Supplementary methods

To characterize the nature of our sequencing data (GEO GSE2182; (1)) we studied the distribution of the coverage using a kernel density estimation (KDE), also termed the *Parzen–Rosenblatt window* method (2,3), with a gaussian kernel. First, we assigned to each nucleotide of the genome its corresponding coverage. Then, we divided the genome in windows with different bin size from 100 nt to 10,000 nt and calculated the coverage normalized against the size of the window, thus coverage level do not depend on window size but on read enrichment. After that, we calculated the kernel density estimation. Supplementary Figure S1 shows that KDE of the coverage using small window bin size (100 or 300 nt), reflects mostly genomic sites not enriched in the BrdU-reads, meaning that bin size is not large enough to discriminate the BrdU-enriched sites. Increasing the window bin size (700 or 900 nt) allows distinguishing the regions enriched in BrdU-reads from those with a low coverage as a crest in the density distribution. The largest windows (5,000 or 10,000 nt) do not detect any zones enriched in BrdU-reads, suggesting that the sites with more BrdU-reads are smaller than the window bin size. These results point to a distribution of the reads around peaks of 700-900 nt and ruled out the possibility of frequent large zones enriched in the sequenced reads.

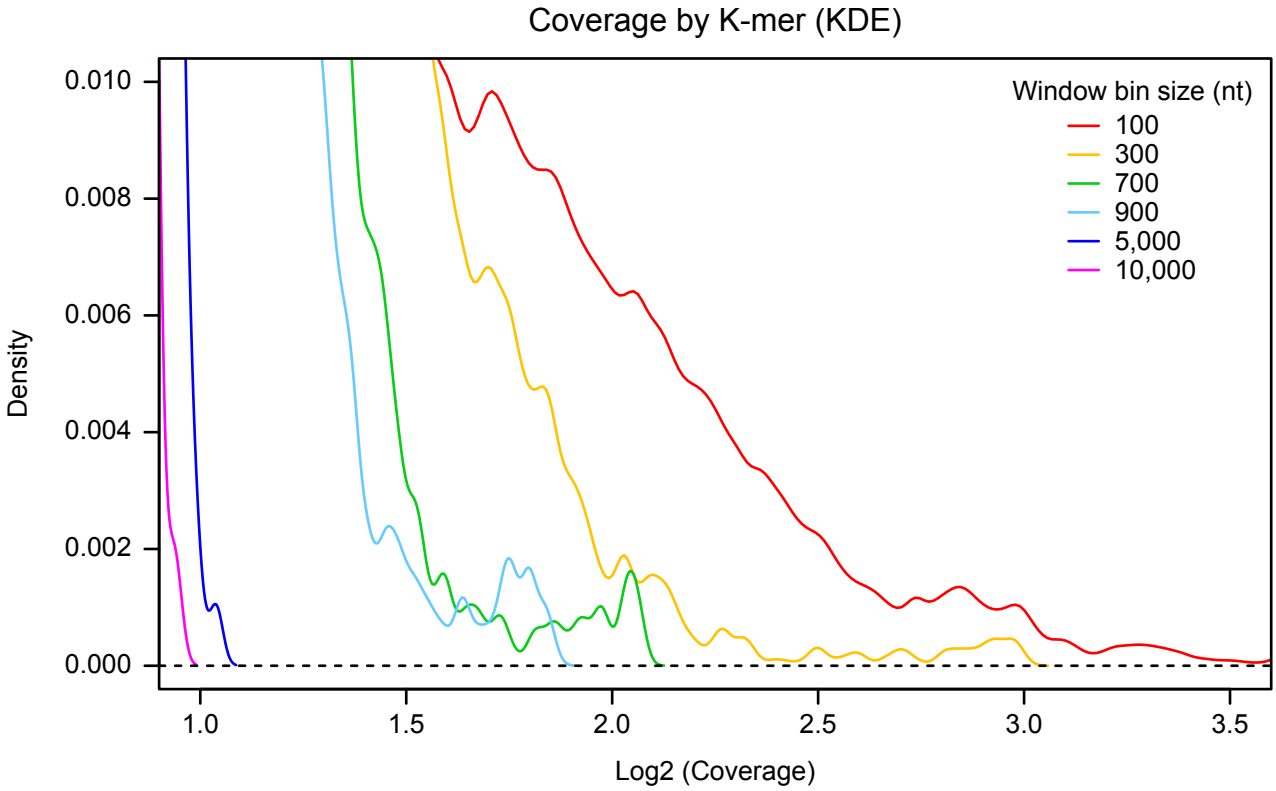
There is a large variety of peak-calling algorithms designed to identify peaks out of sequencing reads. They use different parameters, algorithms and normalizations, but the results differ basically in the number and width of the peaks returned. The aim of the present study was to identify ORIs in the heterochromatic regions of the Arabidopsis genome for which we did not have any previous information on their number and nature, although our KDE analysis suggested that most peaks should not be very large in size. In any case, we used and compared several peak-calling algorithms developed to identify sharp or broader peaks, namely: MACS (versions 1.4 and 2.0; (4)), BayesPeak (5), Homer (6), SICER (7) and T-PIC (8). We found that the peak size distribution in each case showed, at the most, a 2-4-fold difference in the mean value (Supplementary Fig. S2A), which ranged between 537 nt in the case of MACS2.0 to 1232 nt for T-PIC. This indicated that the peak size does not change dramatically with different algorithms and is relatively small, consistent with the compact organization and small size of the Arabidopsis genome. Interestingly, the median of the medians for the six peak-calling algorithms evaluated corresponds to the window bin size of the KDE in which a crest of the density distribution was observed (700-900 nt). Thus, the vast majority of ORIs corresponds to discrete initiation sites rather than to large initiation zones. Furthermore, independently of the algorithm used, the peak size in peri- and non-pericentromeric regions was very similar, in each case (Supplementary Figure S2A). Measurement of the interorigin distance using different algorithms (Supplementary Figure S2B) did not reveal large differences. The total number of peaks (origins) retrieved in each case is different: SICER and T-PICs returned the largest values. A pair-wise comparison

