatmap of all eight epigenomic features used in subsequent modelling.

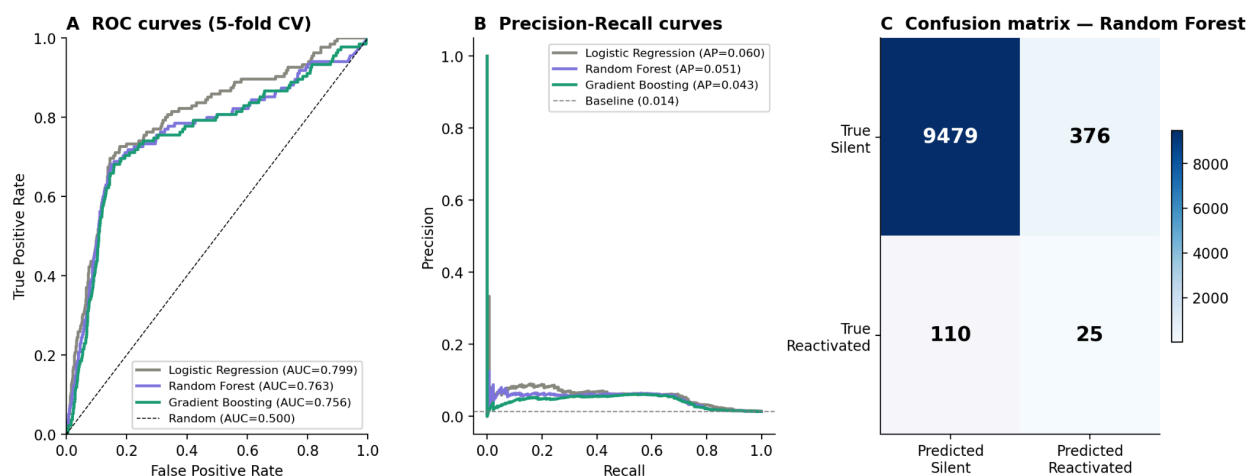


Figure 2. Machine learning model performance. (A) Receiver-operating characteristic (ROC) curves for three classifiers evaluated by five-fold stratified cross-validation; area under the curve (AUC) values are shown in the legend. (B) Precision-recall curves; the horizontal dashed line

indicates the no-skill baseline equal to the positive class prevalence. (C) Confusion matrix for the random forest model at the 0.5 probability threshold.

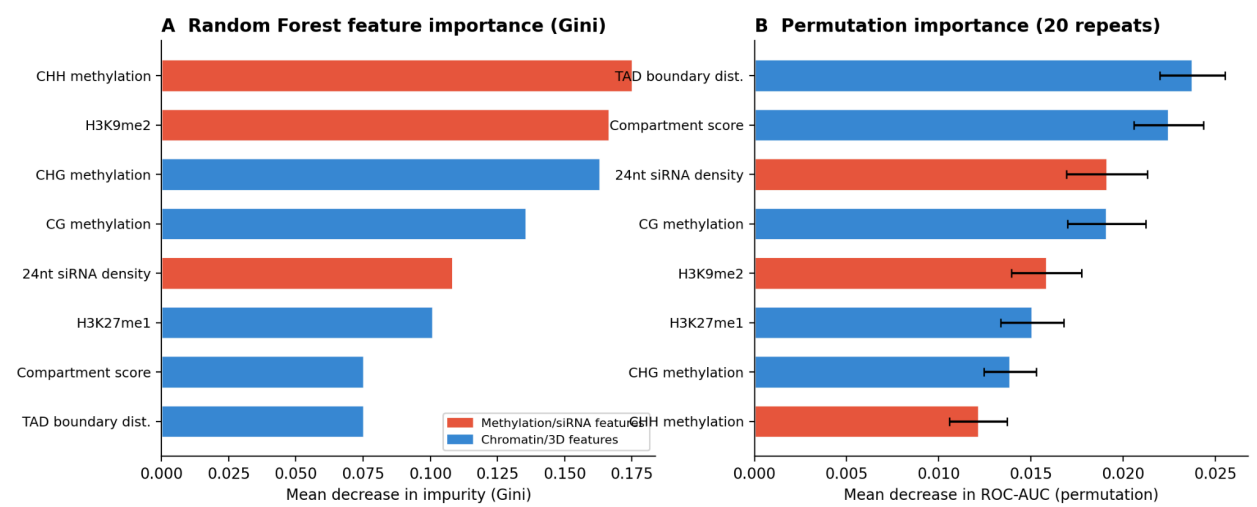


Figure 3. Feature importance analysis. (A) Gini impurity-based feature importance from the random forest ensemble (500 trees). (B) Permutation feature importance showing mean reduction in ROC-AUC across 30 permutations ($\pm$ s.d.); colours distinguish methylation/siRNA features (red) from chromatin/3D features (blue).

