

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 *Arabidopsis* accessions. Most TE variants were rare, and we find a burden of rare variants associated with local extremes of 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 variants as a source of novel genetic diversity. Nearly 200 common TE variants were associated with significantly altered expression of nearby genes, and a major fraction of inter-accession DNA methylation differences were associated with nearby TE insertions. Overall, this demonstrates that TE variants are a rich source of genetic diversity that likely plays an important role in facilitating epigenomic and transcriptional differences between individuals, and indicates a strong genetic basis for epigenetic variation.

13 Introduction

14 Transposable elements (TEs) are mobile genetic elements present in nearly all studied organisms,
15 and comprise a large fraction of most eukaryotic genomes. The two types of TEs are retrotransposons
16 (type I elements), which transpose via an RNA intermediate requiring a reverse transcription reaction,
17 and DNA transposons (type II elements), which transpose via either a cut-paste or, in the case of
18 Helitrons, a rolling circle mechanism with no RNA intermediate [1]. TE activity poses mutagenic
19 potential as a TE insertion may disrupt functional regions of the genome. Consequently, safeguard
20 mechanisms have evolved to suppress this activity, including the methylation of cytosine nucleotides
21 (DNA methylation) to produce 5-methylcytosine (mC), a modification that can induce transcriptional
22 silencing of the methylated locus. In *Arabidopsis thaliana* (Arabidopsis), DNA methylation occurs in all
23 three DNA sequence contexts: mCG, mCHG, and mCHH, where H is any base but G. Establishment
24 of DNA methylation marks can be carried out by two distinct pathways – the RNA-directed DNA
25 methylation pathway guided by 24 nucleotide (nt) small RNAs (smRNAs), and the DDM1/CMT2
26 pathway [2, 3]. A major function of DNA methylation in Arabidopsis is in the transcriptional silencing of
27 TEs. Loss of DNA methylation due to mutations in genes essential for its establishment or maintenance
28 leads to 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, rapidly 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 appears to be due to variation in TE content [15, 16]. Therefore identification of TE
43 variants is essential in order to develop a more comprehensive understanding of the genetic variation
44 that exists between genomes, and of the consequences of TE movement on genome and cellular
45 function.

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

56 enabling exploration of TE diversity and links to gene regulation and epigenomic variation.

57 Results

58 Computational identification of TE presence/absence variation

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

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

91 To estimate the rate of false negative TE absence calls made using TEPID, we compared our *Ler* TE
92 absence calls to the set of TE absences in *Ler* genome identified previously by aligning full-length
93 Col-0 TEs to the *Ler* reference using BLAT [16]. We found that 89.6% (173/193) of these TE absences
94 were also identified using TEPID, indicating a false negative rate of ~10% for TE absence calls. To

95 determine the rate of false negative TE insertion calls, we ran TEPID using 90 bp paired-end Col-0
96 reads (Col-0 control samples from [24]), aligning reads to the Ler PacBio assembly. As TEPID
97 requires a high-quality TE annotation to discover TE variants, which is not available for the Ler
98 assembly, we looked for discordant and split read evidence at the known Col-0-specific TE insertion
99 sites [16], and found evidence reaching the TEPIID threshold for a TE insertion call to be made at
100 89.6% (173/193) of these sites, indicating a false negative rate of ~10%. However, it should be noted
101 that this estimate does not take into account the TEPIID refinement step used on large populations,
102 and so the false negative rate for samples analyzed in the population from Schmitz et al. (2013) is
103 likely to be lower than this estimate, as each accession gained on average 4% more insertion calls
104 following this refinement step (Figure 2 - figure supplement 1).

105 Abundant TE positional variation among natural *Arabidopsis* populations

106 TEPIID was used to analyze previously published 100 bp paired-end genome resequencing data for
107 216 different *Arabidopsis* accessions [20], and identified 15,007 TE insertions (Figure 2 - source
108 data 1) and 8,088 TE absence variants (Figure 2 - source data 2) relative to the Col-0 reference
109 accession, totalling 23,095 unique TE variants. A recent study focused on identifying recent TE
110 insertions containing target site duplications in this population [16]. Our goal was to provide a
111 comprehensive assessment of TE presence/absence variation in *Arabidopsis*. In most accessions
112 TEPIID identified 300-500 TE insertions (mean = 378) and 1,000-1,500 TE absence variants (mean =
113 1,279), the majority of which were shared by two or more accessions (Figure 2 - figure supplement
114 2). PCR validations were performed for a random subset of 10 insertion and 10 absence variants in
115 14 accessions (totalling 280 validations), confirming the high accuracy of TE variant discovery using
116 the TEPIID package, with a false positive rate for both TE insertion and TE absence identification of
117 ~9%, similar to that observed using simulated data and the Ler genome analysis (Figure 2 - figure
118 supplement 3). The number of TE insertions identified was positively correlated with sequencing depth
119 of coverage, while the number of TE absence variants identified had no correlation with sequencing
120 coverage (Figure 2 - figure supplement 4A, B), indicating that the sensitivity of TE absence calls is
121 not limited by sequencing depth, while TE insertion identification benefits from high sequencing depth.
122 However, accessions with low coverage gained more TE insertion calls during the TEPIID refinement
123 step (Figure 2 - figure supplement 4C), indicating that these false negatives were effectively reduced
124 by leveraging TE variant information for the whole population.

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

137 shared by other accessions in the population (Table 2). Overall, we identified 15,077 TE insertions,
138 5,856 true TE deletions, and 2,162 TE variants at a high MAF that were unable to be classified as an
139 insertion or deletion (Figure 2 - source data 3).

140 TE insertions and deletions were distributed throughout chromosome 1 in a pattern that was similar
141 to the distribution of all Col-0 TEs (Figure 2A). TE deletions and common TE variants were found in
142 similar chromosomal regions, as deletion variants represent the rare loss of common variants. TE
143 deletions and common variants were more highly enriched in the pericentromeric regions than rare
144 variants or TE insertions. Among TE deletions, type II elements were slightly less biased towards
145 the centromeres in comparison to the distribution of type I elements (Figure 2 - figure supplement 5).
146 The distribution of rare TE variants and TE insertions was similar to that observed for regions of the
147 genome previously identified as being differentially methylated in all DNA methylation contexts (mCG,
148 mCHG, mCHH) between the wild accessions (population C-DMRs), while population CG-DMRs,
149 differentially methylated in the mCG context, less frequently overlapped with all types of TE variants
150 identified [20]. Furthermore, TE variants were depleted within genes and DNase I hypersensitivity
151 sites [25], while they were enriched in gene flanking regions and within other annotated TEs or
152 pseudogenes (Figure 2B). TE deletions and common TE variants were enriched within the set of TE
153 variants found in gene bodies (Figure 2C, D). No significant enrichment was found for TE variants
154 within the *KNOT ENGAGED ELEMENT* (KEE) regions, previously identified as regions that may act
155 as a “TE sink” [26] (Figure 2 - figure supplement 6). This may indicate that these regions do not act as
156 a “TE sink” as has been previously proposed, or that the “TE sink” activity is restricted to very recent
157 insertions, as the insertions we analysed in this population were likely older than those used in the
158 KEE study [26].

159 Among the identified TE variants, several TE superfamilies were over- or under-represented compared
160 to the number expected by chance given the overall genomic frequency of different TE types (Figure
161 2E). In particular, both TE insertions and deletions in the RC/Helitron superfamily were less numerous
162 than expected, with an 11.5% depletion of RC/Helitron elements in the set of TE variants. In contrast,
163 TEs belonging to the LTR/Gypsy superfamily were more frequently deleted than expected, with
164 a 17% enrichment in the set of TE deletions. This was unlikely to be due to a differing ability of
165 the detection method to identify TE variants of different lengths, as the TE variants identified had
166 a similar distribution of lengths as all *Arabidopsis* TEs annotated in the Col-0 reference genome
167 (Figure 2 - figure supplement 7). These enrichments suggest that the RC/Helitron TEs have been
168 relatively dormant in recent evolutionary history, while the LTR/Gypsy, which are highly enriched in
169 the pericentromeric regions, are frequently lost from the *Arabidopsis* genome. At the family level,
170 we observed similar patterns of TE variant enrichment or depletion (Figure 2 - figure supplement 8;
171 source data 4).

172 We further examined *Arabidopsis* (Col-0) DNA sequencing data from a transgenerational stress
173 experiment to investigate the possible minimum number of generations required for TE variants to
174 arise [24]. In one of the three replicates subjected to high salinity stress conditions, we identified a
175 single potential TE insertion in a sample following 10 generations of single-seed descent, while no
176 TE variants were identified in any of the three control single-seed descent replicate sets. However,
177 without experimental validation it remains unclear if this represents a true variant. Therefore, we
178 conclude that TE variants may arise at a rate less than 1 insertion in 60 generations under laboratory
179 conditions. Further experimental work will be required to precisely determine the rate of transposition
180 in *Arabidopsis*.