between the algorithms showed that MACS, BayesPeak and Homer rendered quite similar results as revealed by the percentage of coincidence (Supplementary Figure S2C). The summary of the total genomic regions covered by peaks (Supplementary Table S1) shows that SICER returned a high number of peaks and T-PIC of larger size, which cover a very large genomic percentage. This is unlikely to be the case. Measurement of SNS enrichment by PCR of various peaks is consistent with peaks of a smaller size, also in TEs.

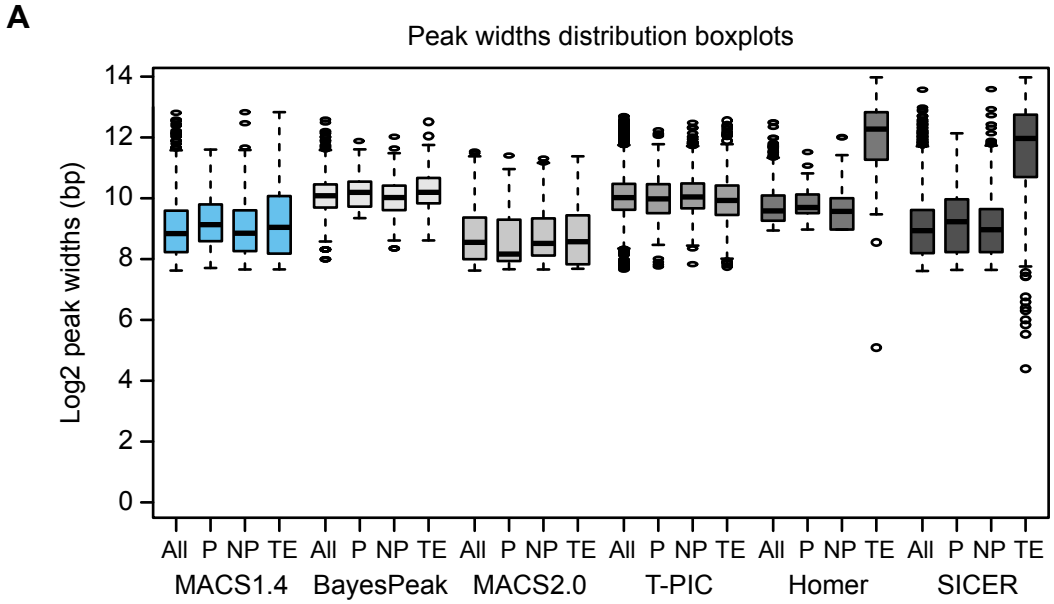
These results led us to conclude that most peaks in Arabidopsis are relatively small, that is origins do not correspond with large initiation zones, which is consistent with the compact organization and small size of its genome. This together with the visual inspection of the original reads in the IGB viewer across all chromosomes, led us to conclude that MACS1.4 provided the most realistic identification of origins, in the case of Arabidopsis.

