

Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation

Tim Stuart¹, Steven R. Eichten², Jonathan Cahn¹, Yuliya Karpievitch¹, Justin Borevitz² and Ryan Lister¹

¹ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, Australia

²ARC Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, Australia

Corresponding author: Ryan Lister ryan.lister@uwa.edu.au

Author ORCID IDs:

0000-0002-3044-0897 (TS)

0000-0003-2268-395X (SRE)

0000-0002-5006-741X (JC)

0000-0001-6637-7239 (RL)

¹ Abstract

Variation in the presence or absence of transposable elements (TEs) is a major source of genetic variation between individuals. Here, we identified 23,095 TE presence/absence variants between 216 wild *Arabidopsis* accessions. Most variants are rare, and we find a burden of rare variants associated with extremes of both gene expression and DNA methylation levels within the population. Of the common alleles identified, two thirds were not in linkage disequilibrium with nearby SNPs, implicating these TE variants as a source of novel genetic diversity. Nearly 200 common TE variants were associated with significantly altered expression of nearby genes, with potential functional consequences including decreased pathogen resistance. A major fraction of inter-accession DNA methylation differences were associated with nearby TE insertion variants, with DNA, indicating an important role of TE variants in facilitating epigenomic variation.

12 Introduction

13 Transposable elements (TEs) are mobile genetic elements present in nearly all studied organisms, and
14 make up a large fraction of most eukaryotic genomes. The two classes of TEs are retrotransposons
15 (class I elements), which transpose via an RNA intermediate requiring a reverse transcription reaction,
16 and DNA transposons (class II elements), which transpose via either a cut-paste or, in the case of
17 Helitrons, a rolling circle mechanism with no RNA intermediate [1]. TE activity poses mutagenic
18 potential as a TE insertion may disrupt essential regions of the genome. Consequently, safeguards
19 have evolved in order to suppress this activity. These safeguards include epigenetic transcriptional
20 silencing mechanisms, chiefly involving the methylation of cytosine nucleotides (DNA methylation) to
21 produce 5-methylcytosine (mC), a mark that can signal transcriptional silencing of the methylated
22 locus. In *Arabidopsis thaliana* (Arabidopsis), DNA methylation occurs in three DNA sequence contexts:
23 mCG, mCHG, and mCHH, where H is any base but G. Establishment of DNA methylation marks
24 can be carried out by two distinct pathways – the RNA-directed DNA methylation pathway guided by
25 24 nucleotide (nt) small RNAs (smRNAs), and the DDM1/CMT2 pathway [2, 3]. A major function of
26 DNA methylation in Arabidopsis is in the transcriptional silencing of TEs. Loss of DNA methylation
27 due to mutations in genes essential for DNA methylation establishment and maintenance leads to
28 expression of previously silent TEs, and in some cases transposition [2, 4–8].

29 TEs are thought to play an important role in evolution, not only because of the disruptive potential of
30 their transposition. The release of transcriptional and post-transcriptional silencing of TEs can lead to
31 bursts of TE activity, quickly generating new genetic diversity [9]. TEs may carry regulatory information
32 such as promoters and transcription factor binding sites, and their mobilization may lead to the
33 creation or expansion of gene regulatory networks [10–13]. Furthermore, the transposase enzymes
34 required and encoded by TEs have frequently been domesticated and repurposed as endogenous
35 proteins, such as the *DAYSLEEPER* gene in Arabidopsis, derived from a hAT transposase enzyme
36 [14]. Clearly, the activity of TEs can have widespread and unpredictable effects on the host genome.
37 However, the identification of TE presence/absence variants in genomes has remained difficult to
38 date. It is challenging to identify the structural changes in the genome caused by TE mobilization
39 using current short-read sequencing technologies as these reads are typically mapped to a reference
40 genome, which has the effect of masking structural changes that may be present. However, in terms
41 of the number of base pairs affected, a large fraction of genetic differences between Arabidopsis
42 accessions appear to be due to variation in TE content [15, 16]. Therefore identification of TE variants
43 is essential in order to develop a more comprehensive understanding of the genetic variation that
44 exists between genomes, and of the consequences of TE movement on genome and cellular function.

45 The tools developed previously for identification of novel TE insertion sites have several limitations.
46 They either require a library of active TE sequences, cannot identify TE absence variants, or are not
47 designed with population studies in mind, or suffer from a high degree of false-negatives [16–19].
48 In order to accurately map the locations of TE presence/absence variants with respect to a refer-
49 ence genome, we have developed a novel algorithm, TEPID (Transposable Element Polymorphism
50 IDentification), which is designed for population studies. We tested our algorithm using both sim-
51 ultated and real Arabidopsis sequencing data, finding that TEPID is able to accurately identify TE
52 presence/absence variants with respect to the Col-0 reference genome. We applied our TE analysis
53 method to existing genome resequencing data for 216 different wild Arabidopsis accessions, and
54 identified widespread TE variation amongst these accessions [20].

55 **Results**

56 **Computational identification of TE presence/absence variation**

57 We developed TEPID, an analysis pipeline capable of detecting TE presence/absence variants from
58 paired end DNA sequencing data. TEPIPID integrates split and discordant read mapping information,
59 read mapping quality, sequencing breakpoints, as well as local variations in sequencing coverage
60 to identify novel TE presence/absence variants with respect to a reference TE annotation (Figure
61 1; see methods). This typically takes 5-10 minutes using Arabidopsis sequencing data at 20-40x
62 coverage (excluding the read mapping step). After TE variant discovery has been performed, TEPIPID
63 then includes a second refinement step designed for population studies. This examines each region
64 of the genome where there was a TE insertion identified in any member of the group, and checks for
65 evidence of this insertion in all other members of the group. In this way, TEPIPID leverages TE variant
66 information for a group of related samples to correct false negative calls within the group. Testing
67 of TEPIPID using simulated TE variants in the Arabidopsis genome showed that it was able to reliably
68 detect simulated TE variants at sequencing coverage levels commonly used in genomics studies
69 (Figure 1 - figure supplement 1).

70 In order to further assess the sensitivity and specificity of TE variant discovery using TEPIPID, we
71 identified TE variants in the *Landsberg erecta* (*Ler*) accession, and compared these with the *Ler*
72 genome assembly created using long PacBio sequencing reads [21]. Previously published 100
73 bp paired-end *Ler* genome resequencing reads [22] were first analyzed using TEPIPID, enabling
74 identification of 446 TE insertions (Figure 1 - source data 1) and 758 TE absence variants (Figure 1 -
75 source data 2) with respect to the Col-0 reference TE annotation. Reads providing evidence for these
76 variants were then mapped to the *Ler* reference genome that was generated by *de novo* assembly
77 using Pacific Biosciences P5-C3 chemistry with a 20 kb insert library [21], using the same alignment
78 parameters as was used to map reads to the Col-0 reference genome. This resulted in 98.7% of
79 reads being aligned concordantly to the *Ler* reference, whereas 100% aligned discordantly or as
80 split reads to the Col-0 reference genome (Table 1). To find whether reads mapped to homologous
81 regions in both the Col-0 and *Ler* reference genomes, we conducted a blast search [23] using the
82 DNA sequence between read pair mapping locations in the *Ler* genome against the Col-0 genome,
83 and found the top blast result for 80% of reads providing evidence for TE insertions, and 89% of
84 reads providing evidence for TE absence variants in *Ler*, to be located within 200 bp of the TE variant
85 reported by TEPIPID. We conclude that reads providing evidence for TE variants map discordantly or
86 as split reads when mapped to the Col-0 reference genome, but map concordantly to homologous
87 regions of the *Ler* *de novo* assembled reference genome, indicating that structural variation is present
88 at the sites identified by TEPIPID, and that this is resolved in the *de novo* assembled genome.

89 To estimate the rate of false negative TE absence calls made using TEPIPID, we compared our *Ler* TE
90 absence calls to the set of TE absences in *Ler* genome identified previously by aligning full-length
91 Col-0 TEs to the *Ler* reference using BLAT [16]. We found that 89.6% (173/193) of these TE absences
92 were also identified using TEPIPID, indicating a false negative rate of around 10% for TE absence calls.
93 To determine the rate of false negative TE insertion calls, we ran TEPIPID using 90 bp paired-end
94 Col-0 reads (Col-0 control samples from [24]), aligning reads to the *Ler* PacBio assembly. As TEPIPID
95 requires a high-quality TE annotation to discover TE variants, which is not available for the *Ler*
96 assembly, we simply looked for discordant and split read evidence at the known Col-0-specific TE
97 insertion sites [16], and found evidence reaching the TEPIPID threshold for a TE insertion call to be

98 made at 89.6% (173/193) of these sites, indicating a false negative rate of 10%. However, we note
99 that this estimate does not take into account the TEPID refinement step used on large populations,
100 and so the false negative rate for samples analyzed in the population from Schmitz et al. (2013) is
101 likely to be slightly lower than this estimate, as each accession gained on average 4% more insertion
102 calls following this refinement step (Figure 2 - figure supplement 1).

103 Abundant TE positional variation among natural *Arabidopsis* populations

104 We used TEPID to analyze previously published 100 bp paired-end genome resequencing data
105 for 216 different *Arabidopsis* accessions [20], and identified 15,007 TE insertions and 8,088 TE
106 absence variants relative to the Col-0 reference accession, totalling 23,095 unique TE variants. In
107 most accessions we identified 300-500 TE insertions (mean = 378) and 1,000-1,500 TE absence
108 variants (mean = 1,279), the majority of which were shared by two or more accessions (Figure 2 -
109 figure supplement 2). PCR validations were performed for a random subset of 10 insertions and 10
110 absence variants in 14 accessions (totalling 280 validations), and confirmed the high accuracy of TE
111 variant discovery using the TEPID package, with results similar to that observed using simulated data
112 and the Ler genome analysis (Figure 2 - figure supplement 3). The number of TE insertions identified
113 was positively correlated with sequencing depth of coverage, while the number of TE absence variants
114 identified had no correlation with sequencing coverage (Figure 2 - figure supplement 4A, B), indicating
115 that the sensitivity of TE absence calls is not limited by sequencing depth, while TE insertion calls
116 benefitted from high sequencing depth. However, accessions with low coverage gained more TE
117 insertion calls during the TEPID refinement step (Figure 2 - figure supplement 4C), indicating that
118 these false negatives were effectively reduced by leveraging TE variant information for the whole
119 population.

120 As TE insertion and TE absence calls represent an arbitrary comparison to the Col-0 reference
121 genome, we sought to remove these arbitrary comparisons and classify each variant as a new TE
122 insertion or true deletion of an ancestral TE in the population. To do this, we examined the minor
123 allele frequency (MAF) of each variant in the population, under the expectation that the minor allele is
124 the derived allele. We re-classified common TE absences relative to Col-0 as rare TE insertions in
125 Col-0, and common TE insertions relative to Col-0 as true TE deletions in Col-0. Cases where the TE
126 variant was at an high MAF (>20%) were assigned NA calls, as it could not be determined if these
127 were cases where the variant was most likely to be due to a true TE deletion or due to a new TE
128 insertion. While these classifications are not definitive, as there will be rare cases where a true TE
129 deletion has spread through the population and becomes the common allele, we will correctly classify
130 most TE variants. Overall, we found 72.3% of the TE absence variants identified with respect to the
131 Col-0 reference genome were likely due to a true TE deletion in these accessions, while 4.8% were
132 due to new insertions in Col-0 not shared by others in the population (Table 2). Overall, we identified
133 15,077 new TE insertions, 5,856 true TE deletions, and 2,162 TE variants at a high MAF that were
134 unable to be classified as an insertion or deletion.

135 TE insertions and deletions were distributed throughout chromosome 1 in a pattern that was similar
136 to the distribution of all Col-0 TEs (Figure 2A). TE deletions and common TE variants were found in
137 similar chromosomal regions, as deletion variants represent the rare loss of common variants. TE
138 deletions and common variants were more highly enriched in the pericentromeric regions than rare
139 variants or TE insertions. The distribution of rare TE variants and TE insertions was somewhat similar
140 to that observed for regions of the genome previously identified as being differentially methylated

in all DNA methylation contexts (mCG, mCHG, mCHH) between the wild accessions (population C-DMRs), while population CG-DMRs, differentially methylated in the mCG context, less frequently overlapped with all types of TE variants identified [20]. Furthermore, TE variants were depleted within genes and DNase I hypersensitivity sites [25], while they were enriched in gene flanking regions and within other annotated TEs or pseudogenes (Figure 2B). We found that TE deletions and common TE variants were enriched within the set of TE variants found in gene bodies (Figure 2C, D). We did not find any significant enrichment of TE variants within the *KNOT ENGAGED ELEMENT* (KEE) regions, previously identified as regions that may act as a “TE sink” [26] (Figure 2 - figure supplement 5). This may indicate that these regions do not act as a “TE sink” as has been previously proposed, or that the “TE sink” activity is restricted to very recent insertions, as the insertions we analysed in this population were likely older than those used in the study by Grob et al.

Among the identified TE variants, several TE superfamilies were over- or under-represented compared to the number expected by chance given the overall genomic frequency of different TE types (Figure 2E). In particular, both TE insertions and deletions in the RC/Helitron superfamily were less numerous than expected, with an 11.5% depletion of RC/Helitron elements in the set of TE variants. In contrast, TEs belonging to the LTR/Gypsy superfamily were more frequently deleted than expected, with a 17% enrichment in the set of TE deletions. This was unlikely to be due to a differing ability of our detection methods to identify TE variants of different lengths, as the TE variants identified had a similar distribution of lengths as all *Arabidopsis* TEs annotated in the Col-0 reference genome (Figure 2 - figure supplement 6). These enrichments suggest that the RC/Helitron TEs have been relatively dormant in recent evolutionary history, while the LTR/Gypsy, which are highly enriched in the pericentromeric regions, are frequently lost from the *Arabidopsis* genome. At the family level, we observed similar patterns of TE variant enrichment or depletion (Figure 2 - figure supplement 7; source data 5).

We further examined *Arabidopsis* (Col-0) DNA sequencing data from a transgenerational stress experiment to investigate the possible minimum number of generations required for TE variants to arise [24]. We identified a single potential TE insertion in a sample following 10 generations of single-seed descent under high salinity stress conditions, and no TE variants in the control single-seed descent set. However, without experimental validation it remains unclear if this represents a true variant. Therefore, we conclude that TE variants may arise at a rate less than 1 insertion in 30 generations under laboratory conditions. Further experimental work will be required to precisely determine the rate of transposition in *Arabidopsis*.

Relationship between TE variants and single nucleotide polymorphisms

Although thousands of TE variants were identified, they may be linked (i.e. ‘tagged’) by the previously identified single nucleotide polymorphisms (SNPs), or unlinked from SNPs across the accessions. We tested how often common TE variants (>3 % MAF; 7 accessions) were linked to adjacent SNPs to determine when they would represent a previously unassessed source of genetic variation between accessions. SNPs that were previously identified between the accessions [20] were compared to the presence/absence of individual TE variants. For the testable TE variants in the population, the nearest 300 flanking SNPs upstream and downstream of the TE variant site were analyzed for local linkage disequilibrium (LD, r^2 ; see methods). TE variants were classified as being either ‘low’, ‘mid’, or ‘high’ LD variants by comparing ranked r^2 values of TE variant to SNPs against the median ranked r^2 value for all between SNP comparisons (SNP-SNP) to account for regional variation in the extent

of SNP-SNP LD (Figure 3A, B) due to recombination rate variation or selection [27]. The majority (61%) of testable TE variants had low LD with nearby SNPs, and represent a source of genetic diversity not previously assessed by current SNP-based genotype calling methods (Figure 3C). 29% of TE variants displayed high levels of LD and are tagged by nearby SNPs, while only 10% had intermediate levels of LD. We observed a positive correlation between TE variant MAF and LD state, with variants of a high minor allele frequency more often classified as high-LD (Figure 3D). While the proportion of TE variants classified as high, mid, or low-LD was mostly the same for both TE insertions and TE deletions, TE variants with a high MAF (>20%) that were unable to be classified as either true deletions or as new insertions had a much higher proportion of high-LD variants (Figure 3E). This was consistent with the observation that the more common alleles were more often in a high-LD state. TE variants displayed a similar distribution over chromosome one regardless of linkage classification (Figure 3 - figure supplement 1). Overall, this analysis revealed an abundance of previously unexplored genetic variation that exists amongst *Arabidopsis* accessions caused by the presence or absence of TEs, and illustrates the importance of identifying TE variants alongside other genetic diversity such as SNPs.