181 Relationship between TE variants and single nucleotide polymorphisms

182 Although thousands of TE variants were identified, they may be linked to the previously identified
183 single nucleotide polymorphisms (SNPs), or unlinked from SNPs across the accessions. We tested
184 how frequently common TE variants (>3% MAF, >7 accessions) were linked to adjacent SNPs to
185 determine when they would represent a previously unassessed source of genetic variation between
186 accessions. SNPs that were previously identified between the accessions [20] were compared to the
187 presence/absence of individual TE variants. For the testable TE variants in the population, the nearest
188 flanking 300 SNPs upstream and 300 SNPs downstream of the TE variant site were analyzed for local
189 linkage disequilibrium (LD, r^2 ; see methods). TE variants were classified as being either 'low', 'mid',
190 or 'high' LD variants by comparing ranked r^2 values of TE variant to SNPs against the median ranked
191 r^2 value for all between SNP comparisons (SNP-SNP) to account for regional variation in the extent of
192 SNP-SNP LD (Figure 3A, B) due to recombination rate variation or selection [27]. The majority (61%)
193 of testable TE variants had low LD with nearby SNPs, and represent a source of genetic diversity
194 not previously assessed by SNP-based genotype calling methods (Figure 3C). 29% of TE variants
195 displayed high levels of LD and are tagged by nearby SNPs, while only 10% had intermediate levels
196 of LD. We observed a positive correlation between TE variant MAF and LD state, with variants of
197 a high minor allele frequency more often classified as high-LD (Figure 3D). While the proportion
198 of TE variants classified as high, mid, or low-LD was mostly the same for both TE insertions and
199 TE deletions, TE variants with a high MAF (>20%) that were unable to be classified as either true
200 deletions or as new insertions had a much higher proportion of high-LD variants (Figure 3E). This was
201 consistent with the observation that the more common alleles were more often in a high-LD state. TE
202 variants displayed a similar distribution over chromosome 1 regardless of linkage classification (Figure
203 3 - figure supplement 1). Overall, this analysis revealed an abundance of previously uncharacterized
204 genetic variation that exists amongst *Arabidopsis* accessions caused by the presence or absence of
205 TEs, and illustrates the importance of identifying TE variants alongside other genetic diversity such as
206 SNPs.

207 TE variants affect gene expression

208 To determine whether the newly discovered TE variants may affect nearby gene expression, the
209 steady state transcript abundance within mature leaf tissue was compared between accessions with
210 and without TE insertions or deletions, for genes with TE variants located in the 2 kb gene upstream
211 region, 5' UTR, exons, introns, 3' UTR or 2 kb downstream region (Figure 4A). While the steady state
212 transcript abundance of most genes appeared to be unaffected by the presence of a TE, 196 genes
213 displayed significant differences in transcript abundance linked with the presence of a TE variant,
214 indicating a role for these variants in the local regulation of gene expression (1% false discovery rate;
215 >2-fold change in transcript abundance; Figure 4A, Figure 4 - source data 1). No functional category
216 enrichments in this set of differentially expressed genes were identified. As rare TE variants with
217 a MAF less than 3% may also be associated with a difference in transcript abundance, but were
218 unable to be statistically tested due to their rarity, a burden test for enrichment of rare variants in the
219 extremes of expression was performed [28]. Briefly, this method counts the frequency of rare variants
220 within each gene expression rank in the population, and aggregates this information over the entire
221 population to determine whether an enrichment of rare variants exists within any gene expression

rank. A strong enrichment for gene expression extremes was observed 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 had 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 indicates 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 potential positional effects 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 simply a change in total transcript abundance of 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* has been 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]. Examination of pathogen resistance phenotype data [30] revealed 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.

Some TE variants were also 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 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 the presence of a TE belonging to the *HELI TRON1* family, indicates that this element may carry regulatory information that alters 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 linked to genetic differences between the accessions using only SNP data. This TE variant was also upstream of *QPT* (AT2G01350), involved in NAD biosynthesis [31], which did not show alterations in 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.

265 **TE variants explain many DNA methylation differences between accessions**

266 As TEs are frequently highly methylated in Arabidopsis [32–35], the DNA methylation state surrounding
267 TE variant sites was assessed to determine whether TE variants might be responsible for differences
268 in DNA methylation patterns previously observed between the wild accessions [20]. TE variants were
269 often physically close to DMRs (Figure 6A). Furthermore, C-DMRs were more often close to a TE
270 variant than expected, whereas CG-DMRs were rarely close to TE insertions or TE deletions (Table
271 3). Overall, 54% of the 13,482 previously reported population C-DMRs were located within 1 kb of
272 a TE variant (predominantly TE insertions), while only 15% of CG-DMRs were within 1 kb of a TE
273 variant (Table 3). For C-DMRs, this was significantly more than expected by chance, while it was
274 significantly less than expected for CG-DMRs ($p < 1 \times 10^{-4}$, determined by resampling 10,000 times).
275 Of the C-DMRs that were not close to a TE variant, 3,701 (27% of all C-DMRs) were within 1 kb
276 of a non-variable TE. Thus, 81% of C-DMRs are within 1 kb of a TE when considering both fixed
277 and variable TEs in the population. Of the remaining 19% of C-DMRs, most were found in genes or
278 intergenic regions.

279 To determine whether DMR methylation levels were dependent on the presence/absence of nearby
280 TE variants, Pearson correlation coefficients were calculated between the DNA methylation level at
281 each DMR and the presence/absence of the nearest TE variant. A negative correlation was observed
282 between the distance from a C-DMR to the nearest TE insertion and the correlation between the
283 DNA methylation level at the C-DMR with the presence/absence of the TE insertion (Figure 6B). This
284 suggests a distance-dependent effect of TE insertion presence on C-DMR methylation. In contrast, no
285 such relationship was found for TE deletions on C-DMRs, or for insertions or deletions on CG-DMRs
286 (Figure 6B). DNA methylation levels at C-DMRs located within 1 kb of a TE insertion (TE-DMRs) were
287 more often positively correlated with the presence/absence of a TE insertion than the DNA methylation
288 levels at C-DMRs further than 1 kb from a TE insertion (non-TE-DMRs). This was evident from the
289 distribution of correlations for non-TE-DMRs being centred around zero, whereas for TE-DMRs this
290 distribution was skewed to the right (Figure 6C, $D=0.24$). For TE deletions, such a difference was not
291 observed in the distributions of correlation coefficients between TE-DMRs and non-TE-DMRs, nor for
292 CG-DMRs and their nearby TE insertions or deletions (Figure 6C, $D=0.07-0.10$). Furthermore, DNA
293 methylation levels were often higher in the presence of the nearby TE insertion, while this relationship
294 was generally not observed for C-DMRs further than 1 kb from a TE variant, for TE deletions, or for
295 CG-DMRs (Figure 6 - figure supplement 1).

296 As the above correlations between TE presence/absence and DMR methylation level rely on the TE
297 variants having a sufficiently high MAF, this precludes analysis of the effect of rare variants on DMR
298 methylation levels. To determine the effect that these rare TE variants may have on DMR methylation
299 levels, a burden test for enrichment of DMR methylation extremes at TE-DMRs was performed,
300 similar to the analysis undertaken to test the effect of rare variants on gene expression. A strong
301 enrichment was observed for high C-DMR and CG-DMR methylation level ranks for TE insertions,
302 while TE deletions were associated with both high and low extremes of DNA methylation levels at
303 C-DMRs, and less so at CG-DMRs (Figure 6D). This further indicates that the presence of a TE
304 insertion is associated with higher C-DMR methylation levels, while TE deletions appear to have more
305 variable effects on DMR methylation levels. This enrichment was completely absent after repeating
306 the analysis with randomized accession names (Figure 6 - figure supplement 2). A slight enrichment
307 was also observed for low DMR methylation ranks for TE insertions near CG-DMRs, indicating that

308 the insertion of a TE was sometimes associated with reduced CG methylation in nearby regions (<1
309 kb from the TE). Closer examination of these TE insertions revealed that some TE insertions were
310 associated with decreased transcript abundance of nearby genes, with a corresponding loss of gene
311 body methylation, offering a potential explanation for the decreased CG methylation observed near
312 some TE insertions (Figure 6 - figure supplement 3).

313 To further assess the effects of TE variants upon local DNA methylation patterns, the levels of methy-
314 lation were examined in regions flanking all TE variants regardless of the presence or absence of a
315 population DMR call. While DNA methylation levels around pericentromeric TE insertions and dele-
316 tions (<3 Mb from a centromere) seemed to be unaffected by the presence of a TE insertion (Figure
317 7A), TE insertions in the chromosome arms were associated with an increase in DNA methylation
318 levels in all sequence contexts (Figure 7A, B). In contrast, TE deletions in the chromosome arms did
319 not affect patterns of DNA methylation, as the flanking methylation level in all contexts appeared to
320 remain high following deletion of the TE (Figure 7A, C). As the change in DNA methylation levels
321 around most TE variant sites appeared to be restricted to regions <200 bp from the insertion site, DNA
322 methylation levels in 200 bp regions flanking TE variants were correlated with the presence/absence
323 of TE variants. DNA methylation levels were often positively correlated with the presence of a TE
324 insertion when the insertion was distant from a centromere (Figure 7D). TE deletions were more vari-
325 ably correlated with local DNA methylation levels, but also showed a bias towards positive correlations
326 for TE deletions distant from the centromeres. As methylome data was available for both leaf and
327 bud tissue for 12 accessions, this analysis was repeated comparing between tissue types, but no
328 differences were observed in the patterns of methylation surrounding TE variant sites between the
329 two tissues (Figure 7 - figure supplement 1).

330 These results indicate that local DNA methylation patterns are influenced by the differential TE content
331 between genomes, and that the DNA methylation-dependent silencing of TEs may frequently lead to
332 the formation of DMRs between wild *Arabidopsis* accessions. TE insertions appear to be important
333 in defining local patterns of DNA methylation, while DNA methylation levels often remain elevated
334 following a TE deletion, and so are independent from the presence or absence of TEs in these cases.
335 Importantly, the distance from a TE insertion to the centromere appears to have a strong impact
336 on whether an alteration of local DNA methylation patterns will occur. This is likely due to flanking
337 sequences being highly methylated in the pericentromeric regions, and so the insertion of a TE
338 cannot further increase levels of DNA methylation. Overall, a large fraction of the population C-DMRs
339 previously identified between wild accessions are correlated with the presence of local TE variants.
340 CG-DMR methylation levels appear to be mostly independent from the presence/absence of common
341 TE variants, while rare TE variants have an impact on DNA methylation levels at both C-DMRs and
342 CG-DMRs. This analysis was aided by the high sensitivity of TEPID to detect TE variants, allowing a
343 thorough assessment of the impact of TE variants on DNA methylation patterns.