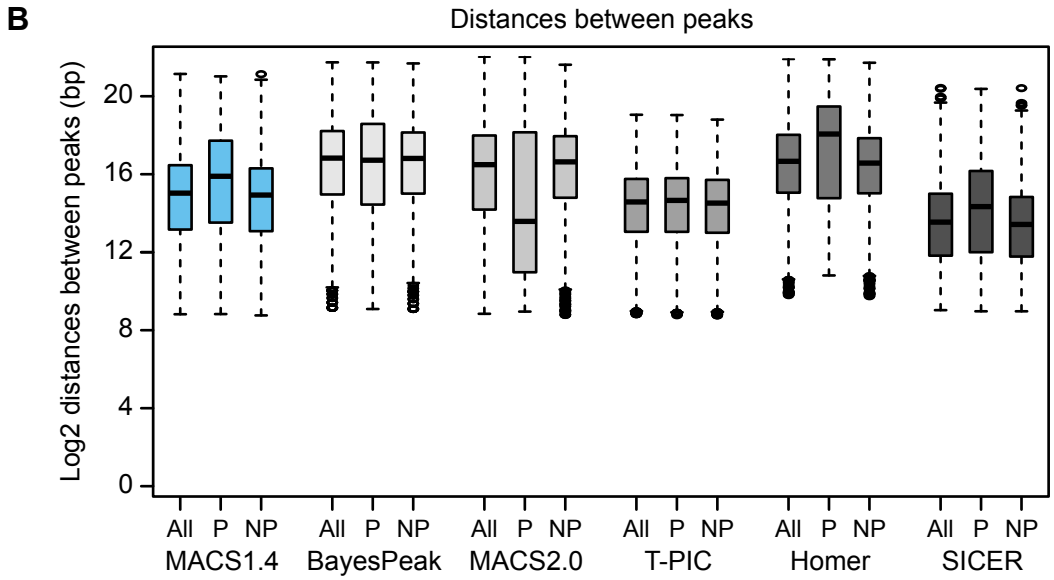
References

1. Costas, C., de la Paz Sanchez, M., Stroud, H., Yu, Y., Oliveros, J.C., Feng, S., Benguria, A., Lopez-Vidriero, I., Zhang, X., Solano, R. *et al.* (2011) Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat Struc Mol Biol*, **18**, 395-400.
2. Rosenblatt, M. (1956) Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.*, **27**, 832.
3. Parzen, E. (1962) On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.*, **33**, 1065.
4. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.
5. Cairns, J., Spyrou, C., Stark, R., Smith, M.L., Lynch, A.G. and Tavare, S. (2011) BayesPeak--an R package for analysing ChIP-seq data. *Bioinformatics*, **27**, 713-714.
6. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576-589.
7. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K. and Peng, W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952-1958.
8. Hower, V., Evans, S.N. and Pachter, L. (2011) Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics*, **12**, 15.



Supplementary Figure S1. Distribution of the coverage using a kernel density estimator (KDE).

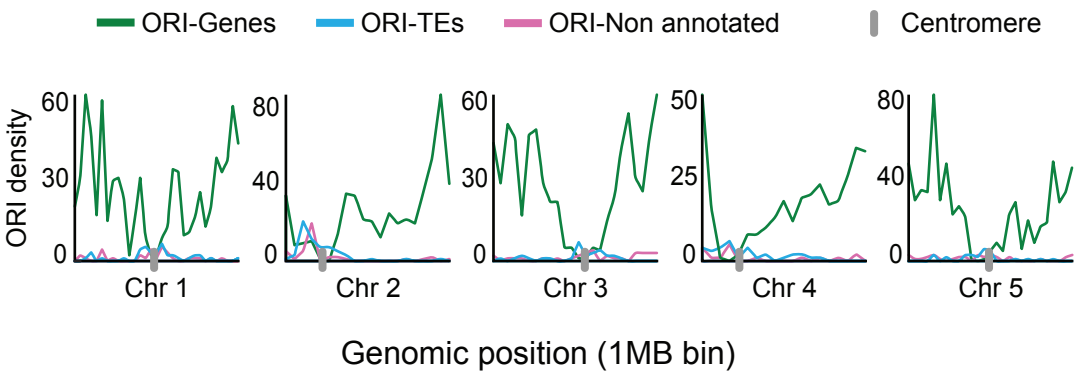




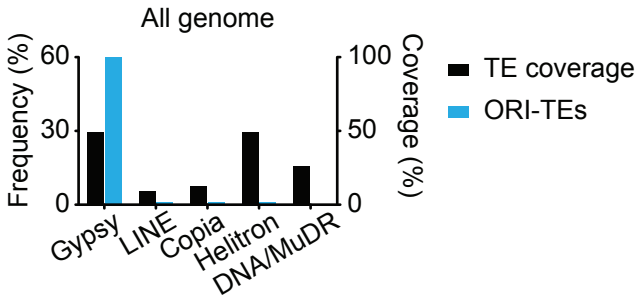
C

	BayesPeak	Homer	MACS1.4	MACS2.0	SICER	T-PIC
BayesPeak	1131 (100/100)	653 (57.7/48.6)	1036 (91.6/77.1)	756 (66.8/56.2)	743 (65.7/8.8)	1040 (91.9/17.8)
Homer		1344 (100/100)	1074 (79.9/33.2)	839 (62.4/64.1)	643 (47.8/7.6)	1141 (84.9/19.5)
MACS1.4			3230 (100/100)	1182 (35.6/90.4)	1107 (34.3/13.1)	2467 (76.4/42.3)
MACS2.0				1308 (100/100)	670 (51.2/7.9)	1245 (95.2/21.3)
SICER					8442 (100/100)	1309 (15.5/22.4)
T-PIC						5836 (100/100)