TE variants affect gene expression

To determine whether the TE variants identified affected nearby gene expression, we compared the steady state transcript abundance within mature leaf tissue between accessions with and without TE insertions or deletions, for genes with TE variants located in the 2 kb gene upstream region, 5' UTR, exon, intron, 3' UTR or 2 kb downstream region (Figure 4A). While the steady state transcript abundance of most genes appeared to be unaffected by the presence of a TE, 196 genes displayed significant differences in transcript abundance linked with the presence of a TE variant, indicating a role for these variants in the local regulation of gene expression (1% false discovery rate; greater than 2-fold change in transcript abundance; Figure 4A, Figure 4 - source data 1). We did not find any functional category enrichments in this set of differentially expressed genes. As rare TE variants with a MAF less than 3% may also be associated with a difference in transcript abundance, but were unable to be statistically tested due to their rarity, we performed a burden test for enrichment of rare variants in the extremes of expression [28]. Briefly, this method counts the frequency of rare variants within each gene expression rank in the population, and aggregates this information over the entire population to determine whether an enrichment of rare variants exists within any gene expression rank. We found a strong enrichment for gene expression extremes for TE variants in all gene features tested (Figure 4B). While TE variants in gene upstream regions showed a strong enrichment of both high and low gene expression ranks, TE variants in exons or gene downstream regions seemed to have a stronger enrichment for low expression ranks than high ranks. Randomization of the accession names removed these enrichments completely (Figure 4 - figure supplement 1), and there was little difference between TE insertions and TE deletions in the gene expression rank enrichments found (Figure 4 - figure supplement 2). This rare variant analysis further indicated that TE variants may alter the transcript abundance of nearby genes.

As both increases and decreases in transcript abundance of nearby genes were observed for TE variants within each gene feature, it appears to be difficult to predict the impact a TE variant may have on nearby gene expression. Furthermore, gene-level transcript abundance measurements may fail to identify the potential positional effect of TE variants upon transcription. To more closely examine changes in transcript abundance associated with TE variants among the accessions, we

inspected a subset of TE variant sites and identified TE variants that appear to have an impact on transcriptional patterns beyond changes in total transcript abundance from a nearby gene. For example, the presence of a TE insertion within an exon of *AtRLP18* (AT2G15040) was associated with truncation of the transcripts at the TE insertion site in accessions possessing the TE variant, as well as silencing of a downstream gene encoding a leucine-rich repeat protein (AT2G15042) (Figure 5A, B). Both genes had significantly lower transcript abundance in accessions containing the TE insertion ($p < 5.8 \times 10^{-10}$, Mann-Whitney U test). *AtRLP18* is reported to be involved in bacterial resistance, with the disruption of this gene by T-DNA insertion mediated mutagenesis resulting in increased susceptibility to the bacterial plant pathogen *Pseudomonas syringae* [29]. We examined pathogen resistance phenotype data [30], and found that accessions containing the TE insertion in the *AtRLP18* exon were more often sensitive to infection by *Pseudomonas syringae* transformed with *avrPpH3* genes (Figure 5C). This suggests that the accessions containing this TE insertion within *AtRLP18* may have an increased susceptibility to certain bacterial pathogens.

We also observed some TE variants associated with increased expression of nearby genes. For example, the presence of a TE within the upstream region of a gene encoding a pentatricopeptide repeat (PPR) protein (AT2G01360) was associated with higher steady state transcript abundance of this gene (Figure 5D, E). Transcription appeared to begin at the TE insertion point, rather than the transcriptional start site of the gene (Figure 5D). Accessions containing the TE insertion had significantly higher AT2G01360 transcript abundance than the accessions without the TE insertion ($p < 1.8 \times 10^{-7}$, Mann-Whitney U test). The apparent transcriptional activation, linked with presence of a TE belonging to the *HELI TRON1* family, indicates that this element may carry sequences or other regulatory information that has altered the expression of genes downstream of the TE insertion site. Importantly, this variant was classified as a low-LD TE insertion, as it is not in LD with surrounding SNPs, and therefore the associated changes in gene transcript abundance would not be identified using only SNP data. This TE variant was also upstream of *QPT* (AT2G01350), involved in NAD biosynthesis [31], which did not show alterations in steady state transcript abundance associated with the presence of the TE insertion, indicating a potential directionality of regulatory elements carried by the TE (Figure 5D, E). Overall, these examples demonstrate that TE variants can have unpredictable, yet important, effects on the expression of nearby genes, and these effects may be missed by studies focused on genetic variation at the level of SNPs.

TE variants explain many DNA methylation differences between accessions

As TEs are frequently highly methylated in *Arabidopsis* [32–35], we next assessed the DNA methylation state surrounding TE variant sites to determine whether TE variants might be responsible for some of the differences in DNA methylation patterns previously observed between the wild accessions [20]. We found that TE variants were often physically close to DMRs (Figure 6A). Furthermore, C-DMRs were more often close to a TE insertion than expected, while they were rarely near a TE deletion (Table 3). CG-DMRs were rarely close to TE insertions or TE deletions. Overall, we found 48% of the 13,482 previously reported population C-DMRs were located within 1 kb of a TE variant (predominantly TE insertions), while only 15% of CG-DMRs were within 1 kb of a TE variant (Table 3). For C-DMRs, this was significantly more than expected by chance, while it was significantly less than expected for CG-DMRs ($p < 1 \times 10^{-4}$, determined by resampling 10,000 times). To determine if DMR methylation levels were dependent on the presence/absence of nearby TE variants, we calculated Pearson correlation coefficients between the DNA methylation level at each DMR and

270 the presence/absence of the nearest TE variant. We observed a negative correlation between the
271 distance from a C-DMR to the nearest TE insertion and the correlation between the DNA methylation
272 level at the C-DMR with the presence/absence of the TE insertion (Figure 6B). This suggests a
273 distance-dependent effect of TE insertion presence on C-DMR methylation. In contrast, we found no
274 such relationship for TE deletions on C-DMRs, or for insertions or deletions on CG-DMRs (Figure
275 6B). DNA methylation levels at C-DMRs located within 1 kb of a TE insertion (TE-DMRs) were more
276 often positively correlated with the presence/absence of a TE insertion than the DNA methylation
277 levels at C-DMRs further than 1 kb from a TE insertion (non-TE-DMRs). This was evident from the
278 distribution of correlations for non-TE-DMRs being centred around zero, whereas for TE-DMRs this
279 distribution was skewed to the right (Figure 6C, $D=0.23$). For TE deletions, we did not observe such a
280 difference in the distributions of correlation coefficients between TE-DMRs and non-TE-DMRs, nor for
281 CG-DMRs and their nearby TE insertions or deletions (Figure 6C, $D=0.07-0.10$). Furthermore, DNA
282 methylation levels were often higher in the presence of the nearby TE insertion, while this relationship
283 was generally not observed for C-DMRs further than 1 kb from a TE variant, for TE deletions, or for
284 CG-DMRs (Figure 6 - figure supplement 1).

285 As the above correlations between TE presence/absence and DMR methylation level rely on the
286 TE variants having a high MAF, this precludes an analysis of the effect of rare variants on DMR
287 methylation levels. To determine the effect that these rare TE variants may have on DMR methylation
288 levels, we performed a burden test for enrichment of DMR methylation extremes at TE-DMRs, similar
289 to as was done to test the effect of rare variants on gene expression. We found a strong enrichment
290 for high C-DMR and CG-DMR methylation level ranks for TE insertions, while TE deletions were
291 associated with both high and low extremes of DNA methylation levels at C-DMRs, and less so
292 at CG-DMRs (Figure 6D). This further indicates that the presence of a TE insertion is associated
293 with higher C-DMR methylation levels, while TE deletions seem to have more variable effects on
294 DMR methylation levels. This enrichment was completely absent after repeating the analysis with
295 randomized accession names (Figure 6 - figure supplement 2). We also observed a slight enrichment
296 for low DMR methylation ranks for TE insertions near CG-DMRs, indicating that the insertion of a
297 TE was sometimes associated with reduced CG methylation in nearby regions (<1 kb from the TE).
298 We examined these TE insertions in a genome browser, and found that some TE insertions were
299 associated with decreased transcript abundance of nearby genes, with a corresponding loss of gene
300 body methylation, offering a potential explanation for the decreased CG methylation observed near
301 some TE insertions (Figure 6 - figure supplement 3).

302 We next examined levels of DNA methylation in regions flanking all TE variants regardless of the
303 presence or absence of a population DMR call. While DNA methylation levels around pericentromeric
304 TE insertions and deletions (<3 Mb from a centromere) seemed to be unaffected by the presence of a
305 TE insertion (Figure 7A), TE insertions in the chromosome arms were associated with an increase in
306 DNA methylation levels in all contexts (Figure 7A, B). In contrast, TE deletions in the chromosome
307 arms did not affect patterns of DNA methylation, as the flanking methylation level in all contexts
308 seemed to remain high following deletion of the TE (Figure 7A, C). As the change in DNA methylation
309 levels around TE variant sites appeared to be restricted to regions <200 bp from the insertion site, we
310 correlated DNA methylation levels in 200 bp regions flanking TE variants with the presence/absence
311 of TE variants. DNA methylation levels were often positively correlated with the presence of a
312 TE insertion when the insertion was distant from a centromere (Figure 7D). TE deletions were more
313 variably correlated with local DNA methylation levels, but also showed a slight bias towards positive
314 correlations for TE deletions distant from the centromeres. As methylome data was available for both
315 leaf and bud tissue for 12 accessions, we repeated this analysis comparing between tissue types, but

316 did not observe any difference in the patterns of methylation surrounding TE variant sites between
317 the two tissues (Figure 7 - figure supplement 1).

318 These results indicate that local DNA methylation patterns are influenced by the differential TE content
319 between genomes, and that the DNA methylation-dependent silencing of TEs may lead to formation
320 of DMRs between wild *Arabidopsis* accessions. TE insertions appear to be important in defining
321 local patterns of DNA methylation, while DNA methylation levels often remain elevated following a TE
322 deletion, and so are independent from the presence or absence of TEs in these cases. Importantly,
323 the distance from a TE insertion to the centromere appears to have a strong impact on whether an
324 alteration of local DNA methylation patterns will occur. This is likely due to flanking sequences being
325 highly methylated in the pericentromeric regions, and so the insertion of a TE cannot further increase
326 levels of DNA methylation. Overall, a large fraction of the population C-DMRs previously identified
327 between wild accessions are correlated with the presence of local TE insertions, but not TE deletions.
328 CG-DMR methylation levels seem to be mostly independent from the presence/absence of common
329 TE variants, while rare TE variants have an impact on DNA methylation levels at both C-DMRs and
330 CG-DMRs.

331 **Genome-wide analysis of TE variant association with methylation levels high- 332 lights distant and local control**

333 To better quantify the above results, an association scan was conducted for all common TE variants
334 (>3% MAF) and all population C-DMRs for the 124 accessions with both DNA methylation and TE
335 variant data available. To test the significance of each pairwise correlation, we collected bootstrap
336 p-value estimates based on 500 permutations of accession labels. TE-DMR associations were
337 deemed significant if they had a real association more extreme than any of the 500 permutations ($p <$
338 1/500). A band of significant associations was observed for TE insertions and their nearby C-DMRs,
339 signifying a local association between TE insertion presence/absence and C-DMR methylation (Figure
340 8A). This local association was not as strong for TE deletions (Figure 8B), consistent with our above
341 findings. While TE variants and DNA methylation showed a local association, it is also possible that
342 TE variation may influence DNA methylation state more broadly in the genome, perhaps through
343 production of *trans*-acting smRNAs or inactivation of genes involved in DNA methylation establishment
344 or maintenance. To identify any potential enrichment of C-DMRs regulated in *trans*, we summed the
345 total number of significant associations for each TE variant across the whole genome (Figure 8A and
346 B, top panels). At many sites, we found far more significant associations than expected due to the
347 false positive rate alone. This suggested the existence of many putative *trans* associations between
348 TE variants and genome-wide C-DMR methylation levels. We further examined these C-DMRs that
349 appeared to be associated with a TE insertion in *trans*, checking for TE insertions near these C-DMRs
350 that were present in the same accessions as the *trans* associated TE, as these could lead to a false
351 *trans* association. These were extremely rare, with only 4 such cases for TE insertions, and 38 cases
352 for TE deletions, and so were unable to explain the high degree of *trans* associations found.

353 **Discussion**

354 Here we discovered widespread differential TE content between wild *Arabidopsis* accessions, and
355 explore the impact of these variants at the level of individual accessions. Most TE variants were due
356 to the *de novo* insertion of TEs, while a smaller subset was likely due to the deletion of ancestral TE
357 copies, mostly around the pericentromeric regions. A subset (32%) of TE variants with a minor allele
358 frequency above 3% were able to be tested for linkage with nearby SNPs. The majority of these TE
359 variants were not in LD with SNPs, indicating that they represent genetic variants currently overlooked
360 in genomic studies. We found a marked depletion of TE variants within gene bodies and DNase I
361 hypersensitivity sites (putative regulatory regions), indicating that the more deleterious TE insertions
362 may have been removed from this population through selection. Of those TE variants found in gene
363 bodies, TE deletions were overrepresented, indicating that the loss of ancestral TEs inserted within
364 genes may be more frequent, or perhaps less deleterious, than the *de novo* insertion of TEs into
365 genes.

366 The identification of a large number of TE variants in this population gave an opportunity to form
367 statistically robust correlations between TE presence/absence and transcript abundance from nearby
368 genes, as well as genome-wide patterns of DNA methylation. We were able to identify examples where
369 TE variants appear to have an effect upon gene expression, both in the disruption of transcription
370 and in the spreading or disruption of regulatory information leading to the transcriptional activation of
371 genes, indicating that these TE variants can have important consequences upon the expression of
372 protein coding genes (Figure 5). In one case, these changes in gene expression could be linked with
373 phenotypic changes, with accessions containing a TE insertion more frequently sensitive to bacterial
374 infection. Further experiments will be needed to establish a causal link between this TE insertion and
375 the associated phenotype. An analysis of rare TE variants, present at a low MAF, further strengthened
376 this relationship between TE presence/absence and altered transcript abundance, as we were able to
377 identify a strong enrichment of rare TE variants in accessions with extreme gene expression ranks in
378 the population.

379 Perhaps most importantly, we provide evidence that differential TE content between genomes of
380 *Arabidopsis* accessions underlies a large fraction of the previously reported population C-DMRs, in
381 agreement with recent similar findings [16]. Thus, the frequency of pure epialleles, independent of
382 underlying genetic variation, may be even more rare than previously anticipated [36]. We did not
383 find evidence of CG-DMR methylation level being altered by the presence of common TE variants,
384 but rather rare TE variants may be more important in shaping patterns of DNA methylation at
385 some CG-DMRs, though the reasons for this distinction remain unclear. The level of local DNA
386 methylation changes associated with TE variants was also related to the distance from a TE variant
387 to the centromere, with variants in the chromosome arms being more strongly correlated with DNA
388 methylation levels. This seems to be due to a higher baseline level of DNA methylation at the
389 pericentromeric regions, which prevent any further increase in DNA methylation level following
390 insertion of a TE. Furthermore, we found an important distinction between TE insertions and TE
391 deletions in the effect that these variants have on nearby DNA methylation levels. While flanking
392 DNA methylation levels appeared to increase following a TE insertion, the deletion of an ancestral TE
393 was often not associated with a corresponding decrease in flanking DNA methylation levels (Figure
394 7). This indicates that high levels of DNA methylation, once established, may be maintained in the
395 absence of the TE insertion that presumably triggered the original change in DNA methylation level. It
396 is then possible that TE variants explain more of the variation in DNA methylation patterns than we

397 find direct evidence for, if some C-DMRs were formed by the insertion of an ancestral TE that is now
398 absent in all the accessions analysed here. These DMRs would then represent the epigenetic “scars”
399 of past TE insertions.

400 Finally, we performed a genome-wide scan of common TE variant association with C-DMR methylation
401 levels, and found further evidence of a strong local association between TE insertion presence/absence and C-DMR methylation level (Figure 8). We were also able to identify some TE
402 variants that appeared to be associated with changes in DNA methylation levels at multiple loci
403 throughout the genome, indicating a possible *trans* regulation of DNA methylation levels linked to
404 certain TE variants. Further experiments will be required to confirm and examine the role of these TE
405 variants in determining genome-wide patterns of DNA methylation. Overall our results show that TE
406 presence/absence variants between wild *Arabidopsis* accessions not only have important effects on
407 nearby gene expression, but can also have a role in determining local patterns of DNA methylation,
408 and explain many regions of differential DNA methylation previously observed in the population.
409

410 Methods

411 TEPID development

412 Mapping

413 FASTQ files are mapped to the reference genome using the ‘tepid-map’ algorithm (Figure 1). This
414 first calls bowtie2 [37] with the following options: ‘–local’, ‘–dovetail’, ‘–fr’, ‘-R5’, ‘-N1’. Soft-clipped and
415 unmapped reads are extracted using Samblaster [38], and remapped using the split read mapper
416 Yaha [39], with the following options: ‘-L 11’, ‘-H 2000’, ‘-M 15’, ‘-osh’. Split reads are extracted from
417 the Yaha alignment using Samblaster [38]. Alignments are then converted to bam format, sorted, and
418 indexed using samtools [40].