344 **Genome-wide association scan highlights distant and local control of DNA 345 methylation**

346 To further investigate the effects of TE variants upon local and distant DNA methylation levels in
347 the genome, an association scan was conducted for all common TE variants (>3% MAF) and all
348 population C-DMRs for the 124 accessions with both DNA methylation and TE variant data available.

349 To test the significance of each pairwise correlation, bootstrap p-value estimates were collected
350 based on 500 permutations of accession labels. TE-DMR associations were deemed significant if
351 they had an association more extreme than any of the 500 permutations ($p < 1/500$). A band of
352 significant associations was observed for TE insertions and their nearby C-DMRs, signifying a local
353 association between TE insertion presence/absence and C-DMR methylation (Figure 8A). This local
354 association was not as strong for TE deletions (Figure 8B), consistent with our above findings. While
355 TE variants and DNA methylation showed a local association, it is also possible that TE variation may
356 influence DNA methylation states more broadly in the genome, perhaps through production of *trans*-
357 acting smRNAs or inactivation of genes involved in DNA methylation establishment or maintenance.
358 To identify any potential enrichment of C-DMRs regulated in *trans*, the total number of significant
359 associations was summed for each TE variant across the whole genome (Figure 8A and B, top
360 panels). At many sites, far more significant associations were found than expected due to the false
361 positive rate alone. This suggested the existence of many putative *trans* associations between TE
362 variants and genome-wide C-DMR methylation levels. These C-DMRs that appeared to be associated
363 with a TE insertion in *trans* were further examined, checking for TE insertions near these C-DMRs
364 that were present in the same accessions as the *trans* associated TE, as these could lead to a false
365 *trans* association. These were extremely rare, with only 4 such cases for TE insertions, and 38 cases
366 for TE deletions, and so were unable to explain the high degree of *trans* associations found. Overall,
367 this analysis suggests that certain TE variants may affect DNA methylation levels more broadly in the
368 genome, as their effects upon DNA methylation are not necessarily limited to nearby DNA sequences.

369 Discussion

370 Here we have discovered widespread differential TE content between wild *Arabidopsis* accessions,
371 and explored the impact of these variants upon transcription and DNA methylation at the level of
372 individual accessions. Most TE variants were due to the *de novo* insertion of TEs, while a smaller
373 subset was likely due to the deletion of ancestral TE copies, mostly around the pericentromeric
374 regions. A subset (32%) of TE variants with a minor allele frequency above 3% were able to be
375 tested for linkage with nearby SNPs. The majority of these TE variants exhibited only low levels of LD
376 with nearby SNPs, indicating that they represent genetic variants currently overlooked in genomic
377 studies. A marked depletion of TE variants within gene bodies and DNase I hypersensitivity sites
378 (putative regulatory regions) is consistent with the more deleterious TE insertions being removed
379 from the population through selection. Of those TE variants found in gene bodies, TE deletions were
380 overrepresented, indicating that the loss of ancestral TEs inserted within genes may be more frequent,
381 or perhaps less deleterious, than the *de novo* insertion of TEs into genes.

382 A previous study focused on recent TE insertions in the *Arabidopsis* population [16], thus the extensive
383 variation between accessions due to older TE insertions or TE deletions has not been explored. We
384 examined both old and new TE variants (common and rare alleles), as well as TE deletions, allowing
385 the detection of 23,095 TE variants in the population. This large sample size gave an opportunity to
386 form statistically robust correlations between TE presence/absence and transcript abundance from
387 nearby genes, as well as genome-wide patterns of DNA methylation. Examples were identified where
388 TE variants appear to have an effect upon gene expression, both in the disruption of transcription
389 and in the spreading or disruption of regulatory information leading to the transcriptional activation

of genes, indicating that these TE variants can have important consequences upon the expression of protein coding genes (Figure 5). In one case, these changes in gene expression could be linked with phenotypic changes, with accessions containing a TE insertion more frequently sensitive to bacterial infection. Further experiments will be needed to establish a causal link between this TE insertion and the associated phenotype. An analysis of rare TE variants, present at a low MAF, further strengthened this relationship between TE presence/absence and altered transcript abundance, as a strong enrichment of rare TE variants in accessions with extreme gene expression ranks in the population was identified. Therefore, the effects of TE insertions appear to be long-lasting, as there was little difference between common (old) and rare (young) variants in the impact upon gene expression (Figure 4).

Importantly, we provide evidence that differential TE content between genomes of *Arabidopsis* accessions underlies a large fraction of the previously reported population C-DMRs. Thus, the frequency of pure epialleles, independent of underlying genetic variation, may be even more rare than previously anticipated [36]. Overall, 81% of all C-DMRs were within 1 kb of a TE, when considering both fixed and variable TEs in the population. We did not find evidence of CG-DMR methylation level being altered by the presence of common TE variants, although rare TE variants may be more important in shaping patterns of DNA methylation at some CG-DMRs, though the reasons for this distinction remain unclear. The level of local DNA methylation changes associated with TE variants was also related to the distance from a TE variant to the centromere, with variants in the chromosome arms being more strongly correlated with DNA methylation levels. This seems to be due to a higher baseline level of DNA methylation at the pericentromeric regions, which prevent any further increase in DNA methylation level following insertion of a TE. Furthermore, we found an important distinction between TE insertions and TE deletions in the effect that these variants have on nearby DNA methylation levels. While flanking DNA methylation levels increase following a TE insertion, the deletion of an ancestral TE was often not associated with a corresponding decrease in flanking DNA methylation levels (Figure 7). This indicates that high levels of DNA methylation, once established, may be maintained in the absence of the TE insertion that presumably triggered the original change in DNA methylation level. It is then possible that TE variants explain more of the inter-accession variation in DNA methylation patterns than we find direct evidence for, if some C-DMRs were formed by the insertion of an ancestral TE that is now absent in all the accessions analysed here. These DMRs would then represent the epigenetic “scars” of past TE insertions.

Finally, a genome-wide scan of common TE variant association with C-DMR methylation levels provides further evidence of a strong local association between TE insertion presence/absence and C-DMR methylation level (Figure 8). The identification of some TE variants that appeared to be associated with changes in DNA methylation levels at multiple loci throughout the genome indicates possible *trans* regulation of DNA methylation state linked to specific TE variants. Further experiments will be required to confirm and examine the role of these TE variants in determining genome-wide patterns of DNA methylation. Overall, our results show that TE presence/absence variants between wild *Arabidopsis* accessions not only have important effects on nearby gene expression, but can also have a role in determining local patterns of DNA methylation, and explain many regions of differential DNA methylation previously observed in the population.

431 **Methods**

432 **TEPID development**

433 *Mapping*

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

440 *TE variant discovery*

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

446 To identify TE insertions with respect to the reference genome, split read alignments are first filtered
447 to remove reads where the distance between split mapping loci is less than 5 kb, to remove split reads
448 due to small indels, or split reads with a mapping quality (MAPQ) less than 5. Split and discordant
449 read mapping coordinates are then intersected using pybedtools [41, 42] with the Col-0 reference TE
450 annotation, requiring 80% overlap between TE and read mapping coordinates. To determine putative
451 TE insertion sites, regions are then identified that contain independent discordant read pairs aligned
452 in an orientation facing one another at the insertion site, with their mate pairs intersecting with the
453 same TE (Figure 1). The total number of split and discordant reads intersecting the insertion site
454 and the TE is then calculated, and a TE insertion predicted where the combined number of reads
455 is greater than a threshold determined by the average sequencing depth over the whole genome
456 (1/10 coverage if coverage is greater than 10, otherwise a minimum of 2 reads). Alternatively, in the
457 absence of discordant reads mapped in orientations facing one another, the required total number of
458 split and discordant reads at the insertion site linked to the inserted TE is set higher, requiring twice
459 as many reads.

460 To identify TE absence variants with respect to the reference genome, split and discordant reads
461 separated >20 kb from one another are first removed, as 99.9% of *Arabidopsis* TEs are shorter than
462 20 kb, and this removes split reads due to larger structural variants not related to TE diversity (Figure
463 2 - figure supplement 7). Col-0 reference annotation TEs that are located within the genomic region
464 spanned by the split and discordant reads are then identified. TE absence variants are predicted
465 where at least 80% of the TE sequence is spanned by a split or discordant read, and the sequencing
466 depth within the spanned region is <10% the sequencing depth of the 2 kb flanking sequence, and
467 there are a minimum number of split and discordant reads present, determined by the sequencing
468 depth (1/10 coverage; Figure 1). A threshold of 80% TE sequence spanned by split or discordant
469 reads is used, as opposed to 100%, to account for misannotation of TE sequence boundaries in the
470 Col-0 reference TE annotation, as well as TE fragments left behind by DNA TEs during cut-paste

471 transposition (TE footprints) that may affect the mapping of reads around annotated TE borders [43].
472 Furthermore, the coverage within the spanned region may be more than 10% that of the flanking
473 sequence, but in such cases twice as many split and discordant reads are required. If multiple TEs are
474 spanned by the split and discordant reads, and the above requirements are met, multiple TEs in the
475 same region can be identified as absent with respect to the reference genome. Absence variants in
476 non-Col-0 accessions are subsequently recategorized as TE insertions present in the Col-0 genome
477 but absent from a given wild accession.