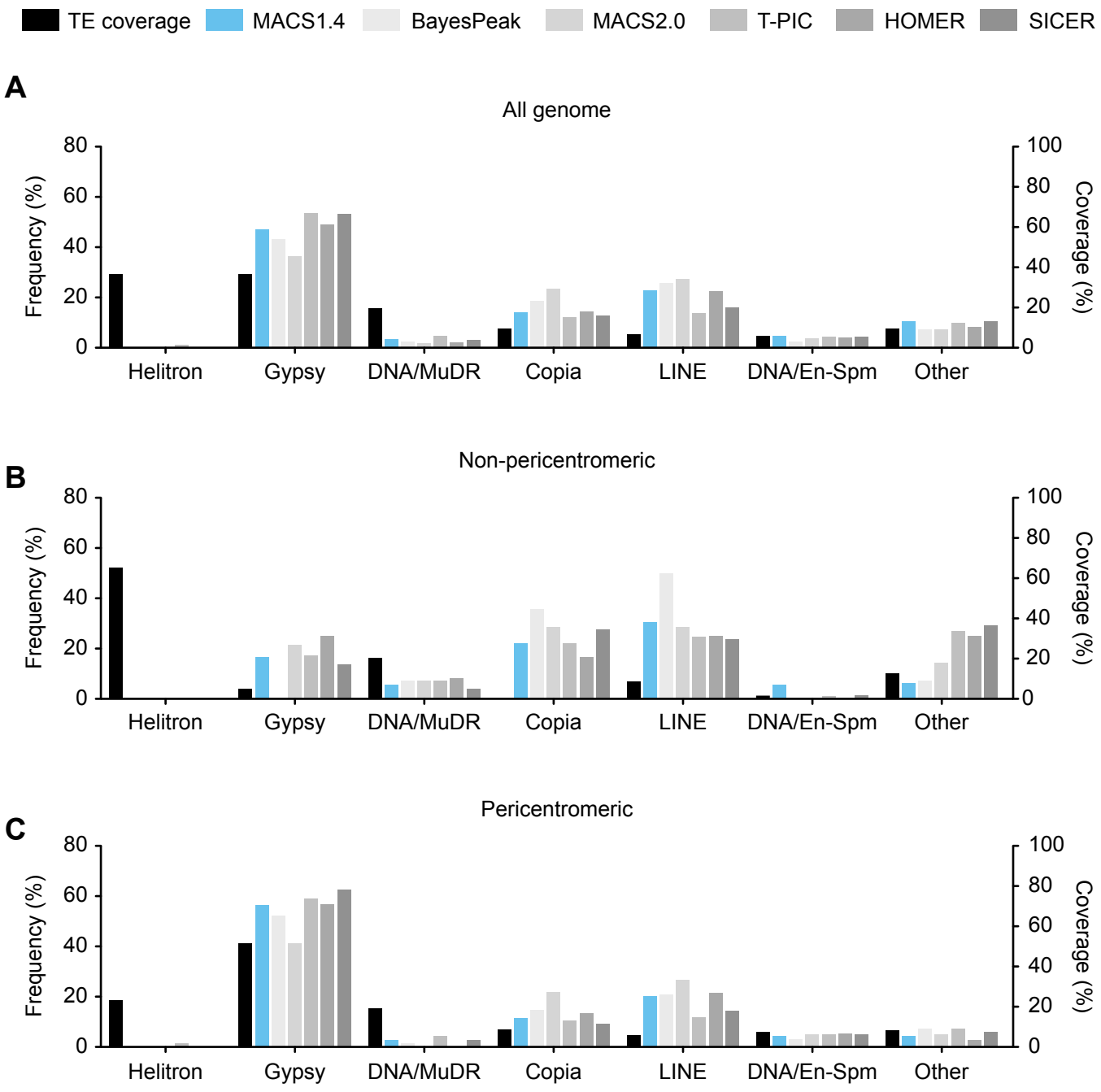
Supplementary Figure S2. Comaprison of various peak-calling algorithms.
A. Distributions of peak width.
B. Dsitributions of interpeak distances.
C. Pair-wise comparison of peaks returned by all the algorithms used.
All = all the genome. P = pericentromeric regions. NP = non-pericentromeric regions.
TE = transposable elements