419 TE variant discovery

420 The ‘tepid-discover’ algorithm examines mapped bam files generated by the ‘tepid-map’ step to identify
421 TE presence/absence variants with respect to the reference genome. Firstly, mean sequencing
422 coverage, mean library insert size, and standard deviation of the library insert size is estimated.
423 Discordant read pairs are then extracted, defined as mate pairs that map more than 4 standard
424 deviations from the mean insert size from one another, or on separate chromosomes.

425 To identify TE insertions with respect to the reference genome, split read alignments are first filtered
426 to remove reads where the distance between split mapping loci is less than 5 kb, to remove split reads
427 due to small indels, or split reads with a mapping quality (MAPQ) less than 5. Split and discordant
428 read mapping coordinates are then intersected using pybedtools [41, 42] with the Col-0 reference TE
429 annotation, requiring 80% overlap between TE and read mapping coordinates. To determine putative
430 TE insertion sites, regions are then identified that contain independent discordant read pairs aligned
431 in an orientation facing one another at the insertion site, with their mate pairs intersecting with the
432 same TE (Figure 1). The total number of split and discordant reads intersecting the insertion site
433 and the TE is then calculated, and a TE insertion predicted where the combined number of reads
434 is greater than a threshold determined by the average sequencing depth over the whole genome
435 (1/10 coverage if coverage is greater than 10, otherwise a minimum of 2 reads). Alternatively, in the

436 absence of discordant reads mapped in orientations facing one another, the required total number of
437 split and discordant reads at the insertion site linked to the inserted TE is set higher, requiring twice
438 as many reads.

439 To identify TE absence variants with respect to the reference genome, split and discordant reads
440 separated >20 kb from one another are first removed, as 99.9% of *Arabidopsis* TEs are shorter
441 than 20 kb, and this removes split reads due to larger structural variants not related to TE diversity.
442 Col-0 reference annotation TEs that are located within the genomic region spanned by the split and
443 discordant reads are then identified. TE absence variants are predicted where at least 80% of the
444 TE sequence is spanned by a split or discordant read, and the sequencing depth within the spanned
445 region is <10% the sequencing depth of the 2 kb flanking sequence, and there are a minimum number
446 of split and discordant reads present, determined by the sequencing depth (1/10 coverage; Figure
447 1). A threshold of 80% TE sequence spanned by split or discordant reads is used, as opposed to
448 100%, to account for misannotation of TE sequence boundaries in the Col-0 reference TE annotation,
449 as well as TE fragments left behind by DNA TEs during cut-paste transposition (TE footprints) that
450 may affect the mapping of reads around annotated TE borders [43]. This was found to improve TE
451 absence detection using simulated data. Furthermore, the coverage within the spanned region may
452 be more than 10% that of the flanking sequence, but in such cases twice as many split and discordant
453 reads are required. If multiple TEs are spanned by the split and discordant reads, and the above
454 requirements are met, multiple TEs in the same region can be identified as absent with respect to the
455 reference genome. Absence variants in non-Col-0 accessions are subsequently recategorized as TE
456 insertions present in the Col-0 genome but absent from a given wild accession.

457 *TE variant refinement*

458 Once TE insertions are identified using the ‘tepid-map’ and ‘tepid-discover’ algorithms, these variants
459 can be refined if multiple related samples are analysed. The ‘tepid-refine’ algorithm is designed to
460 interrogate regions of the genome in which a TE insertion was discovered in other samples but not
461 the sample in question, and check for evidence of that TE insertion in the sample using lower read
462 count thresholds compared to the ‘tepid-discover’ step. In this way, the refine step leverages TE
463 variant information for a group of related samples to reduce false negative calls within the group. This
464 distinguishes TEPID from other similar methods for TE variant discovery utilizing short sequencing
465 reads. A file containing the coordinates of each insertion, and a list of sample names containing the
466 TE insertion must be provided to the ‘tepid-refine’ algorithm, which this can be generated using the
467 ‘merge_insertions.py’ script included in the TEPID package. Each sample is examined in regions
468 where there was a TE insertion identified in another sample in the group. If there is a sequencing
469 breakpoint within this region (no continuous read coverage spanning the region), split reads mapped
470 to this region will be extracted from the alignment file and their coordinates intersected with the TE
471 reference annotation. If there are split reads present at the variant site that are linked to the same
472 TE as was identified as an insertion at that location, this TE insertion is recorded in a new file as
473 being present in the sample in question. If there is no sequencing coverage in the queried region for
474 a sample, an “NA” call is made indicating that it is unknown whether the particular sample contains
475 the TE insertion or not.

476 While the above description relates specifically to use of TEPID for identification of TE variants in
477 *Arabidopsis* in this study, this method can be also applied to other species, with the only prerequisite
478 being the annotation of TEs in a reference genome and the availability of paired-end DNA sequencing
479 data.

480 TE variant simulation

481 To test the sensitivity and specificity of TEPID, 100 TE insertions (50 copy-paste transpositions, 50
482 cut-paste transpositions) and 100 TE absence variants were simulated in the Arabidopsis genome
483 using the RSVSim R package, version 1.7.2 [44], and synthetic reads generated from the modified
484 genome at various levels of sequencing coverage using wgsim [40] (<https://github.com/lh3/wgsim>).
485 These reads were then used to calculate the true positive, false positive, and false negative TE variant
486 discovery rates for TEPIID at various sequencing depths, by running ‘tepid-map’ and ‘tepid-discover’
487 using the simulated reads with the default parameters (Figure 1 - figure supplement 1).

488 Estimation of sensitivity

489 Previously published 100 bp paired end sequencing data for Ler (<http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/current/Ler-1/Reads/>; [22]) was downloaded and analyzed with the
490 TEPIID package to identify TE variants. Reads providing evidence for TE variants were then mapped to
491 the *de novo* assembled Ler genome [21]. To determine whether reads mapped to homologous regions
492 of the Ler and Col-0 reference genome, the *de novo* assembled Ler genome sequence between
493 mate pair mapping locations in Ler were extracted, with repeats masked using RepeatMasker with
494 RepBase-derived libraries and the default parameters (version 4.0.5, <http://www.repeatmasker.org>).
495 A blastn search was then conducted against the Col-0 genome using the following parameters:
496 ‘-max-target-seqs 1’, ‘-evalue 1e-6’ [23]. Coordinates of the top blast hit for each read location were
497 then compared with the TE variant sites identified using those reads. To estimate false negative rates
498 for TEPIID TE absence calls, Ler TE absence calls were compared with a known set of Col-0-specific
500 TE insertions, absent in Ler [16]. For TEPIID TE insertion calls, we mapped Col-0 DNA sequencing
501 reads [24] to the Ler PacBio assembly, and identified sites with read evidence reaching the TEPIID
502 threshold for a TE insertion call to be made.

503 Arabidopsis TE variant discovery

504 We ran the TEPIID, including the insertion refinement step, on previously published sequencing data
505 for 216 different Arabidopsis populations (NCBI SRA SRA012474; Schmitz et al. 2013), mapping
506 to the TAIR10 reference genome and using the TAIR9 TE annotation. The ‘–mask’ option was also
507 used to mask the mitochondrial and plastid genomes. We also ran TEPIID using previously published
508 transgenerational data for salt stress and control conditions (NCBI SRA SRP045804; [24]), using the
509 ‘–mask’ option to mask mitochondrial and plastid genomes, and the ‘–strict’ option for highly related
510 samples.

511 TE variant / SNP comparison

512 SNP information for 216 Arabidopsis accessions was obtained from the 1001 genomes data center
513 (http://1001genomes.org/data/Salk/releases/2013_24_01/; [20]). This was formatted into reference
514 (Col-0 state), alternate, or NA calls for each SNP. Accessions with both TE variant information and
515 SNP data were selected for analysis. Hierarchical clustering of accessions by SNPs as well as

516 TE variants were used to identify essentially clonal accessions, as these would skew minor allele
517 frequency calculations. A single representative from each cluster of similar accessions was kept,
518 leading to a total of 187 accessions for comparison. For each TE variant with minor allele frequency
519 greater than 3%, the nearest 300 upstream and 300 downstream SNPs with a minor allele frequency
520 greater than 3% were selected. Pairwise genotype correlations (r^2 values) for all complete cases were
521 obtained for SNP-SNP and SNP-TE variant states. r^2 values were then ordered by decreasing rank
522 and a median SNP-SNP rank value was calculated. For each of the 600 ranked surrounding positions,
523 the number of times the TE rank was greater than the SNP-SNP median rank was calculated as a
524 relative LD metric of TE to SNP. TE variants with less than 200 ranks over the SNP-SNP median
525 were classified as low-LD insertions. TE variants with ranks between 200 and 400 were classified as
526 mid-LD, while TE variants with greater than 400 ranks above their respective SNP-SNP median value
527 were classified as variants in high LD with flanking SNPs.

528 PCR validations

529 Selection of accessions to be genotyped

530 To assess the accuracy of TE variant calls in accessions with a range of sequencing depths of
531 coverage, we grouped accessions into quartiles based on sequencing depth of coverage and randomly
532 selected a total of 14 accessions for PCR validations from these quartiles. DNA was extracted for
533 these accessions using Edward's extraction protocol [45], and purified prior to PCR using AMPure
534 beads.

535 Selection of TE variants for validation and primer design

536 Ten TE insertion sites and 10 TE absence sites were randomly selected for validation by PCR
537 amplification. Only insertions and absence variants that were variable in at least two of the fourteen
538 accessions selected to be genotyped were considered. For insertion sites, primers were designed
539 to span the predicted TE insertion site. For TE absence sites, two primer sets were designed; one
540 primer set to span the TE, and another primer set with one primer annealing within the TE sequence
541 predicted to be absent, and the other primer annealing in the flanking sequence (Figure 2 - figure
542 supplement 3). Primer sequences were designed that did not anneal to regions of the genome
543 containing previously identified SNPs in any of the 216 accessions [20] or small insertions and
544 deletions, identified using lumpy-sv with the default settings [46](<https://github.com/arq5x/lumpy-sv>),
545 had an annealing temperature close to 52°C calculated based on nearest neighbor thermodynamics
546 (MeltingTemp submodule in the SeqUtils python module; [47]), GC content between 40% and 60%,
547 and contained the same base repeated not more than four times in a row. Primers were aligned to
548 the TAIR10 reference genome using bowtie2 [37] with the '-a' flag set to report all alignments, and
549 those with more than 5 mapping locations in the genome were then removed.

550 PCR

551 PCR was performed with 10 ng of extracted, purified Arabidopsis DNA using Taq polymerase. PCR
552 products were analysed by agarose gel electrophoresis. Col-0 was used as a positive control, water
553 was added to reactions as a negative control.

554 **mRNA analysis**

555 Processed mRNA data for 144 wild *Arabidopsis* accessions were downloaded from NCBI GEO
556 GSE43858 [20]. To find differential gene expression dependent on TE presence/absence variation,
557 we first filtered TE variants to include only those where the TE variant was shared by at least 5
558 accessions with RNA data available. We then grouped accessions based on TE presence/absence
559 variants, and performed a Mann-Whitney U test to determine differences in RNA transcript abundance
560 levels between the groups. We used q-value estimation to correct for multiple testing, using the R
561 qvalue package v2.2.2 with the following parameters: lambda = seq(0, 0.6, 0.05), smooth.df = 4 [48].
562 Genes were defined as differentially expressed where there was a greater than 2 fold difference in
563 expression between the groups, with a q-value less than 0.01. Gene ontology enrichment analysis
564 was performed using PANTHER (<http://pantherdb.org>).

565 **DNA methylation data analysis**

566 Processed base-resolution DNA methylation data for wild *Arabidopsis* accessions were downloaded
567 from NCBI GEO GSE43857 [20], and used to construct MySQL tables in a database.

568 **Rare variant analysis**

569 To assess the effect of rare TE variants on gene expression or DMR DNA methylation levels, we
570 tested for a burden of rare variants in the population extremes, essentially as described previously
571 [28]. For each rare TE variant near a gene or DMR, we ranked the gene expression level or DMR DNA
572 methylation level for all accessions in the population, and tallied the ranks of accessions containing a
573 rare variant. These rank counts were then binned to produce a histogram of the distribution of ranks.
574 We then fit a quadratic model to the counts data, and calculated the R^2 and p-value for the fit of the
575 model.

576 **TE variant and DMR genome-wide association analysis**

577 Accessions were subset to those with both leaf DNA methylation data and TEPID calls. Pairwise
578 correlations were performed for observed data pairs for each TE variant and a filtered set of population
579 C-DMRs, with those C-DMRs removed where more than 15% of the accessions had no coverage.
580 This amounted to a final set of 9,777 C-DMRs. Accession names were then permuted to produce
581 a randomized dataset, and pairwise correlations again calculated. This was repeated 500 times to
582 produce a distribution of expected Pearson correlation coefficients for each pairwise comparison.
583 Correlation values more extreme than any of the 500 permutations were deemed significant.

584 **Data access**

585 TEPID source code can be accessed at <https://github.com/ListerLab/TEPID>. Code and data needed
586 to reproduce this analysis can be found at <https://github.com/timoast/Arabidopsis-TE-variants>. Ler
587 TE variants are available in Figure 1 - source data 1 and 2. TE variants identified among the 216 wild
588 Arabidopsis accessions resequenced by Schmitz et al. (2013) are available in Figure 2 - source data
589 1, 2 and 3.

590 **Acknowledgments**

591 This work was supported by the Australian Research Council (ARC) Centre of Excellence program in
592 Plant Energy Biology CE140100008 (J.B., R.L.). R.L. was supported by an ARC Future Fellowship
593 (FT120100862) and Sylvia and Charles Viertel Senior Medical Research Fellowship, and work in
594 the laboratory of R.L. was funded by the Australian Research Council. T.S. was supported by the
595 Jean Rogerson Postgraduate Scholarship. S.R.E. was supported by an Australian Research Council
596 Discovery Early Career Research Award (DE150101206). We thank Robert J. Schmitz, Mathew
597 G. Lewsey, Ronan C. O’Malley, and Ian Small for their critical reading of the manuscript, and Kevin
598 Murray for his helpful comments regarding the development of TEPID.

599 **Author contributions**

600 R.L. and T.S. designed the research project. R.L. and J.B. supervised research. T.S. developed and
601 tested TEPID. J.C. performed PCR validations of TE variants. T.S. and S.R.E. performed bioinformatic
602 analysis. Y.K. provided statistical guidance. R.L., T.S., J.B. and S.R.E. prepared the manuscript.

603 **Competing financial interests**

604 The authors declare no competing financial interests.

605 **References**

- 606 [1] Thomas Wicker et al. "A unified classification system for eukaryotic transposable elements." In:
607 *Nature Reviews Genetics* 8.12 (Dec. 2007), pp. 973–982. DOI: [10.1038/nrg2165](https://doi.org/10.1038/nrg2165).
- 608 [2] Assaf Zemach et al. "The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltrans-
609 ferases to Access H1-Containing Heterochromatin". In: *Cell* 153.1 (Mar. 2013), pp. 193–205.
610 DOI: [10.1016/j.cell.2013.02.033](https://doi.org/10.1016/j.cell.2013.02.033).
- 611 [3] Marjori A Matzke and Rebecca A Mosher. "RNA-directed DNA methylation: an epigenetic
612 pathway of increasing complexity". In: *Nature Reviews Genetics* 15.6 (May 2014), pp. 394–408.
613 DOI: [10.1038/nrg3683](https://doi.org/10.1038/nrg3683).
- 614 [4] Marie Mirouze et al. "Selective epigenetic control of retrotransposition in Arabidopsis." In: *Nature*
615 461.7262 (Sept. 2009), pp. 427–430. DOI: [10.1038/nature08328](https://doi.org/10.1038/nature08328).
- 616 [5] Asuka Miura et al. "Mobilization of transposons by a mutation abolishing full DNA methylation in
617 Arabidopsis". In: *Nature* 411.6834 (2001), pp. 212–214. DOI: [10.1038/35075612](https://doi.org/10.1038/35075612).
- 618 [6] Hidetoshi Saze, Ortrun Mittelsten Scheid, and Jerzy Paszkowski. "Maintenance of CpG methy-
619 lation is essential for epigenetic inheritance during plant gametogenesis". In: *Nature Genetics*
620 34.1 (Mar. 2003), pp. 65–69. DOI: [10.1038/ng1138](https://doi.org/10.1038/ng1138).
- 621 [7] Zachary Lippman et al. "Role of transposable elements in heterochromatin and epigenetic
622 control." In: *Nature* 430.6998 (July 2004), pp. 471–476. DOI: [10.1038/nature02651](https://doi.org/10.1038/nature02651).
- 623 [8] Jeffrey A Jeddeloh, Trevor L Stokes, and Eric J Richards. "Maintenance of genomic methylation
624 requires a SWI2/SNF2-like protein". In: *Nature Genetics* 22.1 (1999), pp. 94–97. DOI: [10.1038/8803](https://doi.org/10.1038/8803).
- 625 [9] Clémentine Vitte et al. "The bright side of transposons in crop evolution." In: *Briefings in
626 Functional Genomics* 13.4 (July 2014), pp. 276–295. DOI: [10.1093/bfgp/elu002](https://doi.org/10.1093/bfgp/elu002).
- 627 [10] Elizabeth Hénaff et al. "Extensive amplification of the E2F transcription factor binding sites
628 by transposons during evolution of Brassica species." In: *The Plant Journal* 77.6 (Mar. 2014),
629 pp. 852–862. DOI: [10.1111/tpj.12434](https://doi.org/10.1111/tpj.12434).
- 630 [11] Anthony Bolger et al. "The genome of the stress-tolerant wild tomato species". In: *Nature
631 Genetics* 46.9 (July 2014), pp. 1034–1038. DOI: [10.1038/ng.3046](https://doi.org/10.1038/ng.3046).
- 632 [12] Hidetaka Ito et al. "An siRNA pathway prevents transgenerational retrotransposition in plants
633 subjected to stress". In: *Nature* 472.7341 (Mar. 2011), pp. 115–119. DOI: [10.1038/nature09861](https://doi.org/10.1038/nature09861).
- 634 [13] Irina Makarevitch et al. "Transposable Elements Contribute to Activation of Maize Genes in
635 Response to Abiotic Stress". In: *PLoS Genetics* 11.1 (Jan. 2015), e1004915. DOI: [10.1371/journal.pgen.1004915.s016](https://doi.org/10.1371/journal.pgen.1004915.s016).