478 *TE variant refinement*

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

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

501 **TE variant simulation**

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

509 **Estimation of sensitivity**

510 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
511 TEPID package to identify TE variants. Reads providing evidence for TE variants were then mapped to
512 the *de novo* assembled *Ler* genome [21]. To determine whether reads mapped to homologous regions
513 of the *Ler* and *Col-0* reference genome, the *de novo* assembled *Ler* genome sequence between
514 mate pair mapping locations in *Ler* were extracted, with repeats masked using RepeatMasker with
515 RepBase-derived libraries and the default parameters (version 4.0.5, <http://www.repeatmasker.org>).
516 A blastn search was then conducted against the *Col-0* genome using the following parameters:
517 ‘-max-target-seqs 1’, ‘-eval 1e-6’ [23]. Coordinates of the top blast hit for each read location were
518 then compared with the TE variant sites identified using those reads. To estimate false negative rates
519 for TEID TE absence calls, *Ler* TE absence calls were compared with a known set of *Col-0*-specific
520 TE insertions, absent in *Ler* [16]. For TEID TE insertion calls, we mapped *Col-0* DNA sequencing
521 reads [24] to the *Ler* PacBio assembly, and identified sites with read evidence reaching the TEID
522 threshold for a TE insertion call to be made.
523

524 **Arabidopsis TE variant discovery**

525 We ran TEID, including the insertion refinement step, on previously published sequencing data for
526 216 different Arabidopsis populations (NCBI SRA SRA012474; [20]), mapping to the TAIR10 reference
527 genome and using the TAIR9 TE annotation. The ‘–mask’ option was set to mask the mitochondrial
528 and plastid genomes. We also ran TEID using previously published transgenerational data for salt
529 stress and control conditions (NCBI SRA SRP045804; [24]), again using the ‘–mask’ option to mask
530 mitochondrial and plastid genomes, and the ‘–strict’ option for highly related samples.

531 **TE variant / SNP comparison**

532 SNP information for 216 Arabidopsis accessions was obtained from the 1001 genomes data center
533 (http://1001genomes.org/data/Salk/releases/2013_24_01/; [20]). This was formatted into reference
534 (Col-0 state), alternate, or NA calls for each SNP. Accessions with both TE variant information and
535 SNP data were selected for analysis. Hierarchical clustering of accessions by SNPs as well as
536 TE variants were used to identify essentially clonal accessions, as these would skew minor allele
537 frequency calculations. A single representative from each cluster of similar accessions was kept,
538 leading to a total of 187 accessions for comparison. For each TE variant with a minor allele frequency
539 greater than 3%, the nearest 300 upstream and 300 downstream SNPs with a minor allele frequency
540 greater than 3% were selected. Pairwise genotype correlations (r^2 values) for all complete cases were
541 obtained for SNP-SNP and SNP-TE variant states. r^2 values were then ordered by decreasing rank
542 and a median SNP-SNP rank value was calculated. For each of the 600 ranked surrounding positions,
543 the number of times the TE rank was greater than the SNP-SNP median rank was calculated as a
544 relative LD metric of TE to SNP. TE variants with less than 200 ranks over the SNP-SNP median
545 were classified as low-LD insertions. TE variants with ranks between 200 and 400 were classified as

546 mid-LD, while TE variants with greater than 400 ranks above their respective SNP-SNP median value
547 were classified as variants in high LD with flanking SNPs.

548 PCR validations

549 Selection of accessions to be genotyped

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

555 Selection of TE variants for validation and primer design

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

570 PCR

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

574 mRNA analysis

575 Processed mRNA data for 144 wild Arabidopsis accessions were downloaded from NCBI GEO
576 GSE43858 [20]. To find differential gene expression dependent on TE presence/absence variation,
577 we first filtered TE variants to include only those where the TE variant was shared by at least 5
578 accessions with RNA data available. We then grouped accessions based on TE presence/absence
579 variants, and performed a Mann-Whitney U test to determine differences in RNA transcript abundance
580 levels between the groups. We used q-value estimation to correct for multiple testing, using the R
581 qvalue package v2.2.2 with the following parameters: lambda = seq(0, 0.6, 0.05), smooth.df = 4 [48].
582 Genes were defined as differentially expressed where there was a greater than 2 fold difference in

583 expression between the groups, with a q-value less than 0.01. Gene ontology enrichment analysis
584 was performed using PANTHER (<http://pantherdb.org>).

585 DNA methylation data analysis

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

588 Rare variant analysis

589 To assess the effect of rare TE variants on gene expression or DMR DNA methylation levels, we
590 tested for a burden of rare variants in the population extremes, essentially as described previously
591 [28]. For each rare TE variant near a gene or DMR, we ranked the gene expression level or DMR DNA
592 methylation level for all accessions in the population, and tallied the ranks of accessions containing a
593 rare variant. These rank counts were then binned to produce a histogram of the distribution of ranks.
594 We then fit a quadratic model to the counts data, and calculated the R² and p-value for the fit of the
595 model.

596 TE variant and DMR genome-wide association analysis

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

604 Data access

605 TEPIID source code can be accessed at <https://github.com/ListerLab/TEPID>. Code and data needed
606 to reproduce this analysis can be found at <https://github.com/timoast/Arabidopsis-TE-variants>. Ler
607 TE variants are available in Figure 1 - source data 1 and 2. TE variants identified among the 216 wild
608 Arabidopsis accessions resequenced by Schmitz et al. (2013) are available in Figure 2 - source data
609 1, 2 and 3. Source data is available on Dryad (<http://dx.doi.org/10.5061/dryad.187b3>).

610 **Acknowledgments**

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

619 **Author contributions**

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

623 **Competing financial interests**

624 The authors declare no competing financial interests.

625 **References**

- 626 [1] Thomas Wicker et al. "A unified classification system for eukaryotic transposable elements." In:
627 *Nature Reviews Genetics* 8.12 (Dec. 2007), pp. 973–982. DOI: [10.1038/nrg2165](https://doi.org/10.1038/nrg2165).
- 628 [2] Assaf Zemach et al. "The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltrans-
629 ferases to Access H1-Containing Heterochromatin". In: *Cell* 153.1 (Mar. 2013), pp. 193–205.
630 DOI: [10.1016/j.cell.2013.02.033](https://doi.org/10.1016/j.cell.2013.02.033).
- 631 [3] Marjori A Matzke and Rebecca A Mosher. "RNA-directed DNA methylation: an epigenetic
632 pathway of increasing complexity". In: *Nature Reviews Genetics* 15.6 (May 2014), pp. 394–408.
633 DOI: [10.1038/nrg3683](https://doi.org/10.1038/nrg3683).
- 634 [4] Marie Mirouze et al. "Selective epigenetic control of retrotransposition in Arabidopsis." In: *Nature*
635 461.7262 (Sept. 2009), pp. 427–430. DOI: [10.1038/nature08328](https://doi.org/10.1038/nature08328).
- 636 [5] Asuka Miura et al. "Mobilization of transposons by a mutation abolishing full DNA methylation in
637 Arabidopsis". In: *Nature* 411.6834 (2001), pp. 212–214. DOI: [10.1038/35075612](https://doi.org/10.1038/35075612).
- 638 [6] Hidetoshi Saze, Ortrun Mittelsten Scheid, and Jerzy Paszkowski. "Maintenance of CpG methy-
639 lation is essential for epigenetic inheritance during plant gametogenesis". In: *Nature Genetics*
640 34.1 (Mar. 2003), pp. 65–69. DOI: [10.1038/ng1138](https://doi.org/10.1038/ng1138).
- 641 [7] Zachary Lippman et al. "Role of transposable elements in heterochromatin and epigenetic
642 control." In: *Nature* 430.6998 (July 2004), pp. 471–476. DOI: [10.1038/nature02651](https://doi.org/10.1038/nature02651).
- 643 [8] Jeffrey A Jeddeloh, Trevor L Stokes, and Eric J Richards. "Maintenance of genomic methylation
644 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).
- 646 [9] Clémentine Vitte et al. "The bright side of transposons in crop evolution." In: *Briefings in
647 Functional Genomics* 13.4 (July 2014), pp. 276–295. DOI: [10.1093/bfgp/elu002](https://doi.org/10.1093/bfgp/elu002).
- 648 [10] Elizabeth Hénaff et al. "Extensive amplification of the E2F transcription factor binding sites
649 by transposons during evolution of Brassica species." In: *The Plant Journal* 77.6 (Mar. 2014),
650 pp. 852–862. DOI: [10.1111/tpj.12434](https://doi.org/10.1111/tpj.12434).
- 651 [11] Anthony Bolger et al. "The genome of the stress-tolerant wild tomato species". In: *Nature
652 Genetics* 46.9 (July 2014), pp. 1034–1038. DOI: [10.1038/ng.3046](https://doi.org/10.1038/ng.3046).
- 653 [12] Hidetaka Ito et al. "An siRNA pathway prevents transgenerational retrotransposition in plants
654 subjected to stress". In: *Nature* 472.7341 (Mar. 2011), pp. 115–119. DOI: [10.1038/nature09861](https://doi.org/10.1038/nature09861).