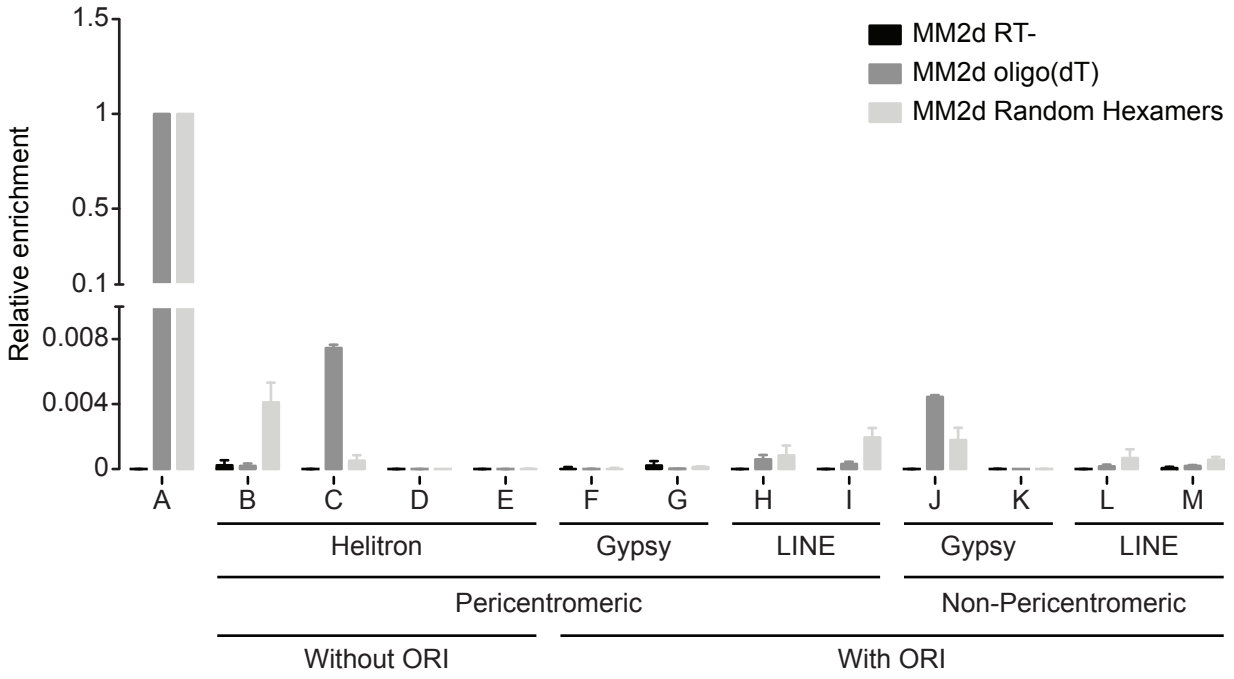
Supplementary Figure S3. Distribution of ORIs across the five Arabidopsis chromosomes. Number of ORIs (per 1 Mb bin) in genes (green), TEs (blue) and non-annotated regions (pink), as indicated, in each five Arabidopsis chromosomes. Grey bar depicts centromere localization.



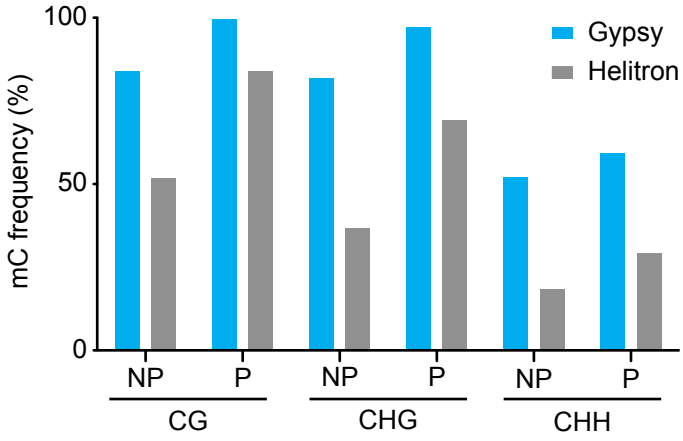
Supplementary Figure S4. Frequency distribution of ORI-TEs in TE families calculated using the multihit sequence reads (blue bars) with the corresponding TE coverage (black bars).



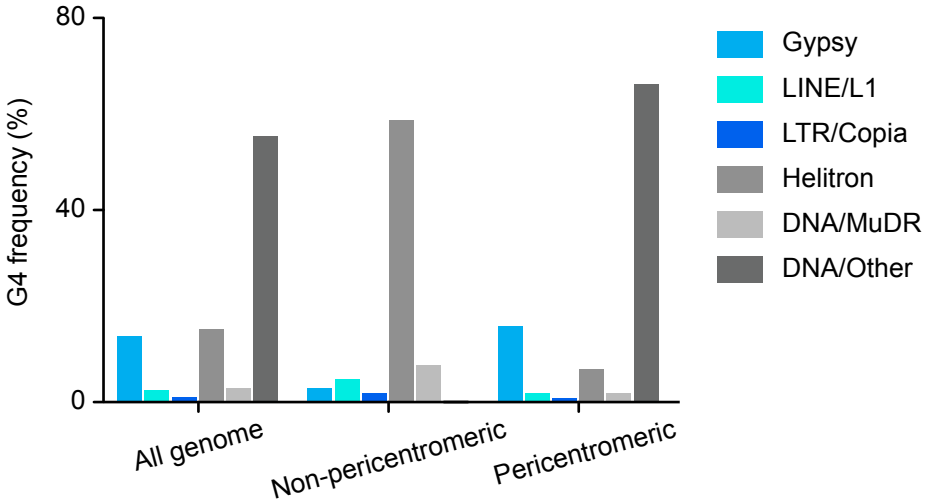
Supplementary Figure S5. Frequency distribution of ORI-TEs (blue bars) in TE families after using different peak-calling algorithms (grey bars) for all the Arabidopsis genome (A), non-pericentromeric regions (B) and pericentromeric regions (blue bars). The respective nucleotide coverage of each family is also shown (black bars).