- 638 [14] Paul Bundock and Paul Hooykaas. "An *Arabidopsis* hAT-like transposase is essential for plant
639 development." In: *Nature* 436.7048 (July 2005), pp. 282–284. DOI: [10.1038/nature03667](https://doi.org/10.1038/nature03667).
- 640 [15] Jun Cao et al. "Whole-genome sequencing of multiple *Arabidopsis thaliana* populations." In:
641 *Nature Publishing Group* 43.10 (Oct. 2011), pp. 956–963. DOI: [10.1038/ng.911](https://doi.org/10.1038/ng.911).
- 642 [16] Leandro Quadrana et al. "The *Arabidopsis thaliana* mobilome and its impact at the species
643 level." In: *eLife* 5 (2016), p. 6919. DOI: [10.7554/eLife.15716](https://doi.org/10.7554/eLife.15716).
- 644 [17] Djie Tjwan Thung et al. "Mobster: accurate detection of mobile element insertions in next
645 generation sequencing data". In: (Oct. 2014), pp. 1–11. DOI: [10.1186/s13059-014-0488-x](https://doi.org/10.1186/s13059-014-0488-x).
- 646 [18] Sofia M. C. Robb et al. "The use of RelocaTE and unassembled short reads to produce high-
647 resolution snapshots of transposable element generated diversity in rice". In: *G3: Genes /*
648 *Genomes / Genetics* (2013). DOI: [10.1534/g3.112.005348/-/DC1](https://doi.org/10.1534/g3.112.005348/-/DC1).
- 649 [19] Elizabeth Hénaff et al. "Jitterbug: somatic and germline transposon insertion detection at single-
650 nucleotide resolution". In: *BMC Genomics* 16.1 (Oct. 2015), pp. 1–16. DOI: [10.1186/s12864-015-1975-5](https://doi.org/10.1186/s12864-015-1975-5).
- 652 [20] Robert J Schmitz et al. "Patterns of population epigenomic diversity". In: *Nature* 495.7440 (Mar.
653 2013), pp. 193–198. DOI: [10.1038/nature11968](https://doi.org/10.1038/nature11968).
- 654 [21] Chen-Shan Chin et al. "Nonhybrid, finished microbial genome assemblies from long-read SMRT
655 sequencing data". In: *Nature Methods* 10.6 (May 2013), pp. 563–569. DOI: [10.1038/nmeth.2474](https://doi.org/10.1038/nmeth.2474).
- 656 [22] Korbinian Schneeberger et al. "Reference-guided assembly of four diverse *Arabidopsis thaliana*
657 genomes". In: *Proceedings of the National Academy of Sciences of the United States of America*
658 108.25 (2011), pp. 10249–10254. DOI: [10.1073/pnas.1107739108](https://doi.org/10.1073/pnas.1107739108).
- 659 [23] Christiam Camacho et al. "BLAST+: architecture and applications." In: *BMC Bioinformatics* 10.1
660 (2009), p. 421. DOI: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421).
- 661 [24] Caifu Jiang et al. "Environmentally responsive genome-wide accumulation of de novo *Arabidopsis*
662 *thaliana* mutations and epimutations." In: *Genome Research* 24.11 (Nov. 2014), pp. 1821–
663 1829. DOI: [10.1101/gr.177659.114](https://doi.org/10.1101/gr.177659.114).
- 664 [25] Alessandra M Sullivan et al. "Mapping and dynamics of regulatory DNA and transcription factor
665 networks in *A. thaliana*." In: *Cell* 8.6 (Sept. 2014), pp. 2015–2030. DOI: [10.1016/j.celrep.2014.08.019](https://doi.org/10.1016/j.celrep.2014.08.019).
- 667 [26] Stefan Grob, Marc W Schmid, and Ueli Grossniklaus. "Hi-C Analysis in *Arabidopsis* Identifies
668 the KNOT, a Structure with Similarities to the flamenco Locus of *Drosophila*". In: *Molecular Cell*
669 (Aug. 2014), pp. 1–16. DOI: [10.1016/j.molcel.2014.07.009](https://doi.org/10.1016/j.molcel.2014.07.009).

- 670 [27] Matthew W Horton et al. "Genome-wide patterns of genetic variation in worldwide Arabidopsis
671 thaliana accessions from the RegMap panel". In: *Nature Publishing Group* 44.2 (Feb. 2012),
672 pp. 212–216. DOI: [10.1038/ng.1042](https://doi.org/10.1038/ng.1042).
- 673 [28] Jing Zhao et al. "A Burden of Rare Variants Associated with Extremes of Gene Expression
674 in Human Peripheral Blood". In: *The American Journal of Human Genetics* 98.2 (Feb. 2016),
675 pp. 299–309. DOI: [10.1016/j.ajhg.2015.12.023](https://doi.org/10.1016/j.ajhg.2015.12.023).
- 676 [29] Guodong Wang et al. "A genome-wide functional investigation into the roles of receptor-like
677 proteins in Arabidopsis." In: *Plant Physiolsy* 147.2 (June 2008), pp. 503–517. DOI: [10.1104/pp.108.119487](https://doi.org/10.1104/pp.108.119487).
- 679 [30] María José Aranzana et al. "Genome-Wide Association Mapping in Arabidopsis Identifies
680 Previously Known Flowering Time and Pathogen Resistance Genes". In: *PLoS Genetics* 1.5
681 (2005), e60–9. DOI: [10.1371/journal.pgen.0010060](https://doi.org/10.1371/journal.pgen.0010060).
- 682 [31] Akira Katoh et al. "Early steps in the biosynthesis of NAD in Arabidopsis start with aspartate
683 and occur in the plastid." In: *Plant Physiolsy* 141.3 (July 2006), pp. 851–857. DOI: [10.1104/pp.106.081091](https://doi.org/10.1104/pp.106.081091).
- 685 [32] Ryan Lister et al. "Highly integrated single-base resolution maps of the epigenome in Arabidopsis."
686 In: *Cell* 133.3 (May 2008), pp. 523–536. DOI: [10.1016/j.cell.2008.03.029](https://doi.org/10.1016/j.cell.2008.03.029).
- 687 [33] Shawn J Cokus et al. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals
688 DNA methylation patterning". In: *Nature* 452.7184 (Feb. 2008), pp. 215–219. DOI: [10.1038/nature06745](https://doi.org/10.1038/nature06745).
- 690 [34] Xiaoyu Zhang et al. "Genome-wide High-Resolution Mapping and Functional Analysis of DNA
691 Methylation in Arabidopsis". In: *Cell* 126.6 (Sept. 2006), pp. 1189–1201. DOI: [10.1016/j.cell.2006.08.003](https://doi.org/10.1016/j.cell.2006.08.003).
- 693 [35] Daniel Zilberman et al. "Genome-wide analysis of Arabidopsis thaliana DNA methylation
694 uncovers an interdependence between methylation and transcription." In: *Nature Genetics* 39.1
695 (Jan. 2007), pp. 61–69. DOI: [10.1038/ng1929](https://doi.org/10.1038/ng1929).
- 696 [36] Eric J Richards. "Inherited epigenetic variation—revisiting soft inheritance." In: *Nature Reviews
697 Genetics* 7.5 (May 2006), pp. 395–401. DOI: [10.1038/nrg1834](https://doi.org/10.1038/nrg1834).
- 698 [37] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature
699 Methods* 9.4 (Mar. 2012), pp. 357–359. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923).
- 700 [38] Gregory G Faust and Ira M Hall. "SAMBLASTER: fast duplicate marking and structural variant
701 read extraction." In: *Bioinformatics* 30.17 (Sept. 2014), pp. 2503–2505. DOI: [10.1093/bioinformatics/btu314](https://doi.org/10.1093/bioinformatics/btu314).

- 703 [39] Gregory G Faust and Ira M Hall. "YAHA: fast and flexible long-read alignment with optimal
704 breakpoint detection." In: *Bioinformatics* 28.19 (Oct. 2012), pp. 2417–2424. DOI: [10.1093/bioinformatics/bts456](https://doi.org/10.1093/bioinformatics/bts456).
- 706 [40] Heng Li et al. "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics* 25.16
707 (Aug. 2009), pp. 2078–2079. DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
- 708 [41] Ryan K Dale, Brent S Pedersen, and Aaron R Quinlan. "Pybedtools: a flexible Python library
709 for manipulating genomic datasets and annotations." In: *Bioinformatics* 27.24 (Dec. 2011),
710 pp. 3423–3424. DOI: [10.1093/bioinformatics/btr539](https://doi.org/10.1093/bioinformatics/btr539).
- 711 [42] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic
712 features." In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. DOI: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).
- 713 [43] Ronald H Plasterk. "The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*."
714 In: *The EMBO Journal* 10.7 (July 1991), pp. 1919–1925.
- 715 [44] Christoph Bartenhagen and Martin Dugas. "RSVSim: an R/Bioconductor package for the
716 simulation of structural variations." In: *Bioinformatics* 29.13 (July 2013), pp. 1679–1681. DOI:
717 [10.1093/bioinformatics/btt198](https://doi.org/10.1093/bioinformatics/btt198).
- 718 [45] K Edwards, C Johnstone, and C Thompson. "A simple and rapid method for the preparation of
719 plant genomic DNA for PCR analysis." In: *Nucleic Acids Research* 19.6 (Mar. 1991), p. 1349.
- 720 [46] Ryan M Layer et al. "LUMPY: a probabilistic framework for structural variant discovery." In:
721 *Genome Biology* 15.6 (2014), R84. DOI: [10.1186/gb-2014-15-6-r84](https://doi.org/10.1186/gb-2014-15-6-r84).
- 722 [47] Peter J A Cock et al. "Biopython: freely available Python tools for computational molecular
723 biology and bioinformatics." In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. DOI: [10.1093/bioinformatics/btp163](https://doi.org/10.1093/bioinformatics/btp163).
- 725 [48] John D Storey and Robert Tibshirani. "Statistical significance for genomewide studies." In:
726 *Proceedings of the National Academy of Sciences of the United States of America* 100.16 (Aug.
727 2003), pp. 9440–9445. DOI: [10.1073/pnas.1530509100](https://doi.org/10.1073/pnas.1530509100).

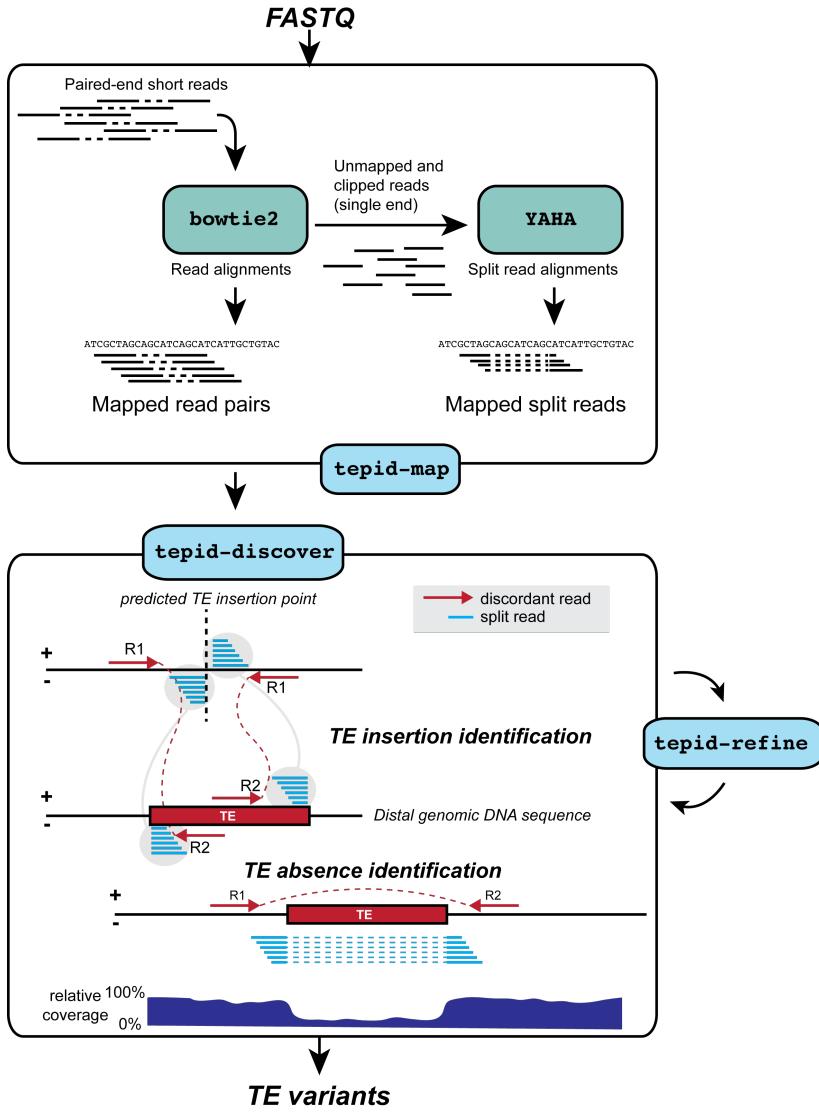


Figure 1: TE variant discovery pipeline

Principle of TE variant discovery using split and discordant read mapping positions. Paired end reads are first mapped to the reference genome using Bowtie2 [37]. Soft-clipped or unmapped reads are then extracted from the alignment and re-mapped using Yaha, a split read mapper [39]. All read alignments are then used by TEPID to discover TE variants relative to the reference genome, in the ‘tepid-discover’ step. When analyzing groups of related samples, these variants can be further refined using the ‘tepid-refine’ step, which examines in more detail the genomic regions where there was a TE variant identified in another sample, and calls the same variant for the sample in question using lower read count thresholds as compared to the ‘tepid-discover’ step, in order to reduce false negative variant calls within a group of related samples.

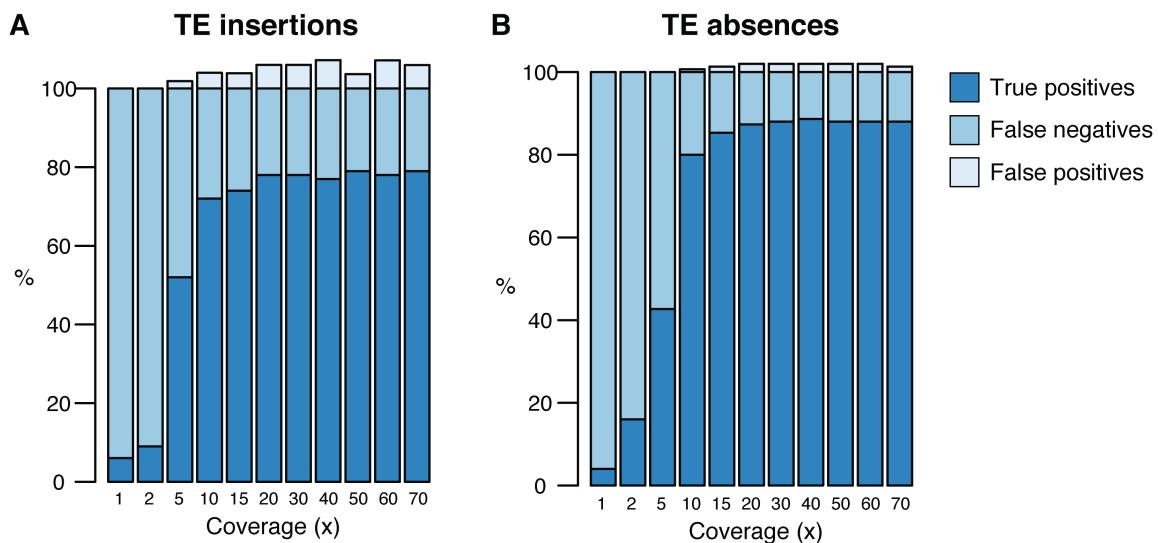


Figure 1: figure supplement 1

737 Testing of the TEPIP pipeline using simulated TE variants in the Arabidopsis Col-0 genome (TAIR10),
 738 for a range of sequencing coverage levels. TE insertions (A) and TE absence calls (B).