- 655 [13] Irina Makarevitch et al. "Transposable Elements Contribute to Activation of Maize Genes in
656 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).
- 658 [14] Paul Bundock and Paul Hooykaas. "An *Arabidopsis* hAT-like transposase is essential for plant
659 development." In: *Nature* 436.7048 (July 2005), pp. 282–284. DOI: [10.1038/nature03667](https://doi.org/10.1038/nature03667).
- 660 [15] Jun Cao et al. "Whole-genome sequencing of multiple *Arabidopsis thaliana* populations." In:
661 *Nature Genetics* 43.10 (Oct. 2011), pp. 956–963. DOI: [10.1038/ng.911](https://doi.org/10.1038/ng.911).
- 662 [16] Leandro Quadrana et al. "The *Arabidopsis thaliana* mobilome and its impact at the species
663 level." In: *eLife* 5 (2016), p. 6919. DOI: [10.7554/eLife.15716](https://doi.org/10.7554/eLife.15716).
- 664 [17] Djie Tjwan Thung et al. "Mobster: accurate detection of mobile element insertions in next
665 generation sequencing data". In: *Genome Biology* (Oct. 2014), pp. 1–11. DOI: [10.1186/s13059-014-0488-x](https://doi.org/10.1186/s13059-014-0488-x).
- 667 [18] Sofia M. C. Robb et al. "The use of RelocaTE and unassembled short reads to produce high-
668 resolution snapshots of transposable element generated diversity in rice". In: *G3: Genes /*
669 *Genomes / Genetics* (2013). DOI: [10.1534/g3.112.005348/-/DC1](https://doi.org/10.1534/g3.112.005348/-/DC1).
- 670 [19] Elizabeth Hénaff et al. "Jitterbug: somatic and germline transposon insertion detection at single-
671 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).
- 673 [20] Robert J Schmitz et al. "Patterns of population epigenomic diversity". In: *Nature* 495.7440 (Mar.
674 2013), pp. 193–198. DOI: [10.1038/nature11968](https://doi.org/10.1038/nature11968).
- 675 [21] Chen-Shan Chin et al. "Nonhybrid, finished microbial genome assemblies from long-read SMRT
676 sequencing data". In: *Nature Methods* 10.6 (May 2013), pp. 563–569. DOI: [10.1038/nmeth.2474](https://doi.org/10.1038/nmeth.2474).
- 677 [22] Korbinian Schneeberger et al. "Reference-guided assembly of four diverse *Arabidopsis thaliana*
678 genomes". In: *Proceedings of the National Academy of Sciences of the United States of America*
679 108.25 (2011), pp. 10249–10254. DOI: [10.1073/pnas.1107739108](https://doi.org/10.1073/pnas.1107739108).
- 680 [23] Christiam Camacho et al. "BLAST+: architecture and applications." In: *BMC Bioinformatics* 10.1
681 (2009), p. 421. DOI: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421).
- 682 [24] Caifu Jiang et al. "Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations." In: *Genome Research* 24.11 (Nov. 2014), pp. 1821–
683 1829. DOI: [10.1101/gr.177659.114](https://doi.org/10.1101/gr.177659.114).
- 685 [25] Alessandra M Sullivan et al. "Mapping and dynamics of regulatory DNA and transcription factor
686 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).

- 688 [26] Stefan Grob, Marc W Schmid, and Ueli Grossniklaus. “Hi-C Analysis in Arabidopsis Identifies
689 the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila”. In: *Molecular Cell*
690 (Aug. 2014), pp. 1–16. DOI: [10.1016/j.molcel.2014.07.009](https://doi.org/10.1016/j.molcel.2014.07.009).
- 691 [27] Matthew W Horton et al. “Genome-wide patterns of genetic variation in worldwide Arabidopsis
692 thaliana accessions from the RegMap panel”. In: *Nature Genetics* 44.2 (Feb. 2012), pp. 212–
693 216. DOI: [10.1038/ng.1042](https://doi.org/10.1038/ng.1042).
- 694 [28] Jing Zhao et al. “A Burden of Rare Variants Associated with Extremes of Gene Expression
695 in Human Peripheral Blood”. In: *The American Journal of Human Genetics* 98.2 (Feb. 2016),
696 pp. 299–309. DOI: [10.1016/j.ajhg.2015.12.023](https://doi.org/10.1016/j.ajhg.2015.12.023).
- 697 [29] Guodong Wang et al. “A genome-wide functional investigation into the roles of receptor-like
698 proteins in Arabidopsis.” In: *Plant Physioloxy* 147.2 (June 2008), pp. 503–517. DOI: [10.1104/pp.108.119487](https://doi.org/10.1104/pp.108.119487).
- 700 [30] María José Aranzana et al. “Genome-Wide Association Mapping in Arabidopsis Identifies
701 Previously Known Flowering Time and Pathogen Resistance Genes”. In: *PLoS Genetics* 1.5
702 (2005), e60–9. DOI: [10.1371/journal.pgen.0010060](https://doi.org/10.1371/journal.pgen.0010060).
- 703 [31] Akira Katoh et al. “Early steps in the biosynthesis of NAD in Arabidopsis start with aspartate
704 and occur in the plastid.” In: *Plant Physioloxy* 141.3 (July 2006), pp. 851–857. DOI: [10.1104/pp.106.081091](https://doi.org/10.1104/pp.106.081091).
- 706 [32] Xiaoyu Zhang et al. “Genome-wide High-Resolution Mapping and Functional Analysis of DNA
707 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).
- 709 [33] Daniel Zilberman et al. “Genome-wide analysis of Arabidopsis thaliana DNA methylation
710 uncovers an interdependence between methylation and transcription.” In: *Nature Genetics* 39.1
711 (Jan. 2007), pp. 61–69. DOI: [10.1038/ng1929](https://doi.org/10.1038/ng1929).
- 712 [34] Shawn J Cokus et al. “Shotgun bisulphite sequencing of the Arabidopsis genome reveals
713 DNA methylation patterning”. In: *Nature* 452.7184 (Feb. 2008), pp. 215–219. DOI: [10.1038/nature06745](https://doi.org/10.1038/nature06745).
- 715 [35] Ryan Lister et al. “Highly integrated single-base resolution maps of the epigenome in Arabidopsis.”
716 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).
- 717 [36] Eric J Richards. “Inherited epigenetic variation—revisiting soft inheritance.” In: *Nature Reviews
718 Genetics* 7.5 (May 2006), pp. 395–401. DOI: [10.1038/nrg1834](https://doi.org/10.1038/nrg1834).
- 719 [37] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature
720 Methods* 9.4 (Mar. 2012), pp. 357–359. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923).

- 721 [38] Gregory G Faust and Ira M Hall. “SAMBLASTER: fast duplicate marking and structural vari-
722 ant read extraction.” In: *Bioinformatics* 30.17 (Sept. 2014), pp. 2503–2505. DOI: [10.1093/bioinformatics/btu314](https://doi.org/10.1093/bioinformatics/btu314).
- 724 [39] Gregory G Faust and Ira M Hall. “YAHA: fast and flexible long-read alignment with optimal
725 breakpoint detection.” In: *Bioinformatics* 28.19 (Oct. 2012), pp. 2417–2424. DOI: [10.1093/bioinformatics/bts456](https://doi.org/10.1093/bioinformatics/bts456).
- 727 [40] Heng Li et al. “The Sequence Alignment/Map format and SAMtools.” In: *Bioinformatics* 25.16
728 (Aug. 2009), pp. 2078–2079. DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
- 729 [41] Ryan K Dale, Brent S Pedersen, and Aaron R Quinlan. “Pybedtools: a flexible Python library
730 for manipulating genomic datasets and annotations.” In: *Bioinformatics* 27.24 (Dec. 2011),
731 pp. 3423–3424. DOI: [10.1093/bioinformatics/btr539](https://doi.org/10.1093/bioinformatics/btr539).
- 732 [42] Aaron R Quinlan and Ira M Hall. “BEDTools: a flexible suite of utilities for comparing genomic
733 features.” In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. DOI: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).
- 734 [43] Ronald H Plasterk. “The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*.”
735 In: *The EMBO Journal* 10.7 (July 1991), pp. 1919–1925.
- 736 [44] Christoph Bartenhagen and Martin Dugas. “RSVSim: an R/Bioconductor package for the
737 simulation of structural variations.” In: *Bioinformatics* 29.13 (July 2013), pp. 1679–1681. DOI:
738 [10.1093/bioinformatics/btt198](https://doi.org/10.1093/bioinformatics/btt198).
- 739 [45] K Edwards, C Johnstone, and C Thompson. “A simple and rapid method for the preparation of
740 plant genomic DNA for PCR analysis.” In: *Nucleic Acids Research* 19.6 (Mar. 1991), p. 1349.
- 741 [46] Ryan M Layer et al. “LUMPY: a probabilistic framework for structural variant discovery.” In:
742 *Genome Biology* 15.6 (2014), R84. DOI: [10.1186/gb-2014-15-6-r84](https://doi.org/10.1186/gb-2014-15-6-r84).
- 743 [47] Peter J A Cock et al. “Biopython: freely available Python tools for computational molecular
744 biology and bioinformatics.” In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. DOI: [10.1093/bioinformatics/btp163](https://doi.org/10.1093/bioinformatics/btp163).
- 746 [48] John D Storey and Robert Tibshirani. “Statistical significance for genomewide studies.” In:
747 *Proceedings of the National Academy of Sciences of the United States of America* 100.16 (Aug.
748 2003), pp. 9440–9445. DOI: [10.1073/pnas.1530509100](https://doi.org/10.1073/pnas.1530509100).

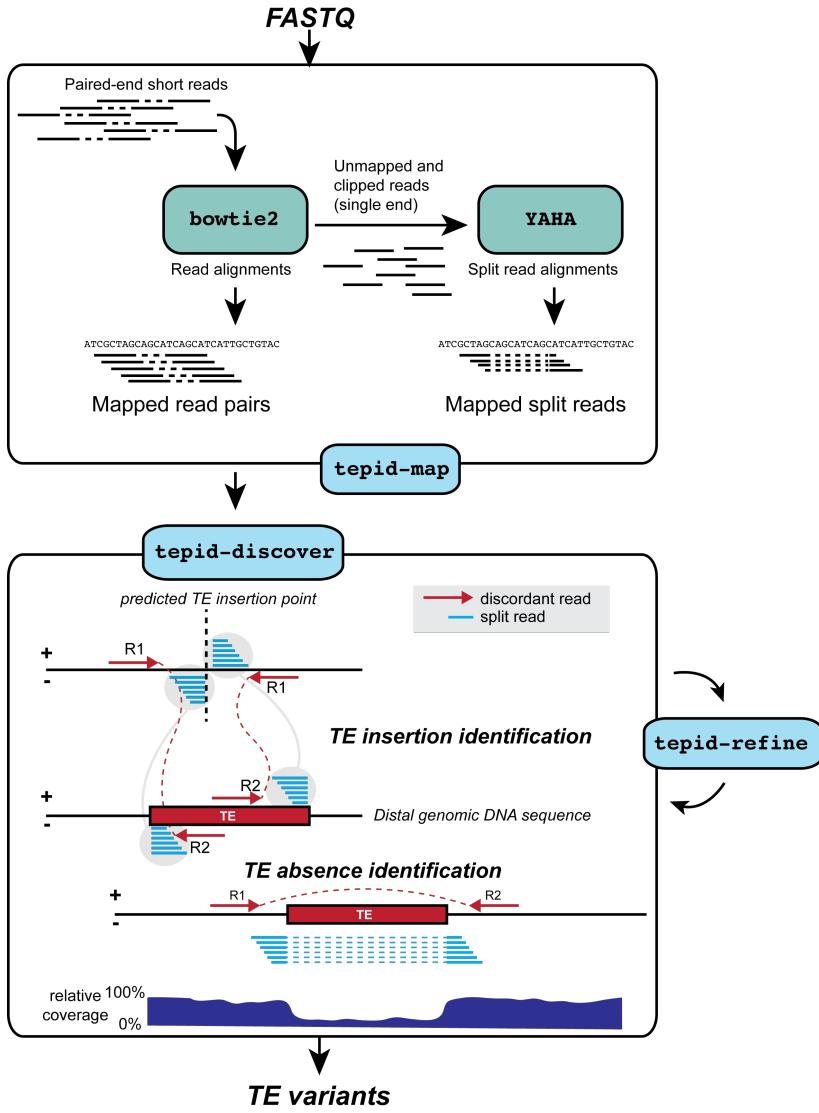


Figure 1: TE variant discovery pipeline

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.