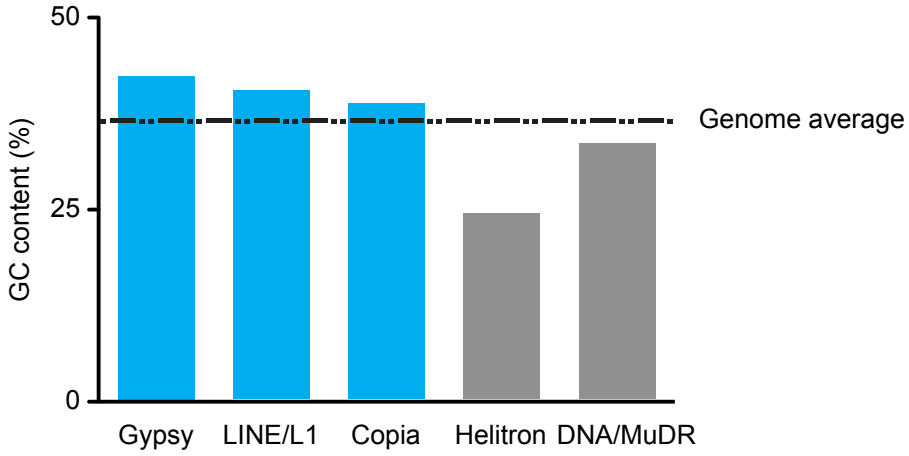
Supplementary Figure S6. RNA levels in Arabidopsis MM2d cultured cells across TEs representative of various families and with or without ORIs. Enrichment values were calculated as described in Methods and normalized against the housekeeping gene *GAPC-2* (*GLICERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C-2*). Two biological replicates were evaluated. The mean values and standard deviation is plotted. The codes for the primer pairs used to identify each TE, according to the list in Supplementary Table 2, are: A, GAPDC2 At1g13440; B - AT2TE13970; C - AT2TE16335; D - AT4TE16735; E - AT4TE17050; F - AT4TE16725-2; G - AT4TE16725-3; H - AT2TE15565-2; I - AT2TE15565-3; J - AT1TE62820-3; K - AT1TE62820-5; L - AT4TE03295-3; M - AT4TE03295-4. RT- = Minus reverse transcriptase control.



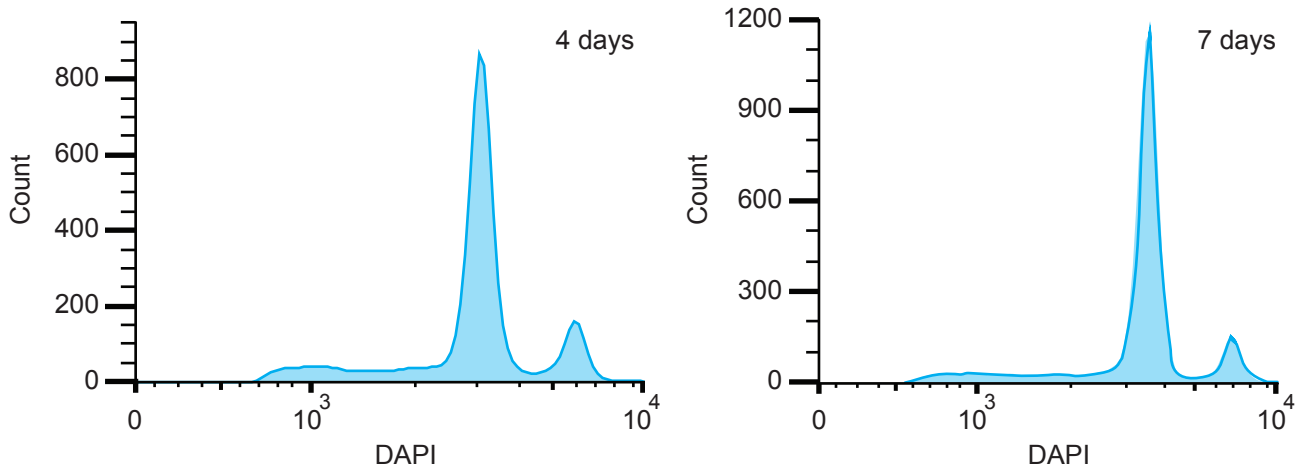
Supplementary Figure S7. Fraction of cytosine methylation. Fraction of methylated cytosines for Gypsy (blue) and Helitron (grey) elements, depending on the context (CG, CHG, CHH, where H=A, T or C) and region (P, pericentromeric; NP, non-pericentromeric).