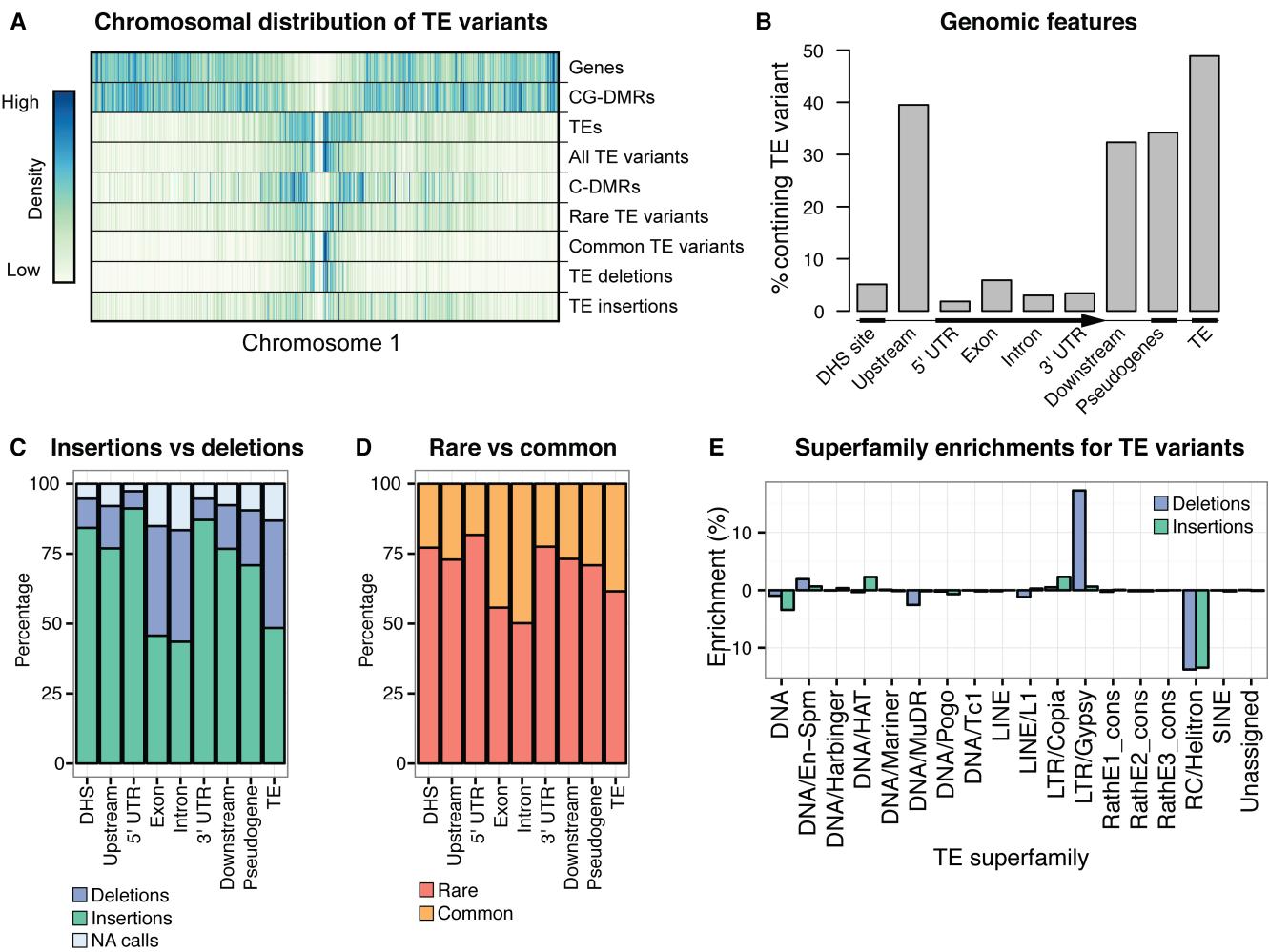


Figure 2: Extensive novel genetic diversity uncovered by TE variant analysis

- (A) Distribution of identified TE variants on chromosome 1, with distributions of all Col-0 genes, Col-0 TEs, and population DMRs.
- (B) Frequency of TE variants at different genomic features.
- (C) Proportion of TE variants within each genomic feature classified as deletions or insertions.
- (D) Proportion of TE variants within each genomic feature classified as rare or common.
- (E) Enrichment and depletion of TE variants categorized by TE superfamily compared to the expected frequency due to genomic occurrence.

TE calls due to TEPID refinement step

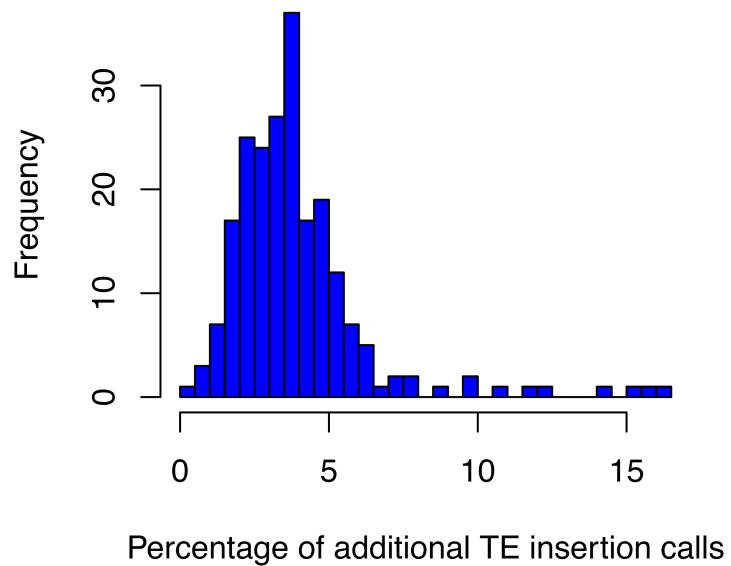


Figure 2: figure supplement 1

746 Number of additional TE insertion calls made due to the TEPID refinement step for each accession in
747 the population.

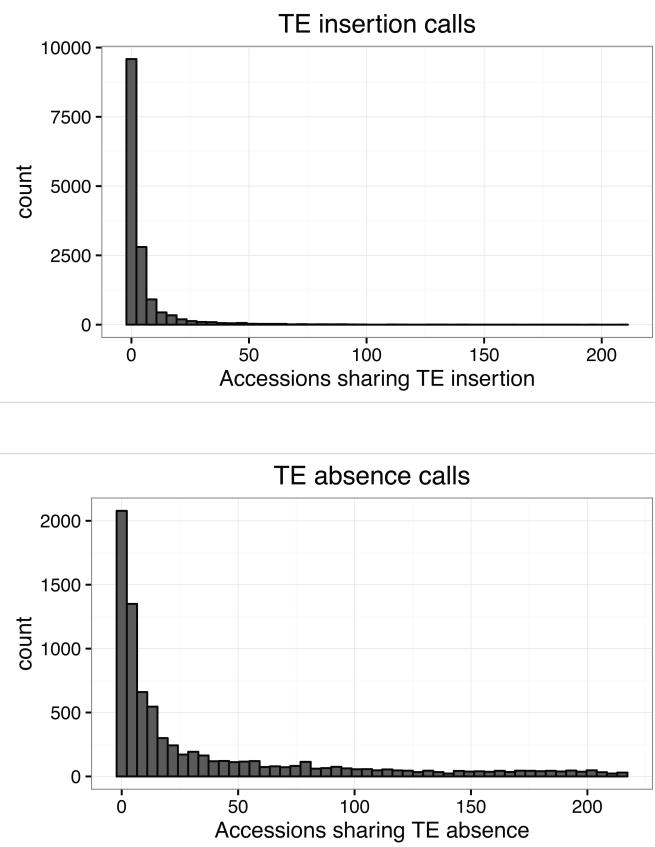


Figure 2: figure supplement 2

748 Minor allele frequency distribution for TE variants.

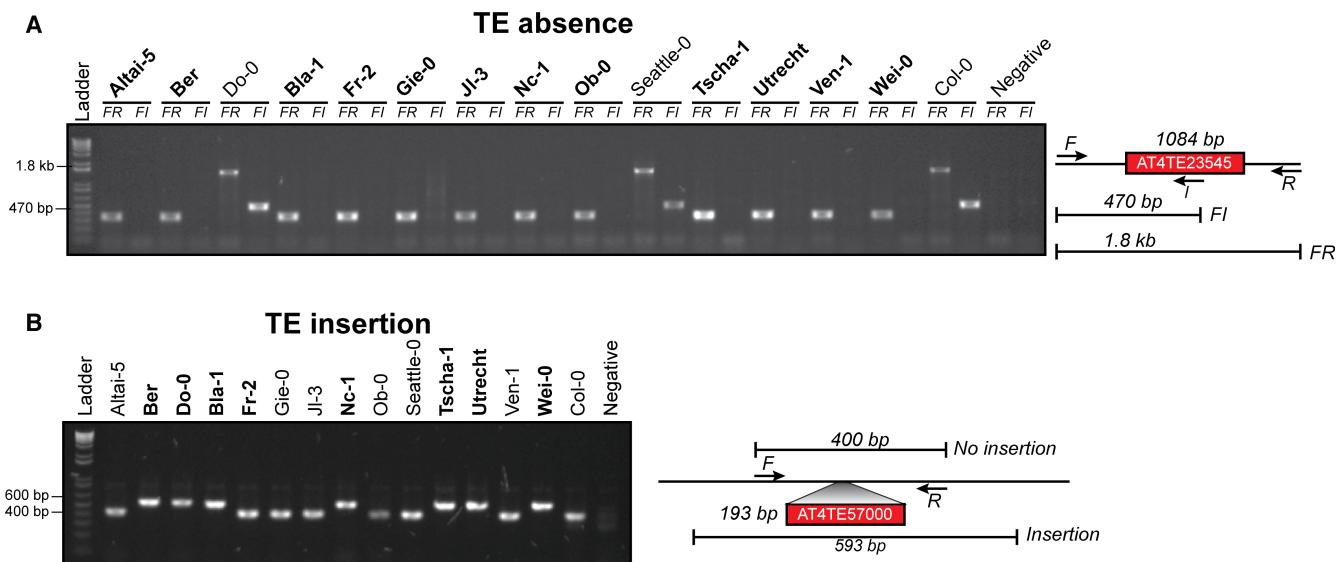


Figure 2: figure supplement 3

- 749 (A) PCR validations for a TE absence variant. Accessions that were predicted to contain a TE
 750 insertion or TE absence are marked in bold. Two primer sets were used; forward (F) and reverse
 751 (R) or internal (I). Accessions with a TE absence will not produce the FI band and produce a
 752 shorter FR product, with the change in size matching the size of the deleted TE.
- 753 (B) PCR validations for a TE insertion variant. One primer set was used, spanning the TE insertion
 754 site. A band shift of approximately 200 bp can be seen, corresponding to the size of the inserted
 755 TE.

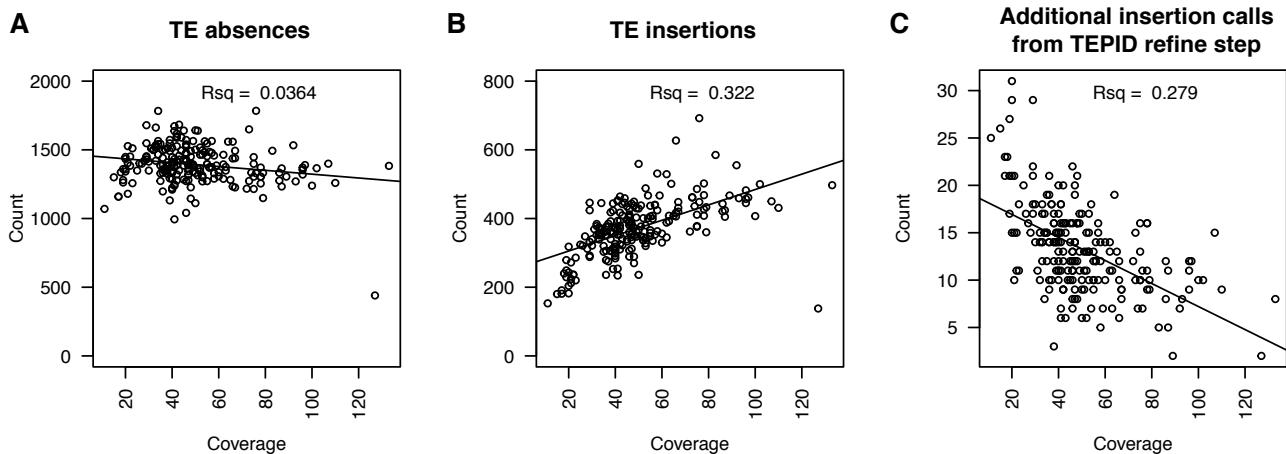
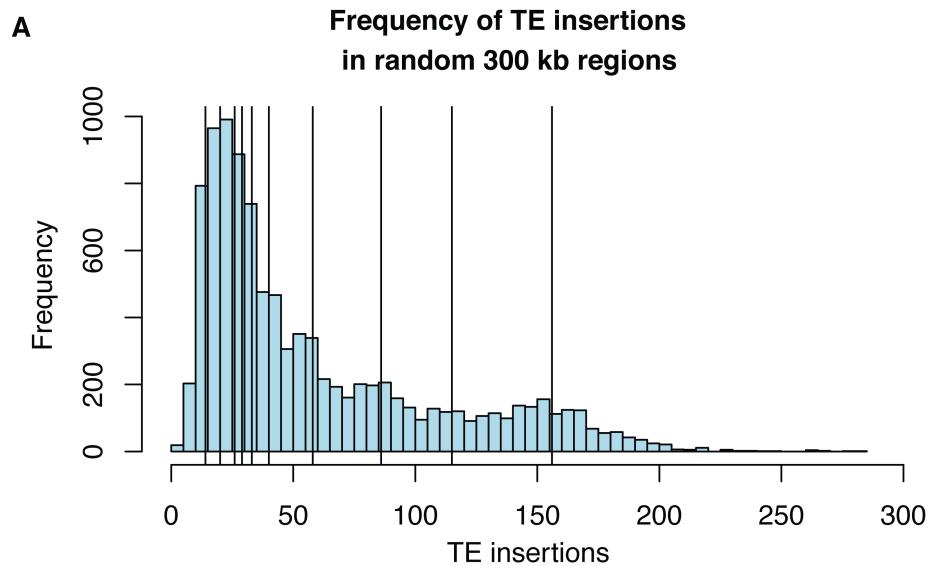


Figure 2: figure supplement 4

- 756 (A) Number of TE absence variants identified versus the sequencing depth of coverage for each
757 accession.
- 758 (B) Number of TE insertion variants identified versus the sequencing depth of coverage for each
759 accession.
- 760 (C) Number of additional TE insertion calls made due to the TEPID refinement step versus se-
761 quencing depth of coverage for all accessions.



B

chr	start	stop	KEE	TE variants	p-value
chr1	6900000	7200000	kee1	29	0.6304
chr2	4025000	4325000	kee2	156	0.0675
chr3	1800000	2100000	kee3	33	0.5672
chr3	2950000	3250000	kee4	14	0.9172
chr3	16537500	16837500	kee5	115	0.1659
chr3	22375000	22675000	kee6	40	0.4927
chr4	10900000	11200000	kee7	58	0.3589
chr4	15387500	15687500	kee8	26	0.6824
chr5	4612500	4912500	kee9	20	0.802
chr5	10162500	10462500	kee10	86	0.2455

Figure 2: figure supplement 5. Frequency of TE insertion in the *KNOT* region

- 762 (A) Number of TE insertion variants within each 300 kb *KNOT ENGAGED ELEMENT (KEE)*,
 763 vertical lines) and the number of TE insertion variants found in 10,000 randomly selected 300
 764 kb windows (histogram).
- 765 (B) Table showing number of TE insertion variants within each *KEE* region, and the associated
 766 p-value determined by resampling 10,000 times.