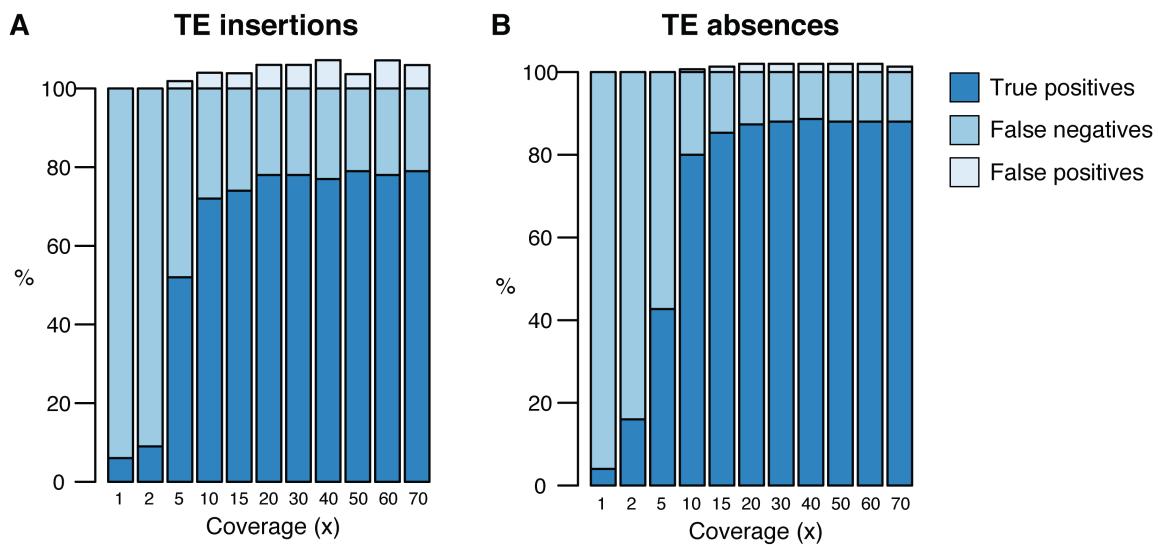


Figure 1: figure supplement 1

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

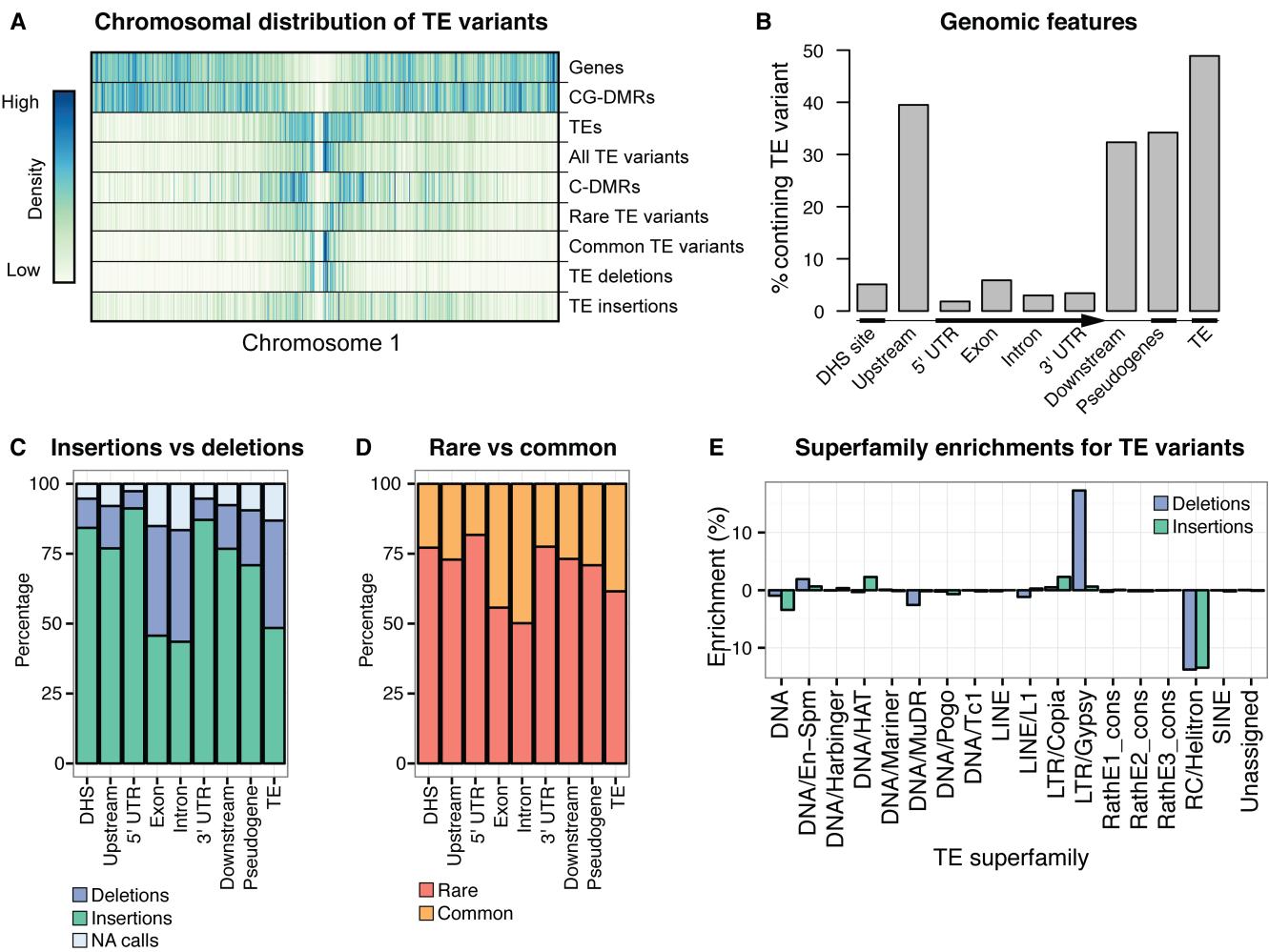


Figure 2: Extensive novel genetic diversity uncovered by TE variant analysis

- (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

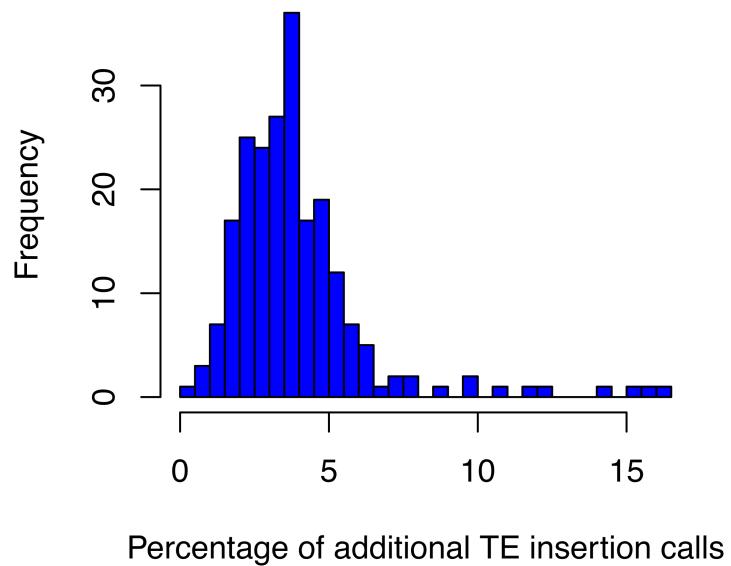


Figure 2: figure supplement 1

767 Percentage of total TE insertion calls that were made due to the TEPID refinement step for each
768 accession in the population.