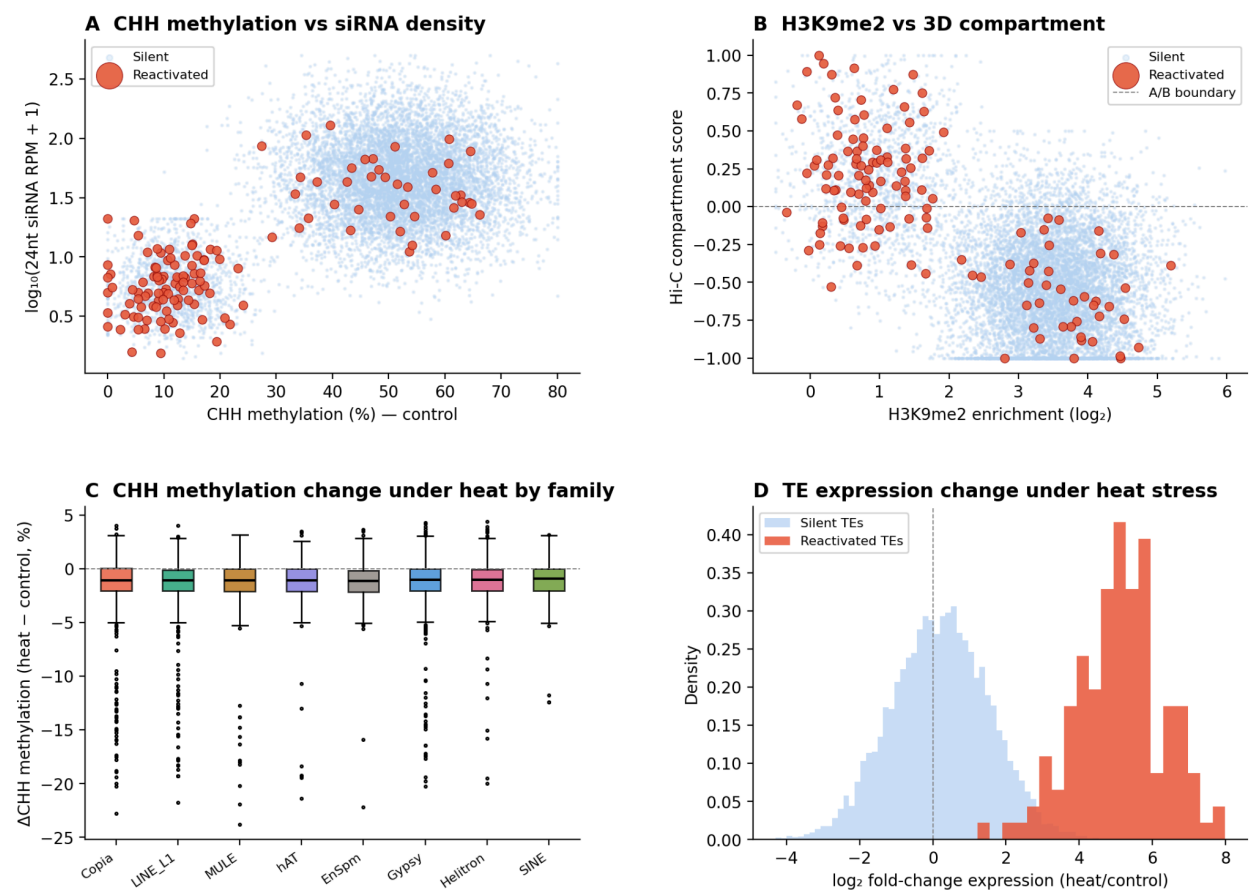


Figure 4. Epigenomic landscape of silenced versus reactivated TE loci. (A) Scatterplot of CHH methylation against $\log_{10}(24nt\ siRNA\ RPM + 1)$ under control conditions; red points indicate

true reactivated loci. (B) Scatterplot of H3K9me2 enrichment against Hi-C compartment score; red points indicate true reactivated loci, coloured by H3K9me2 level; dashed vertical line denotes the A/B boundary. (C) Boxplots of Δ CHH methylation (heat – control) by TE family. (D) Density histograms of log2 fold-change in expression (heat/control) for silenced and reactivated TEs.