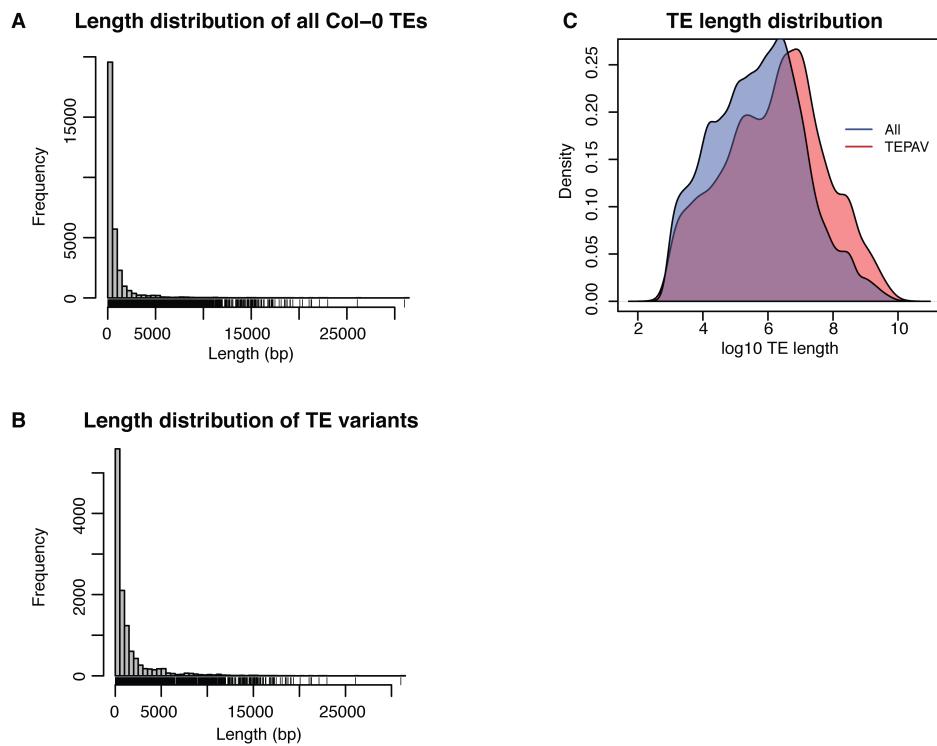


Figure 2: figure supplement 6. Length distribution for all Col-0 TEs and all TE variants

- 767 (A) Histogram showing lengths of all annotated TEs in the Col-0 reference genome.
- 768 (B) Histogram showing lengths of all TE variants.
- 769 (C) Density distribution of \log_{10} TE length for all Col-0 TEs (red) and TE variants (blue).

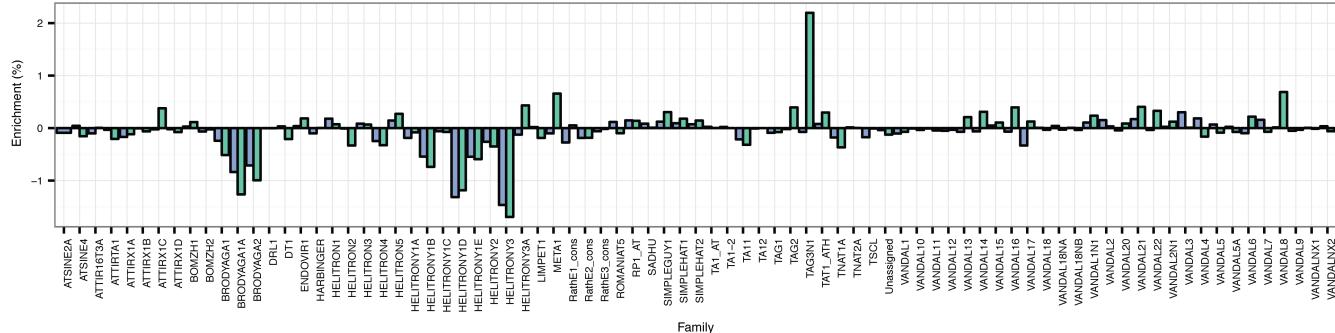
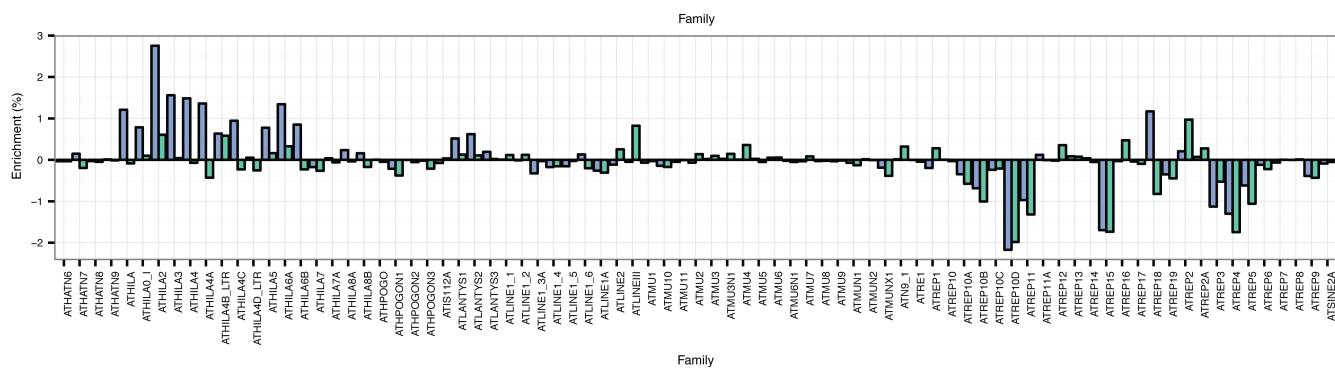
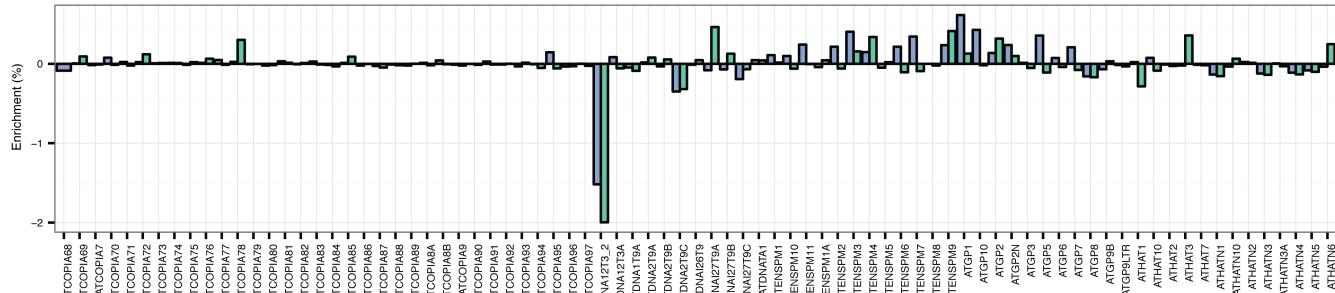
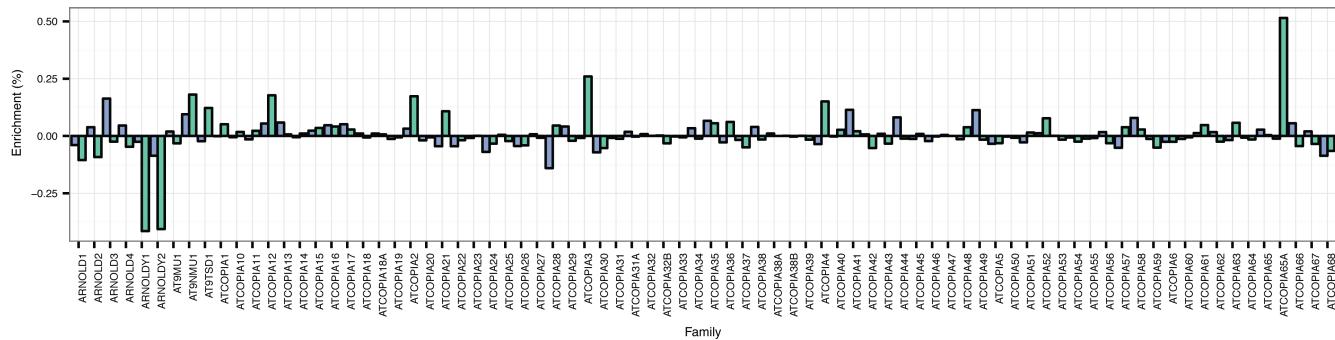


Figure 2: figure supplement 7

770 TE family enrichments and depletions for TE insertions and TE deletions.

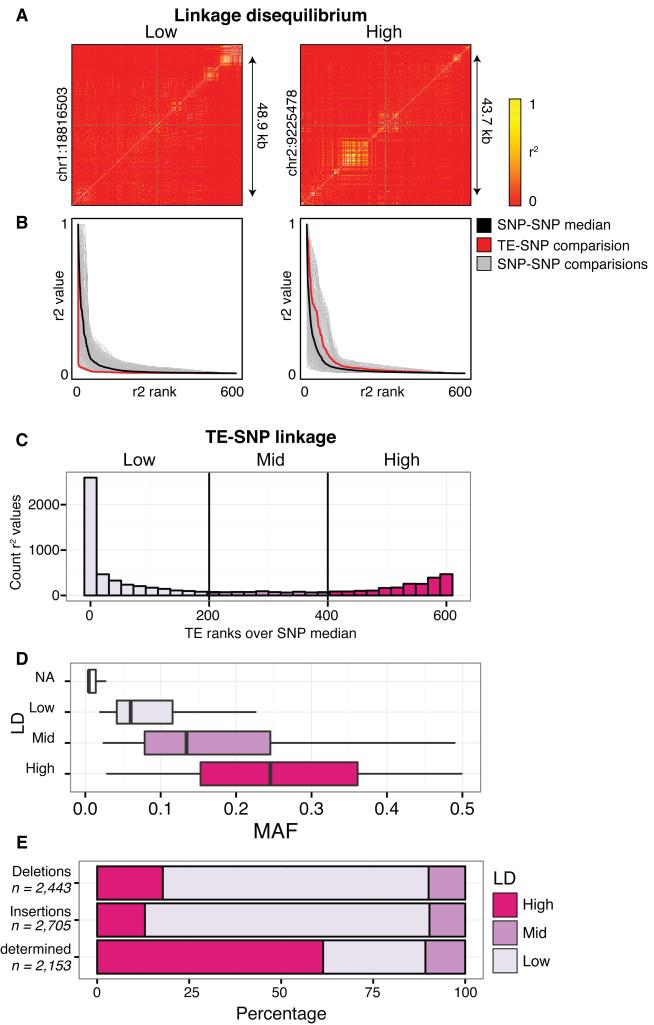


Figure 3: Patterns of TE-SNP linkage

- (A) r^2 correlation matrices for individual representative high and low-LD TE variants showing the background level of SNP-SNP linkage.
- (B) Rank order plots for individual representative high and low-LD TE variants (matching those shown in A). Red line indicates the median r^2 value for each rank across SNP-based values. Blue line indicates r^2 values for TE-SNP comparisons. Grey lines indicate all individual SNP-SNP comparisons.
- (C) Histogram of the number of TE r^2 ranks (0-600) that are above the SNP-based median r^2 value for testable TE variants.
- (D) Boxplots showing distribution of minor allele frequencies for each LD category. Boxes represent the interquartile range (IQR) from quartile 1 to quartile 3. Boxplot upper whiskers represent the maximum value, or the upper value of the quartile 3 plus 1.5 times the IQR (whichever is smaller). Boxplot lower whisker represents the minimum value, or the lower value of the quartile 1 minus 1.5 times the IQR (whichever is larger).
- (E). Proportion of TE insertions, TE deletions, and unclassified TE variants in each LD category.

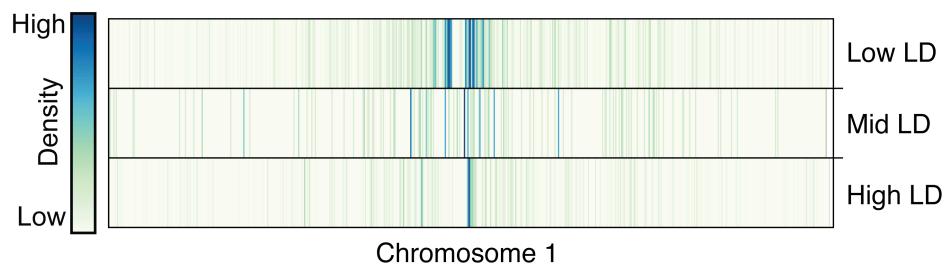


Figure 3: figure supplement 1

785 Distribution of TE variants across chromosome 1 for each LD category (high, mid, low).

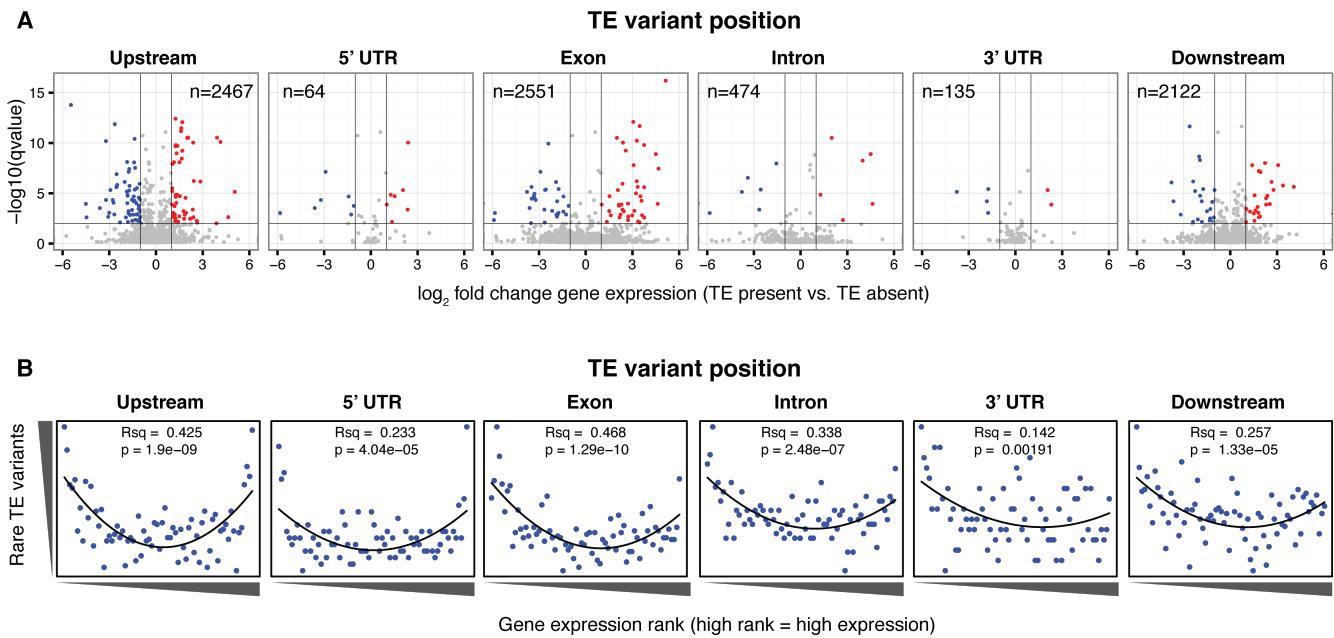


Figure 4: Differential transcript abundance associated with TE variant presence/absence

- 786 (A) Volcano plots showing transcript abundance differences for genes associated with TE insertion
 787 variants at different positions, indicated in the plot titles. Genes with significantly different
 788 transcript abundance in accessions with a TE insertion compared to accessions without a TE
 789 insertion are colored blue (lower transcript abundance in accessions containing TE insertion) or
 790 red (higher transcript abundance in accessions containing TE insertion). Vertical lines indicate
 791 ± 2 fold change in FPKM. Horizontal line indicates the 1% FDR.
- 792 (B) Relationship between TE rare variant counts and gene expression rank. Plot shows the
 793 cumulative number of rare TE variants in equal-sized bins for gene expression ranks, from the
 794 lowest-ranked accession (left) to the highest-ranked accession (right). Lines indicate the fit of a
 795 quadratic model.