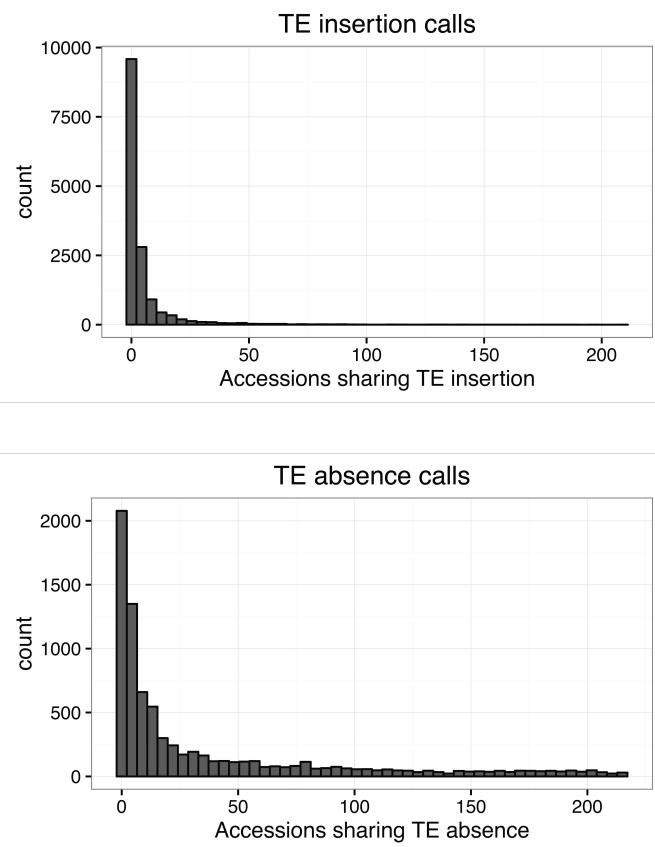


Figure 2: figure supplement 2

769 Number of accessions sharing TE variants identified by TEPID.

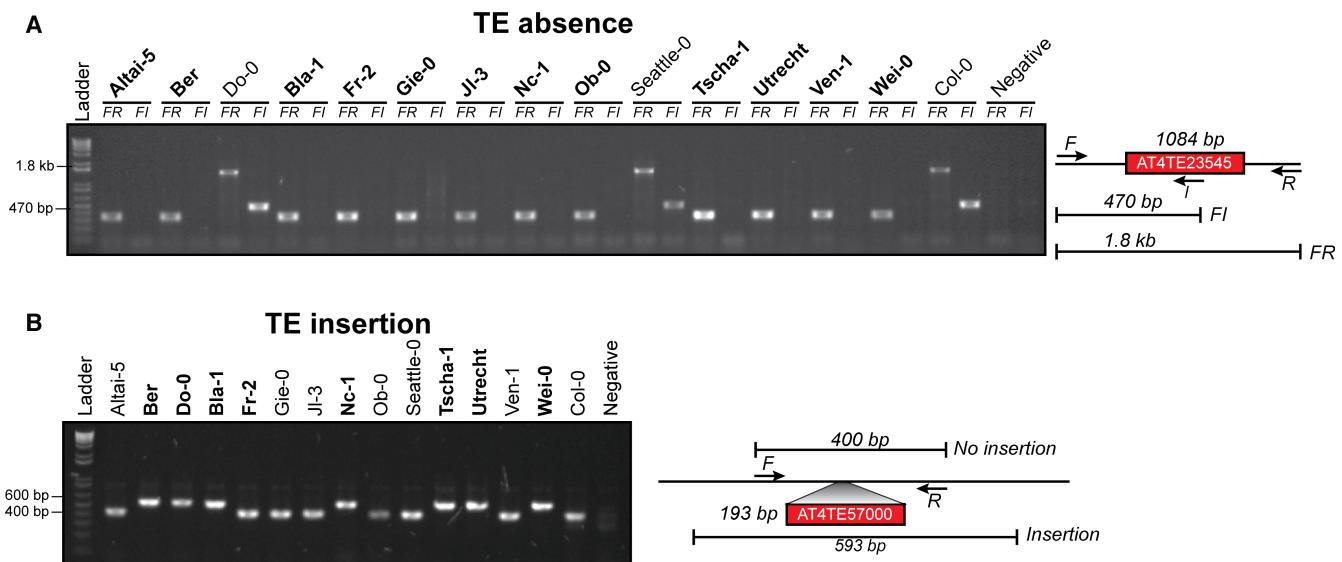


Figure 2: figure supplement 3

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

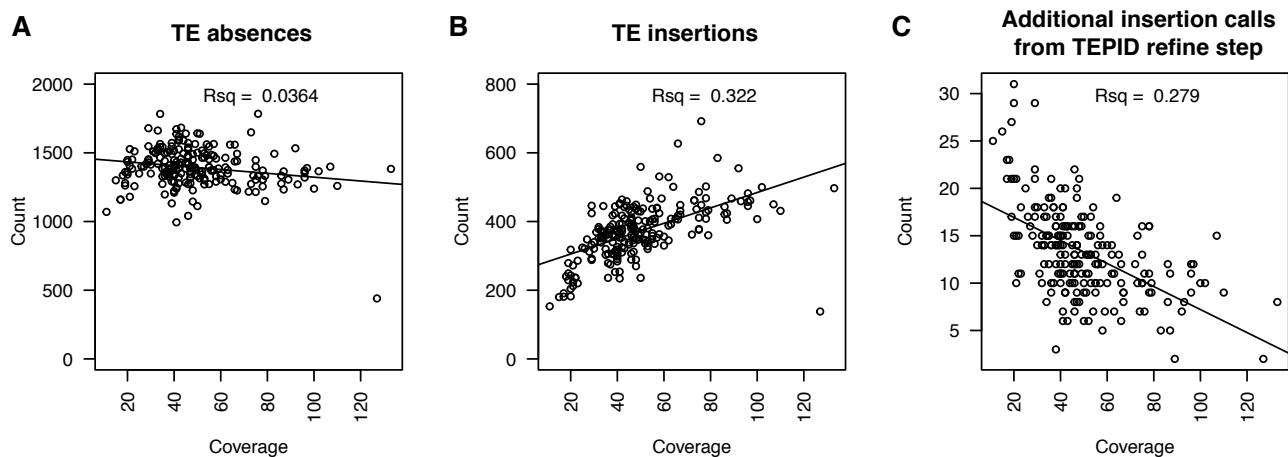


Figure 2: figure supplement 4

- 777 (A) Number of TE absence variants identified versus the sequencing depth of coverage for each
778 accession.
- 779 (B) Number of TE insertion variants identified versus the sequencing depth of coverage for each
780 accession.
- 781 (C) Number of additional TE insertion calls made due to the TEPID refinement step versus se-
782 quencing depth of coverage for all accessions.

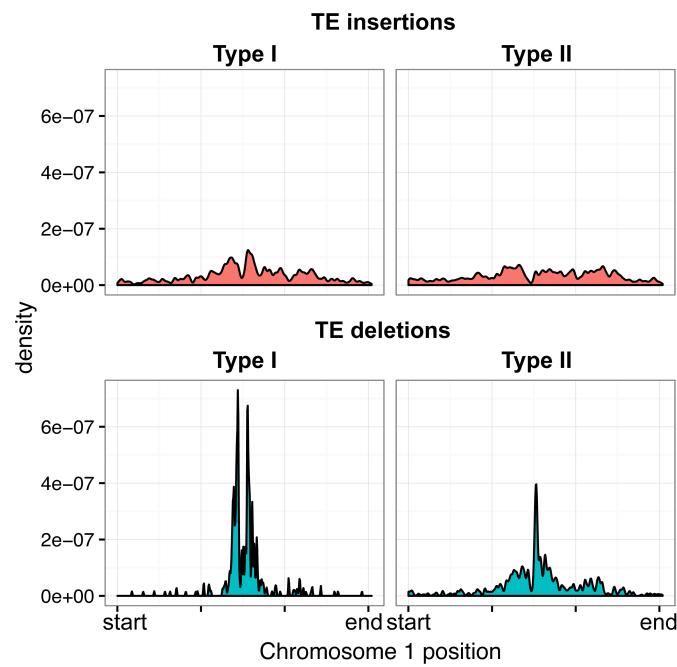
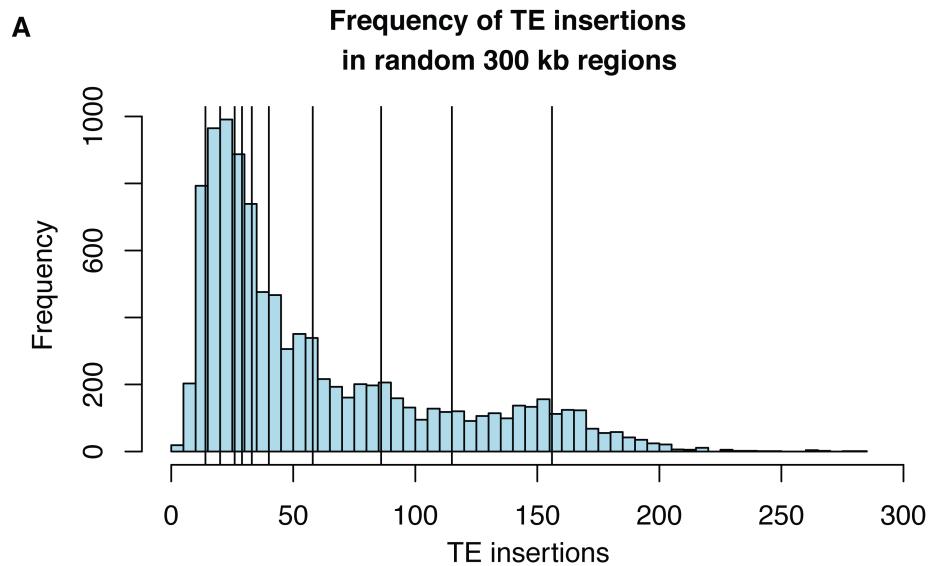


Figure 2: figure supplement 5

783 Distribution of Type I and Type II elements over chromosome 1, for TE insertions and TE deletions.



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 6. Frequency of TE insertion in the *KNOT* region