Supplementary Figure S8. Colocalization of G4 quadruplexes and TEs. Frequency of the predicted G quadruplexes (G4) in various TE families, as indicated, in whole genome, non-pericentromeric and pericentromeric regions.



Supplementary Figure S9. GC content in various Arabidopsis TE families. Note that retrotransposons (blue) tend to have a higher GC content compared to DNA transposons (grey).



Supplementary Figure S10. Flow cytometry profile of Arabidopsis MM2d cells used in this study. The DNA content distribution of DAPI-stained nuclei of cultured Arabidopsis cells was estimated by flow cytometry at the indicated times after subculturing.

Supplementary Table 1. Number of peaks returned by the different peak calling algorithms in (A) all the genome and (B) transposable elements (TEs).

A

Algorithm	Number of peaks	Mean peak length (nt)	Total size (nt)	% genome
BayesPeak	1,131	1,219	1,378,286	1.16
HOMER	1,344	950	1,276,806	1.07
MACS1.4	3,230	653	2,110,158	1.77
MACS2.0	1,308	537	703,174	0.59
SICER	8,442	670	5,658,500	4.75
T-PIC	5,836	1,232	7,189,942	6.03

B

Algorithm	Number of peaks	Mean peak length (nt)	Total size (nt)	% TEs
BayesPeak	81	1,426	115,521	0.10
HOMER	49	5,173	253,494	0.21
MACS1.4	161	912	171,504	0.14
MACS2.0	54	571	30,883	0.03
SICER	372	4,758	1,769,840	1.49
T-PIC	599	1,163	696,548	0.58

Supplementary Table S2. List of primers used in this study indicating the procedures where they have been used: enrichment of nascent strands (SNS), RNA levels (RNA) and chromatin immunoprecipitation (ChIP).

Primer name	Sequence (5' -> 3')	SNS	RNA	ChIP
AT1TE62820 1F	GTCGACAAGCCAAACTGGAT	x		
AT1TE62820 1R	CCGATTTTCGGTTCTCCATA			
AT1TE62820 2F	TCCCGACAACAAAGTGAAGA	x		
AT1TE62820 2R	CAAATAAACGGGAGGGTGTG			
AT1TE62820 3F	ACTGCGGAAAACGCTACTCA	x	x	x
AT1TE62820 3R	ACATCGACCGGGAACATAGC			
AT1TE62820 4F	TCCGAATTGGTTCTCTCCAG	x		
AT1TE62820 4R	TGTCACCACTCGAGCTATGC			
AT1TE62820 5F	TTGCTTCGGAAACTTTCGCC	x	x	x
AT1TE62820 5R	ACCTCGTAGAGCGGGTGATA			
AT1TE62820 6F	TCTTCACCGACTACCGCATA	x		
AT1TE62820 6R	CGATCGGAACGAAATCCTAA			
AT1G51350 1F	AACACAATACCACAAACCAAAG	x		
AT1G51350 1R	AGTCAATGGAGTATAGATAGAG			
AT1G51350 2F	TTCCAATCTAAGCCAAAACTC	x		
AT1G51350 2R	ATCAGAATCGTCAGCATCAGC			
AT1G51350 3F	GTATTATCGCTCATGCTTGTG	x		
AT1G51350 3R	TGACAACTAAGCAAAGACAAG			
AT1G51350 4F	TCAATGGATCCAAATACTCGG	x		
AT1G51350 4R	TCAACAAGATTACGGAGGAGG			
AT1G51350 5F	CTTTACTTGTGCCACTTTTAGA	x		
AT1G51350 5R	ATTTTAATTTTATGTTTTGCCACG			
AT1G51350 6F	GTGGGTTTGAATTTCTGGTAG	x		
AT1G51350 6R	TGTACCCAATAAAAAGGAAATG			
AT4TE16725 1F	TGAAAAGCACTACTGCGCTAA	x		
AT4TE16725 1R	TCAACCGGACTGTTTGTTCA			
AT4TE16725 2F	CAGATCGGAAAAGGGAAGAG	x	x	x
AT4TE16725 2R	CCAAGGTAATTTCCCTCCTTC			
AT4TE16725 3F	CCCGGATTCCAGTTCACCTC	x	x	x
AT4TE16725 3R	TGGAGGCCATTAACGTGGAC			
AT4TE16725 4F	CACCGCGTGAACGAGATAGT	x		
AT4TE16725 4R	GCTACAGGACCATGTGAACCA			
AT4TE16725 5F	CAGTCAAAGGGGCAACCTAC	x		
AT4TE16725 5R	AAGTGTGCCTTGACCTTTGG			
AT4TE16725 6F	CTACAATGCCCTGCTCATCA	x		
AT4TE16725 6R	GCACCGGAGTCGTCAGTTAT			
AT2TE15565 1F	AAAGTGAAACCGGGTCAAAA	x		
AT2TE15565 1R	TCACAAAAATACAAGTGGGTTAAA			
AT2TE15565 2F	TCCTCACACGCTCCACATTA	x	x	x
AT2TE15565 2R	GAAACCTGAGGCTGAGGAGA			
AT2TE15565 3F	ACCCGGACTCCTCTTAGGAC	x	x	x
AT2TE15565 3R	AAGGCACAGTGGTGGTGAAT			