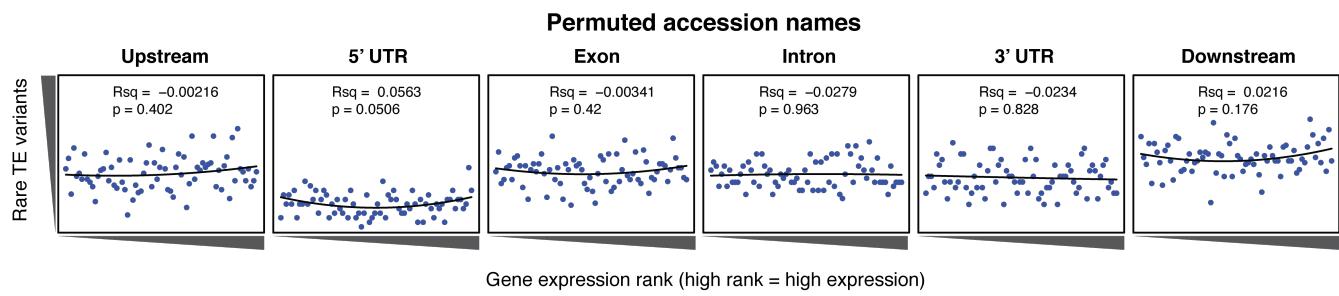


Figure 4: figure supplement 1

796 Relationship between rare TE variants and gene expression rank as for Figure 4B, for permuted TE
 797 variants.

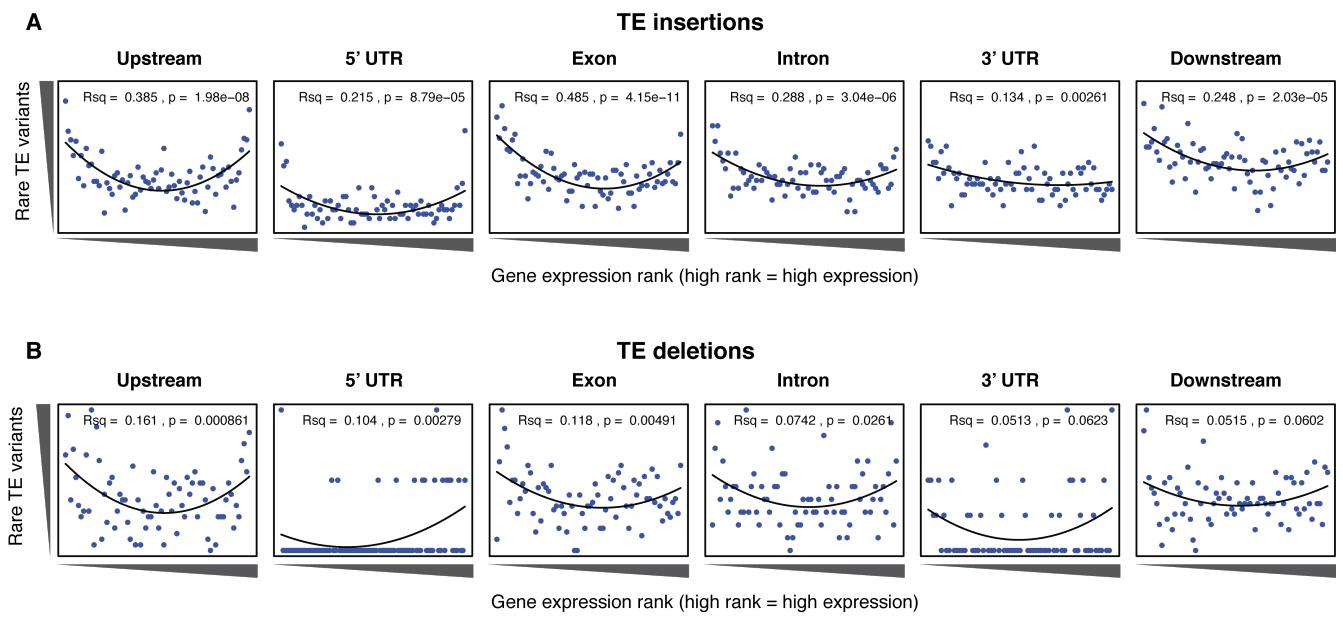


Figure 4: figure supplement 2

798 Relationship between rare TE variants and gene expression rank as for Figure 4B, for TE insertions
 799 (A) and TE deletions (B) separately.

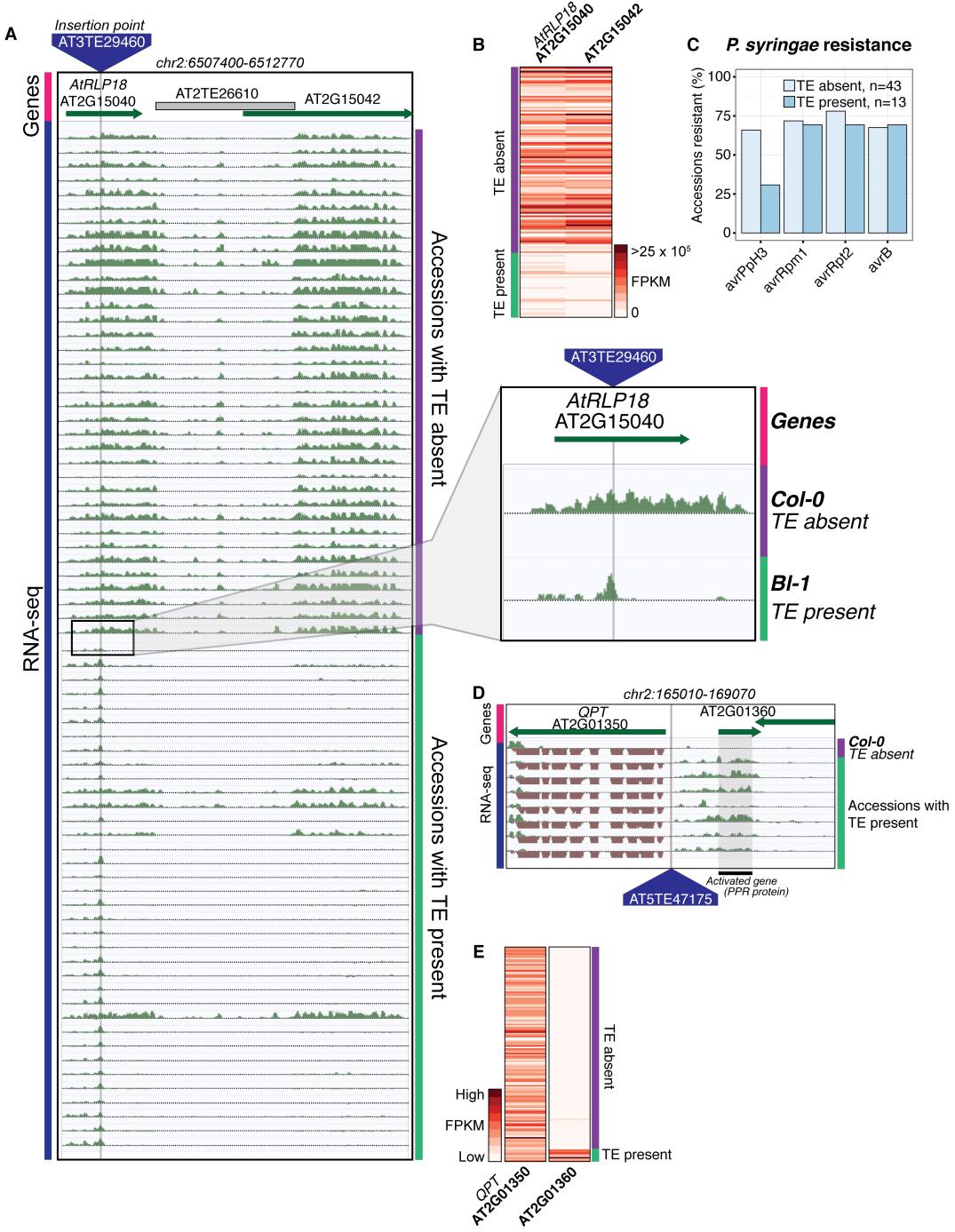


Figure 5: Effects of TE variants on local gene expression

- 800 (A) Genome browser representation of RNA-seq data for genes *AtRLP18* (AT2G15040) and a
801 leucine-rich repeat family protein (AT2G15042) for Db-1, containing a TE insertion within the
802 exon of the gene *AtRLP18*, and for a Col-0 (not containing the TE insertion within the exon of
803 *AtRLP18*). Inset shows magnified view of the TE insertion site.
- 804 (B) Heatmap showing *AtRLP18* and AT2G15042 RNA-seq FPKM values for all accessions.
- 805 (C) Percentage of accessions with resistance to *Pseudomonas syringae* transformed with different
806 *avr* genes, for accessions containing or not containing a TE insertion in* *AtRLP18*.*

- 807 (D) Genome browser representation of RNA-seq data for a PPR protein-encoding gene
808 (AT2G01360) and *QPT* (AT2G01350), showing transcript abundance for these genes in
809 accessions containing a TE insertion variant in the upstream region of these genes.
- 810 (E) Heatmap representation of RNA-seq FPKM values for *QPT* and a gene encoding a PPR protein
811 (AT2G01360), for all accessions. Note that scales are different for the two heatmaps, due to the
812 higher transcript abundance of *QPT* compared to AT2G01360. Scale maximum for AT2G01350
813 is 3.1×10^5 , and for AT2G01360 is 5.9×10^4 .

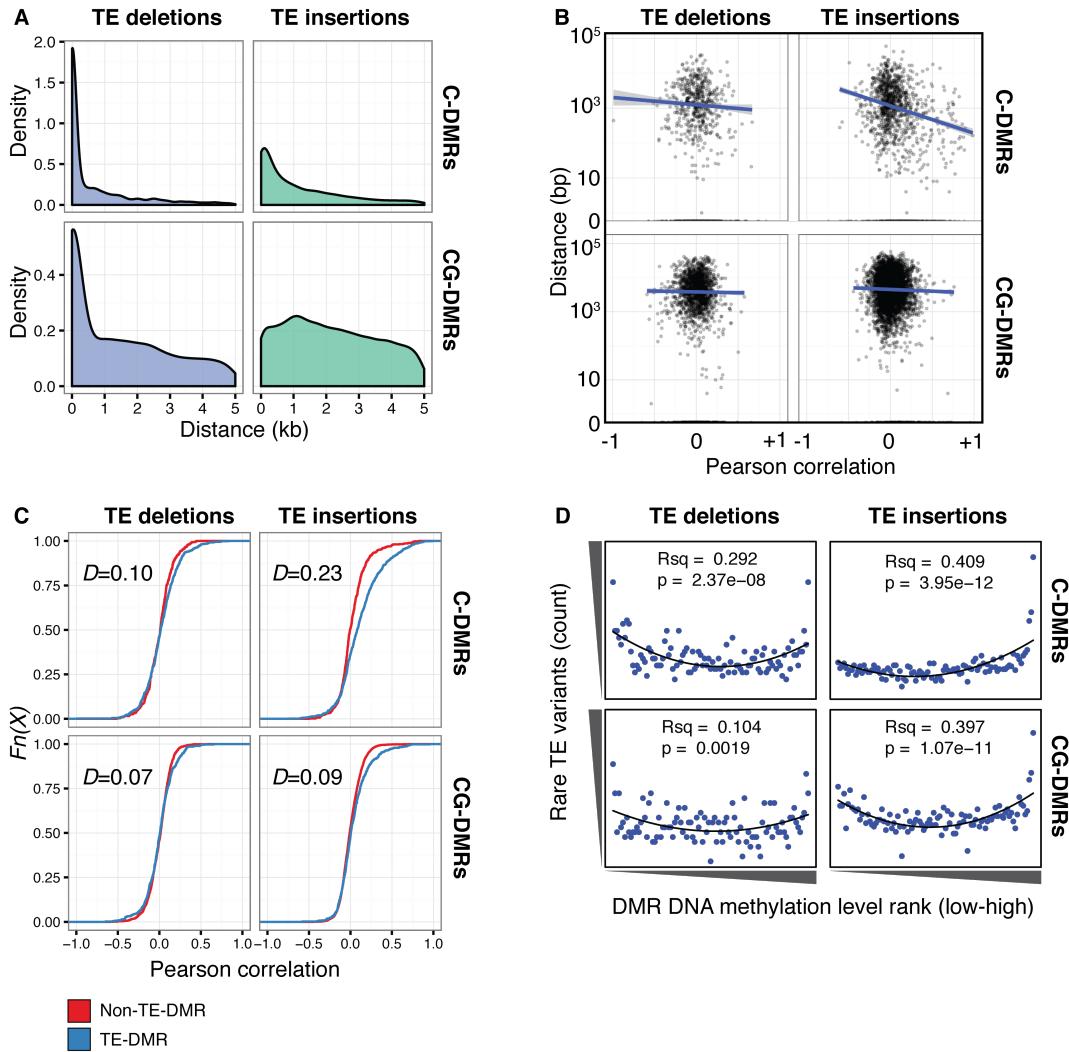


Figure 6: TE variants are associated with nearby DMR methylation levels

- (A) Distribution of distances from TE variants to the nearest population DMR, for TE deletions and TE insertion, C-DMRs and CG-DMRs.
- (B) Pearson correlation between DMR DNA methylation level and TE presence/absence, for all DMRs and their closest TE variant, versus the distance from the DMR to the TE variant (log scale). Blue lines show a linear regression between the correlation coefficients and the \log_{10} distance to the TE variant.
- (C) Empirical cumulative distribution of Pearson correlation coefficients between TE presence/absence and DMR methylation level for TE insertions, TE deletions, C-DMRs and CG-DMRs. The Kolmogorov–Smirnov statistic is shown in each plot, indicated by D .
- (D) Relationship between rare TE variant counts and nearby DMR DNA methylation level ranks, for TE insertions, deletions, C-DMRs, and CG-DMRs. Plot shows the cumulative number of rare TE variants in equal-sized bins of DMR methylation level ranks, from the lowest ranked accession (left) to the highest ranked accession (right). Lines indicate the fit of a quadratic model, and the corresponding R^2 and p values are shown in each plot.

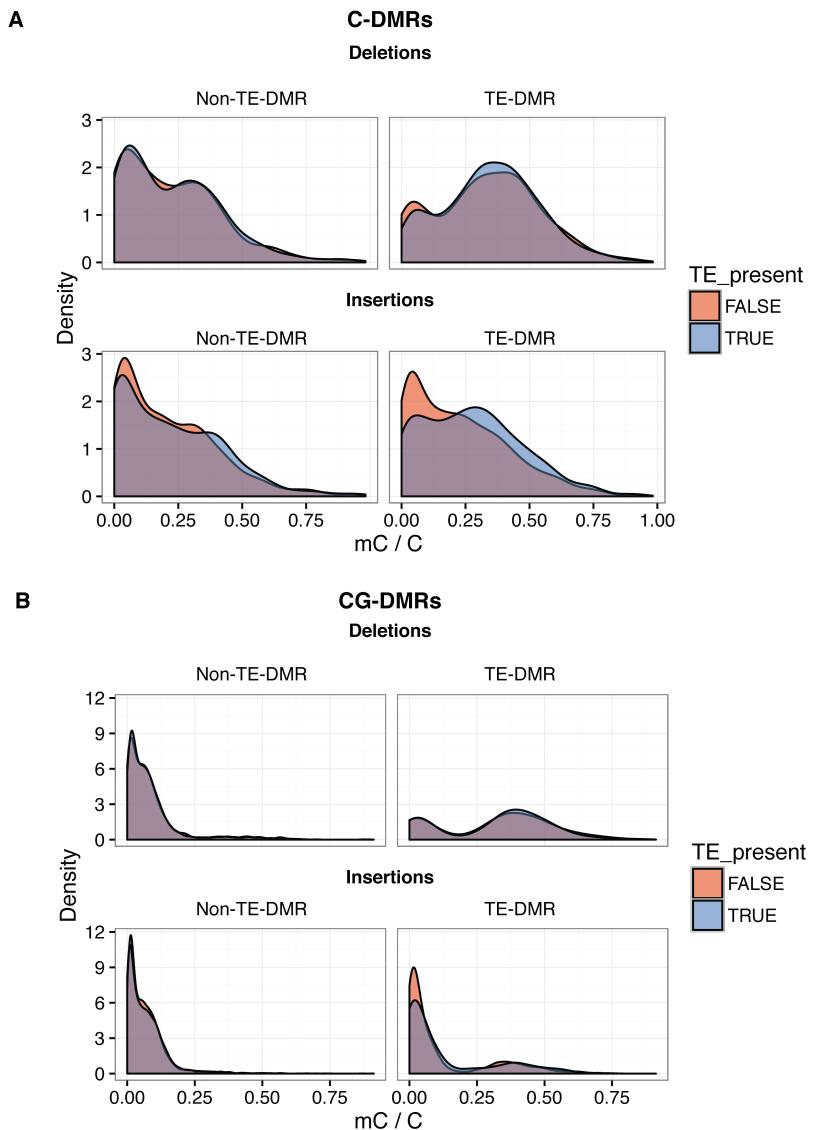


Figure 6: figure supplement 1

- 828 (A) DNA methylation density distribution at C-DMRs within 1 kb of a TE variant (TE-DMRs) or
 829 further than 1 kb from a TE variant (non-TE-DMRs), in the presence or absence of the TE, for
 830 TE insertions and TE deletions.
- 831 (B) As for A, for CG-DMRs.