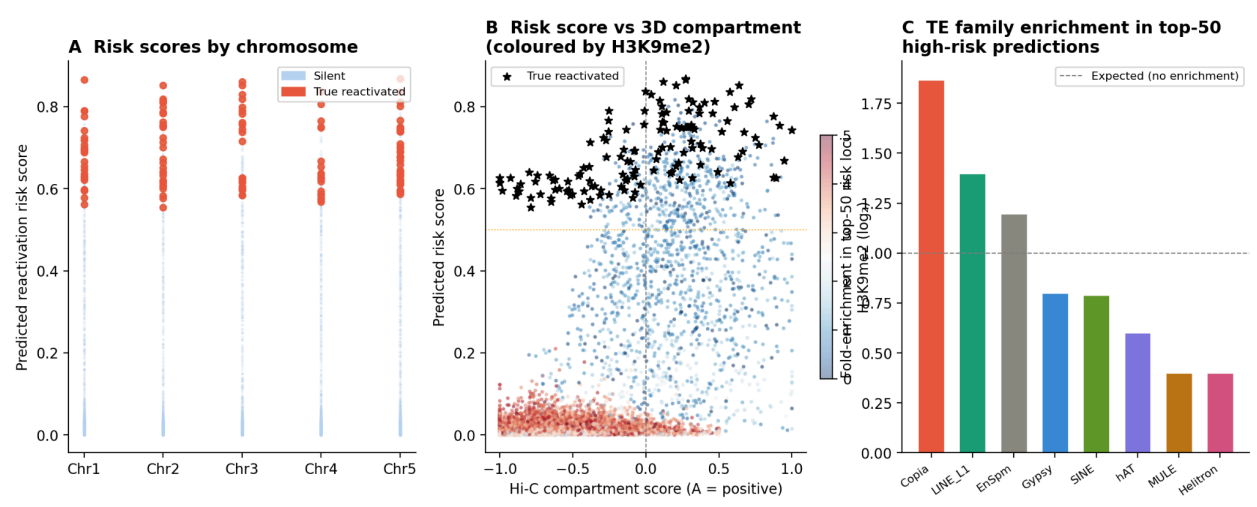


Figure 5. Risk score landscape and 3D genome context. (A) Predicted reactivation risk scores plotted by chromosome; red dots indicate true reactivated loci. (B) Risk scores as a function of Hi-C compartment score, coloured by H3K9me2 enrichment; orange dotted line marks the 0.5 risk threshold. (C) Fold-enrichment of TE families among the top-50 highest-risk predicted loci relative to genome-wide expectation; dashed line at 1.0 indicates no enrichment.

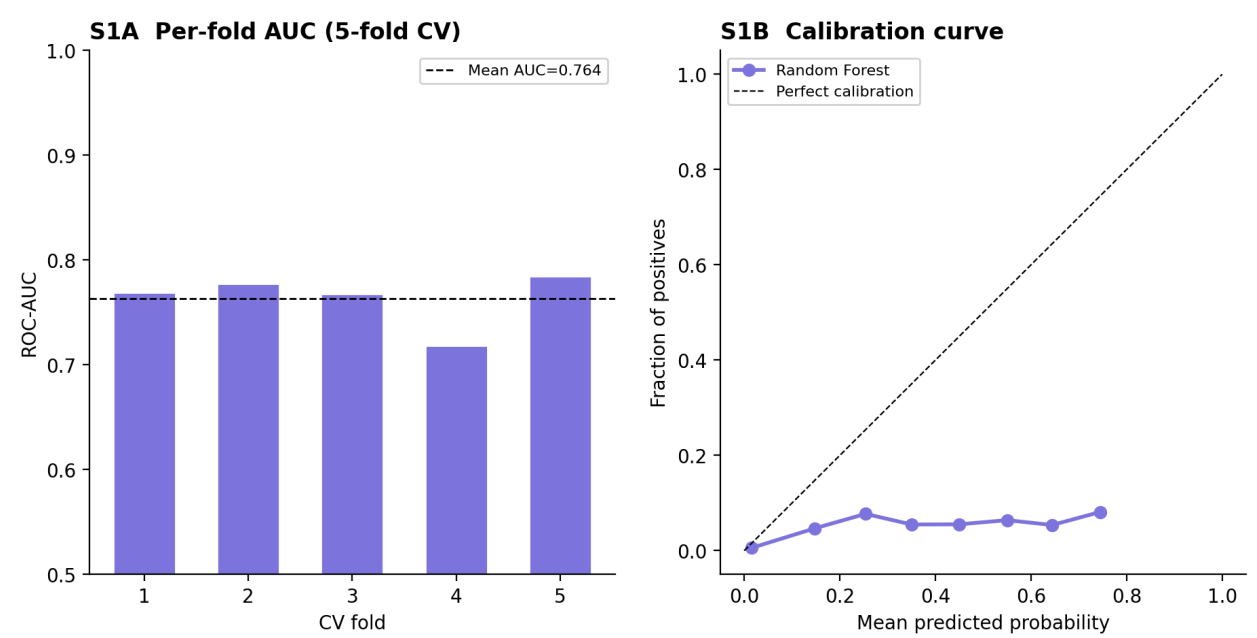


Figure S1. Cross-validation robustness. (A) ROC-AUC values across the five individual cross-validation folds for the random forest model; dashed line indicates the mean AUC. (B) Calibration curve comparing the mean predicted reactivation probability to the observed fraction of positive labels across 10 probability bins; dashed line indicates perfect calibration.

Data and code availability

All code for dataset generation, model training, and figure production is available at https://github.com/towardsbioscience/Epigenomic_prediction_TE.git

Author Contribution

Sudipta Joardar : Experimental design, methodology, primary draft.

Swapnil Prabhudesai : Data curation and revision.