AT2TE15565 4F	CTCGAACCAGTCATCTCATGG	x		
AT2TE15565 4R	GTGGTGGCAGTGGTCAAGA			
AT4TE03295 1F	TTGAAGAGTCCCCATTCCAC	x		
AT4TE03295 1R	CGCGCCAATTTACTGGTATT			
AT4TE03295 2F	TGAACTGGACCGGCTACTTT	x		
AT4TE03295 2R	GGTCCCCGTTAGGAAACAAG			
AT4TE03295 3F	GTAAACCACCTGCGACCATT	x	x	x
AT4TE03295 3R	TGAGCTGCGATACTGTCCAC			
AT4TE03295 4F	AGTCTAGCACGAGAGTGGCT	x	x	x
AT4TE03295 4R	TTCTTTGTTCCAGGCAGCGT			
AT4TE03295 5F	TTCTGCGGAGAGGAGGAGTC	x		
AT4TE03295 5R	GGTATTGGGATCCAAGGGCT			
AT4TE03295 6F	GTGGCTACATCCCTGACGAT	x		
AT4TE03295 6R	ACGGTGGAAAACTCCGGTAT			
AT2TE13970 F	TTCGGCTTGATTTGAGCCAC		x	x
AT2TE13970 R	GTTAGCTGGTGGGAGGACAG			
AT2TE16335 F	TGTGAAAGCCCATTACACTTTACT		x	x
AT2TE16335 R	ACGACGAATAAGGTTTGGGAGA			
AT4TE16735 F	TCGTCAAAATCTTTGGTGCTTGT		x	x
AT4TE16735 R	GCAATTAACGATCACCTAATCATGG			
AT4TE17050 F	ACACCAAATCTGCCACTCCA		x	x
AT4TE17050 R	TGTTTTTGGTGAATGATTGGATAGT			
GAPC2 F	TCCAACGCTAGTTGCACCAC		x	
GAPC2 R	TGGACAGTGGTCATGAGTCC			
Neg AT2G28970 1F	AGAATGATTTGTTCGGACGTG	x		
Neg AT2G28970 1R	GCCAGAAAAGAAGGCACAAG			
Neg AT2G28970 2F	CAAACCACCAATGGTTCACA	x		
Neg AT2G28970 2R	GGAGTTCCAGCCACAACAGT			

Supplementary Table S4. ORI distribution in the indicated genomic locations. ORI location was determined using uniquely mapped reads.

Feature	MACS14 – All genome		Non-pericentromeric		Pericentromeric	
	No. ORIs	% ORIs	No. ORIs	% ORIs	No. ORIs	% ORIs
Genes	2888	89.41	2724	94.98	164	45.3
TEs	161	4.98	39	1.36	122	33.7
Non-annotated	154	4.77	83	2.89	71	19.61
Overlapping	27	0.84	22	0.77	5	1.38
Total	3230	100	2868	100	362	100

Supplementary Table S5. Distribution of ORIs resulting from multihit reads in the different genomic features.

Feature	MACS1.4 Multihit– All genome	
	No. ORIs	% ORIs
Genes	77	31.3
TEs	118	48.0
Non-annotated	47	19.1
Overlapping	4	1.6
Total	246	100

Supplementary Table S6. Nucleotide coverage of TE families with respect to the total TE coverage and their contribution to ORI. Only ORIs overlapping with a unique TE family were included in this analysis.

Family	All genome		Non-pericentromeric		Pericentromeric	
	% nt	% ORI-TEs	% nt	% ORI-TEs	% nt	% ORI-TEs
RC/Helitron	29.42	0.00	52.31	0.00	18.80	0.00
LTR/Gypsy	29.40	46.98	3.97	16.67	41.20	56.64
DNA/MuDR	15.75	3.36	16.39	5.56	15.45	2.65
LTR/Copia	7.67	14.09	8.91	22.22	7.10	11.50
LINE/L1	5.38	22.82	6.83	30.56	4.71	20.35
DNA/EN-SPM	4.63	4.70	1.30	5.56	6.18	4.42
Other families	7.74	10.62	10.30	6.19	6.56	4.42