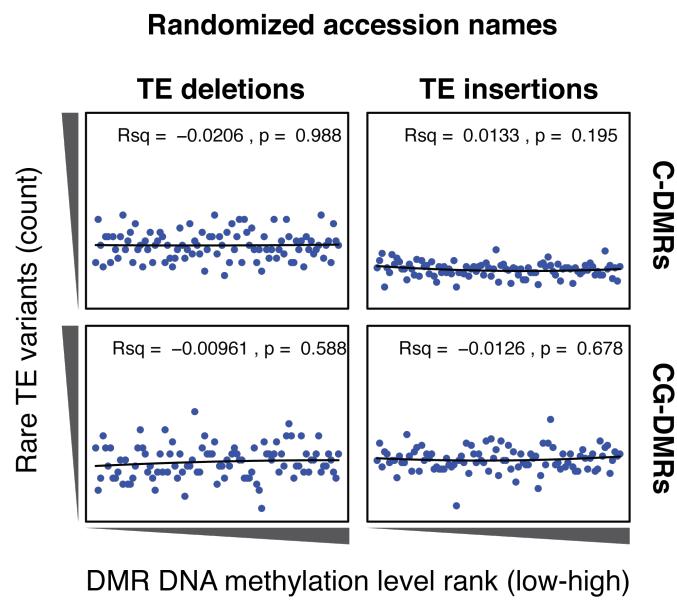


Figure 6: figure supplement 2

832 Cumulative number DMR methylation level ranks for DMRs near rare TE variants with accessions
 833 selected at random. Lines indicate the fit of a quadratic model, and the corresponding R^2 and p values
 834 are shown in each plot.

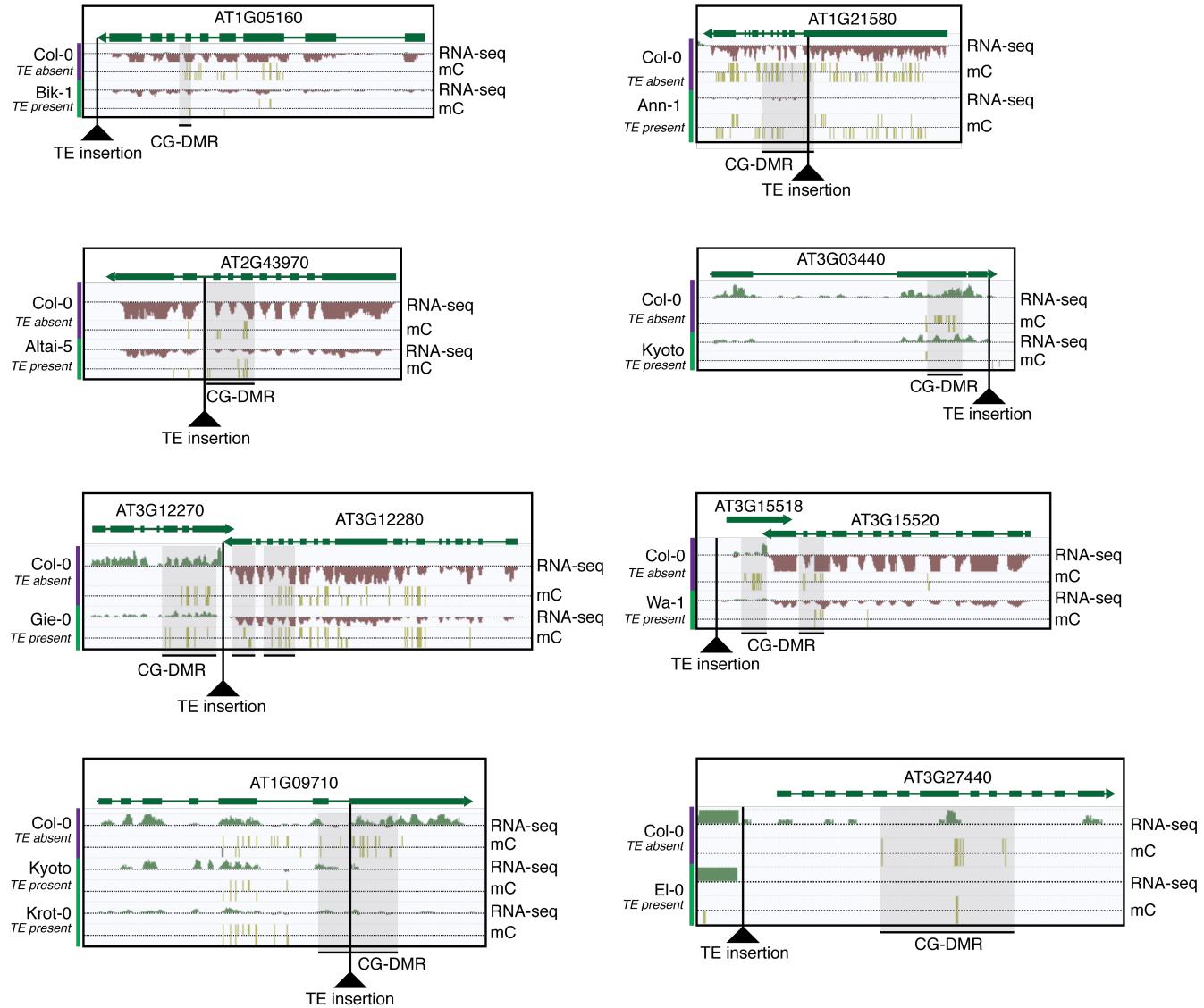


Figure 6: figure supplement 3

835 Selected examples of TE insertions apparently associated with transcriptional downregulation of
 836 nearby genes and loss of gene body CG methylation leading to the formation of a CG-DMR.

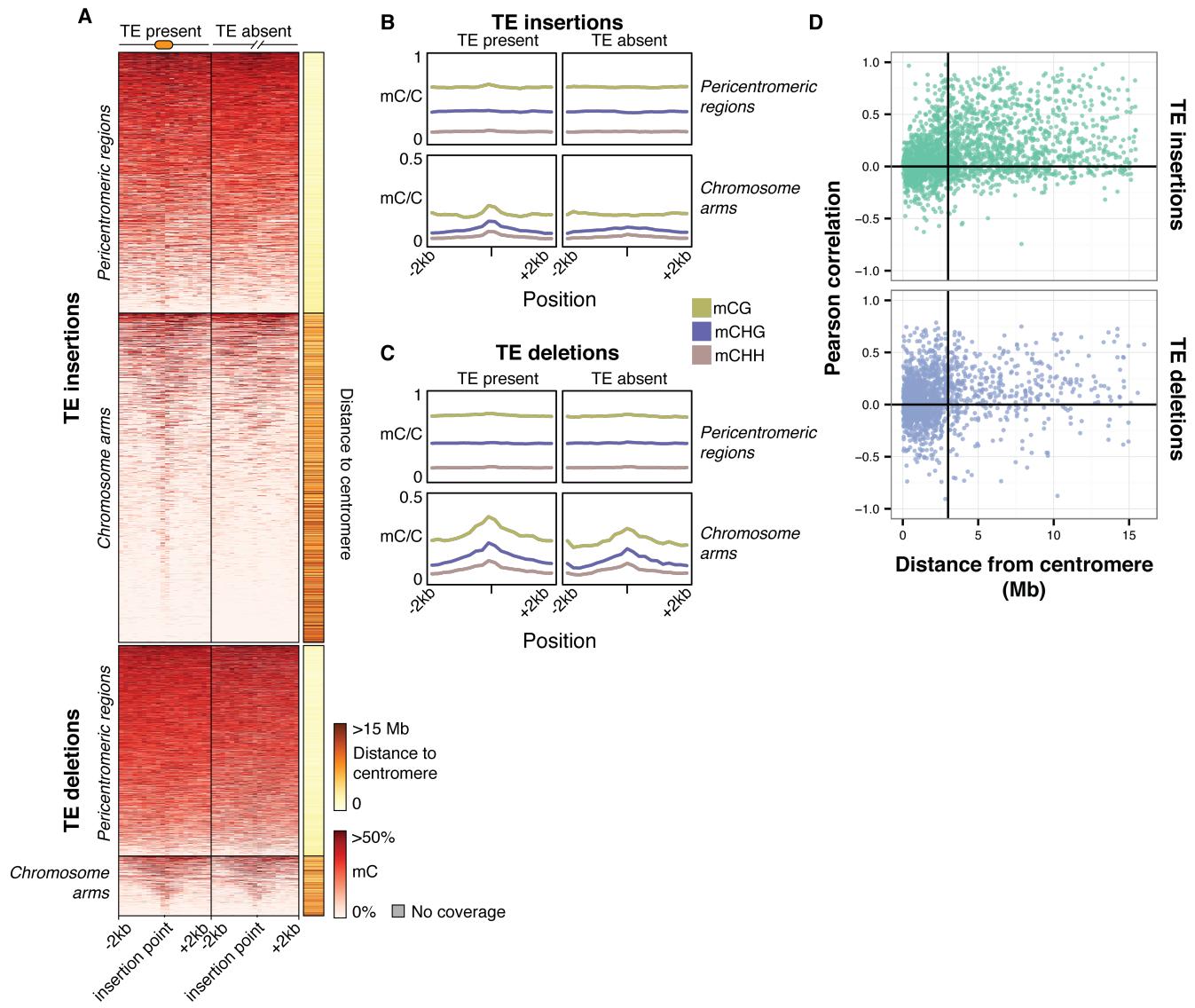


Figure 7: Local patterns of DNA methylation surrounding TE variant sites

- (A) Heatmap showing DNA methylation levels in 200 bp bins flanking TE variant sites, +/- 2 kb from the TE insertion point. TE variants were grouped into pericentromeric variants (<3 Mb from a centromere) or variants in the chromosome arms (>3 Mb from a centromere).
- (B) Line plot showing the DNA methylation level in each sequence context for TE insertion sites, +/- 2 kb from the TE insertion point.
- (C) As for B, for TE deletions.
- (D) Distribution of Pearson correlation coefficients between TE presence/absence and DNA methylation levels in the 200 bp regions flanking TE variant, ordered by distance to the centromere.

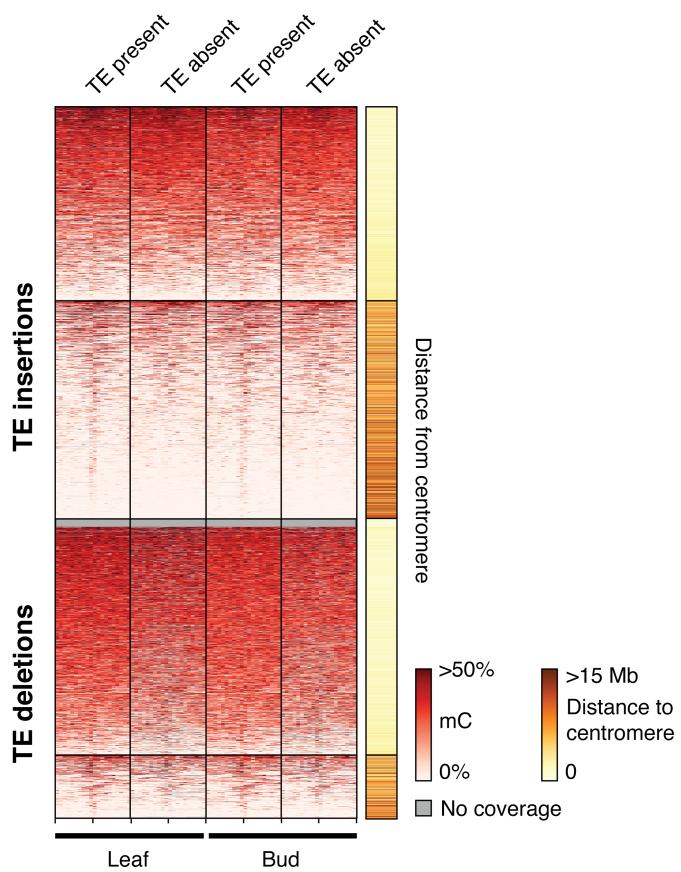


Figure 7: figure supplement 1

845 Heatmap showing DNA methylation levels in 200 bp bins flanking TE variant sites in the 12 accessions
 846 with DNA methylation data for both leaf and bud tissue, +/- 2 kb from the TE insertion point. TE
 847 variants were grouped into pericentromeric variants (<3 Mb from a centromere) or variants in the
 848 chromosome arms (>3 Mb from a centromere).

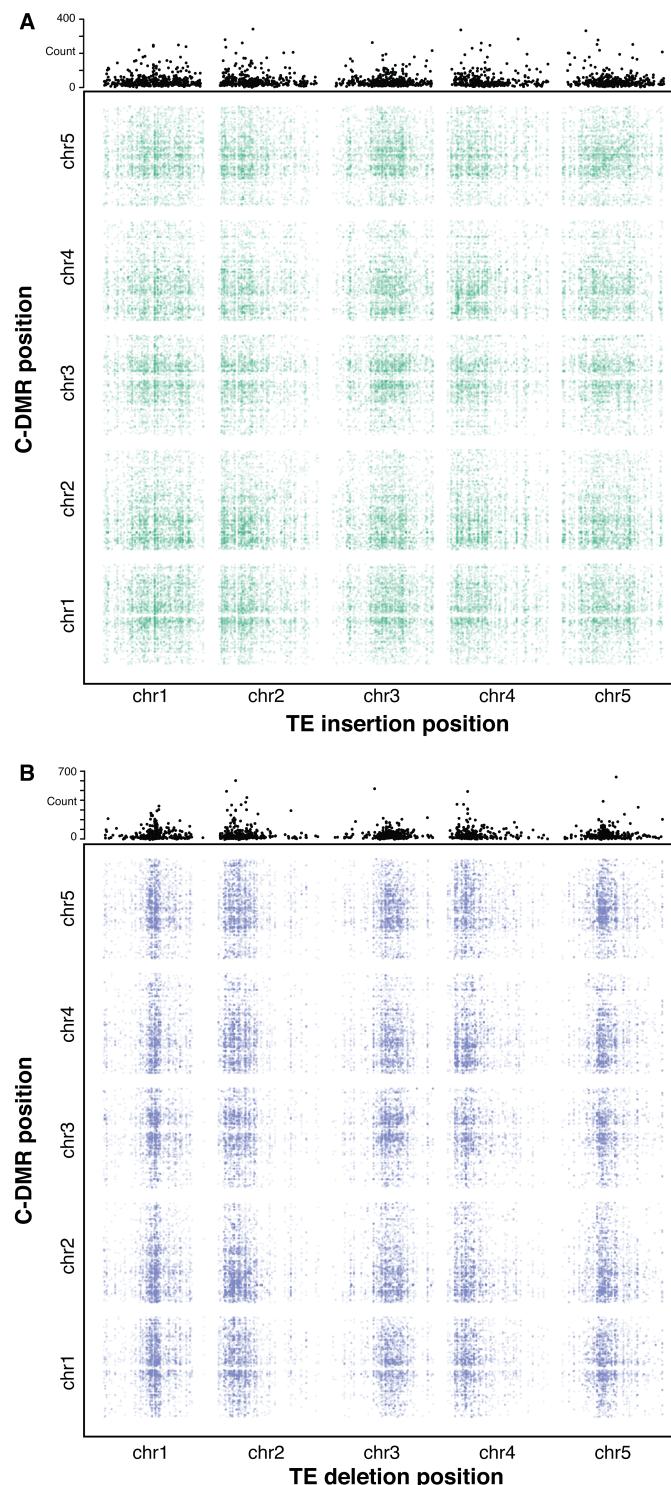


Figure 8: Association scan between TE variants and C-DMR methylation variation

- 849 (A) Significant correlations between TE insertions and C-DMR DNA methylation level. Points
 850 show correlations between individual TE-DMR pairs that were more extreme than any of 500
 851 permutations of the DMR data. Top plots show the total number of significant correlations for
 852 each TE insertion across the whole genome.
- 853 (B) As for (A), for TE deletions.

Table 1: Mapping of paired-end reads providing evidence for TE presence/absence variants in the *Ler* reference genome

	Concordant	Discordant	Split	Unmapped	Total
Col-0 mapped	0	993	9513	0	10206
<i>Ler</i> mapped	10073	92	34	7	10206

Note: Discordant and split read categories are not mutually exclusive, as some discordant reads may have one read in the mate pair split-mapped.

Table 2: Summary of TE variant classifications

TEPID call	TE classification	Count
Insertion	NA	310
	Insertion	14689
	Deletion	8
Absence	NA	1852
	Insertion	388
	Deletion	5848

Table 3: Percentage of DMRs within 1 kb of a TE variant

	C-DMRs			CG-DMRs		
	Observed	Expected	95% CI	Observed	Expected	95% CI
TE deletions	8.7	16	0.0078	4.3	16	0.0041
TE insertions	36	26	0.0089	9.4	26	0.0047
NA calls	3.4	6.2	0.0052	1.7	6.2	0.0027
Total	48	41	0.01	15	42	0.0054