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

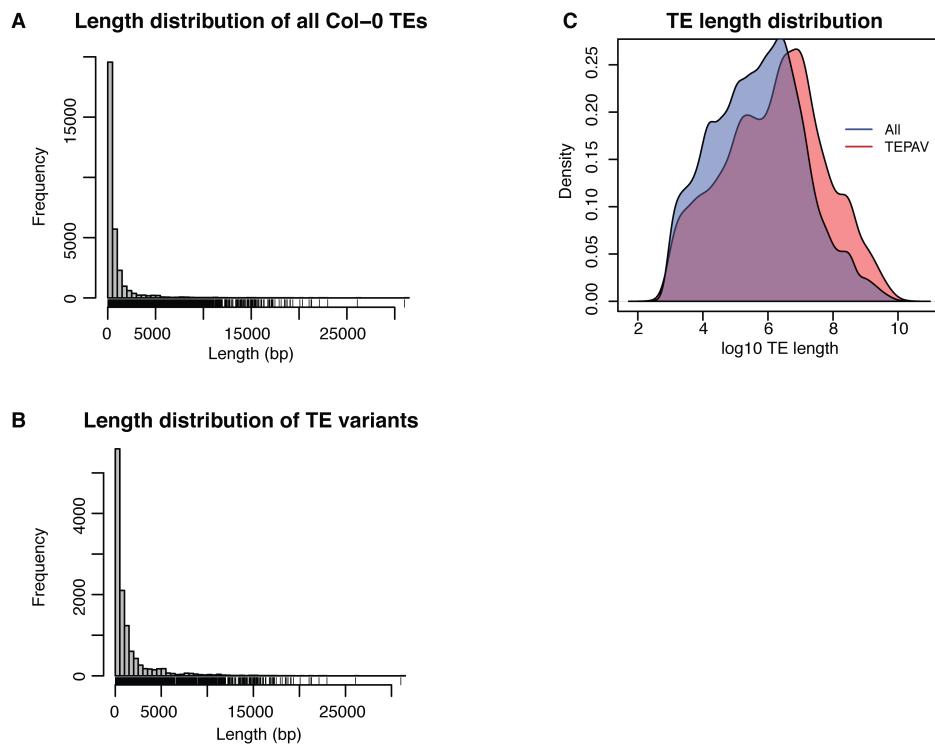


Figure 2: figure supplement 7. Length distribution for all Col-0 TEs and all TE variants

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

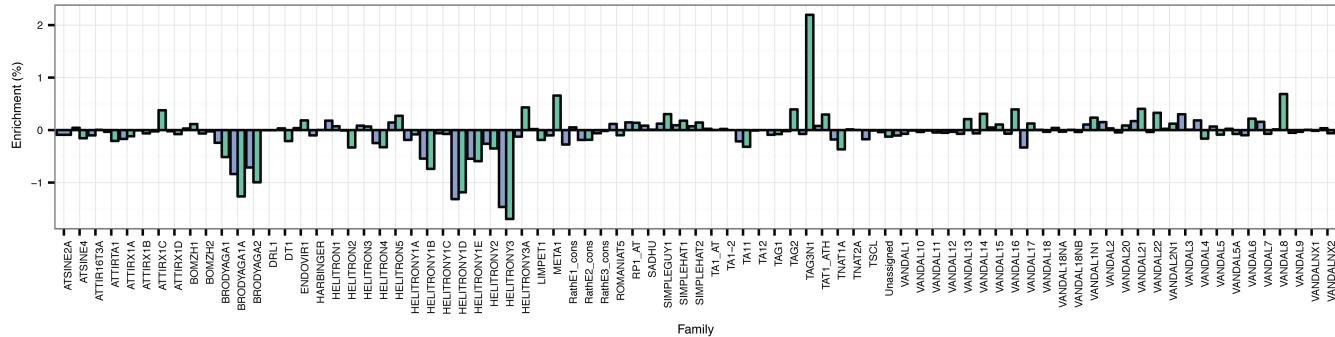
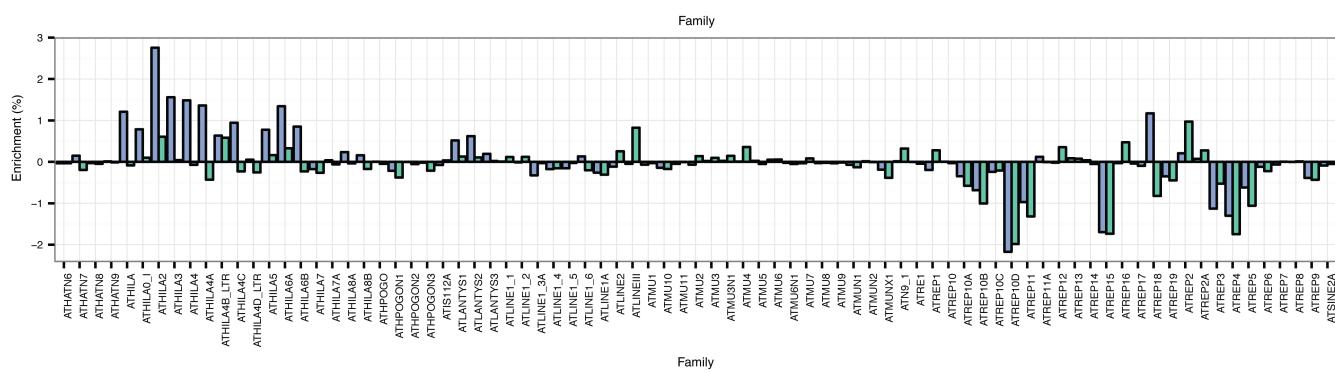
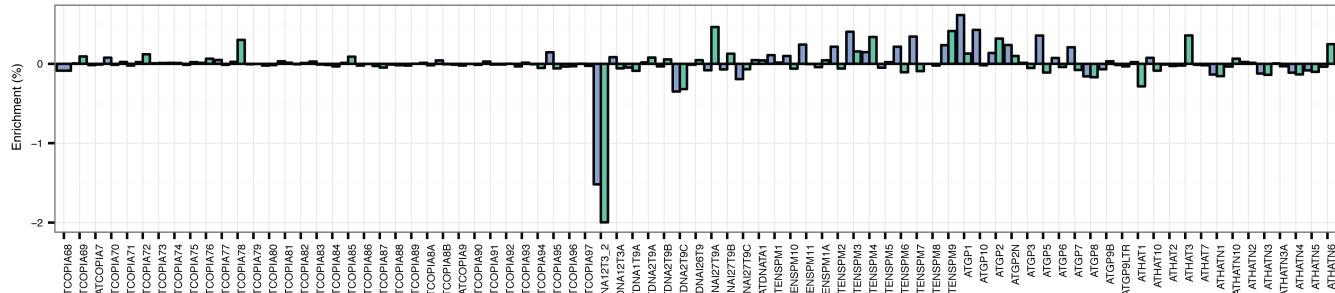
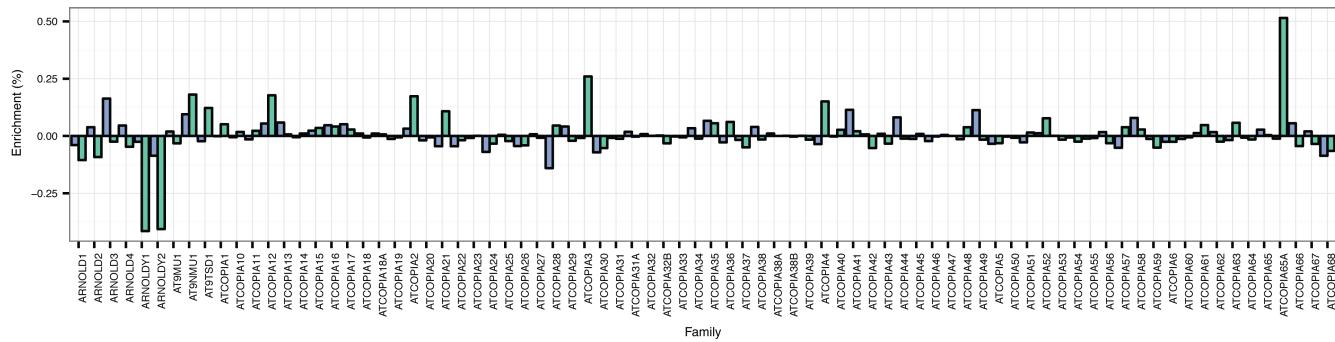


Figure 2: figure supplement 8

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

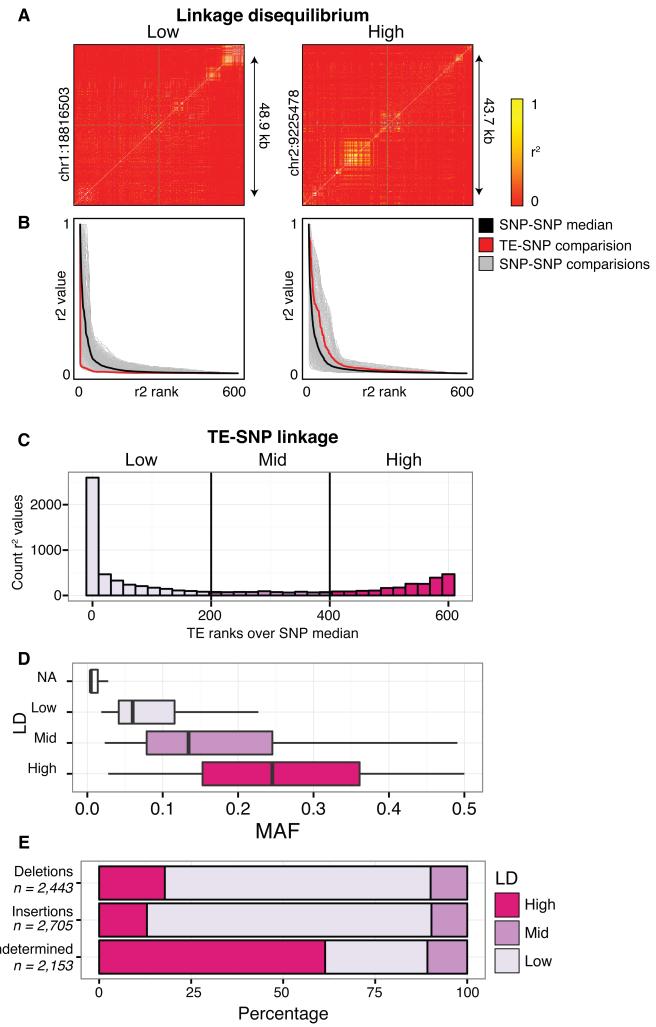


Figure 3: Patterns of TE-SNP linkage

- (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.

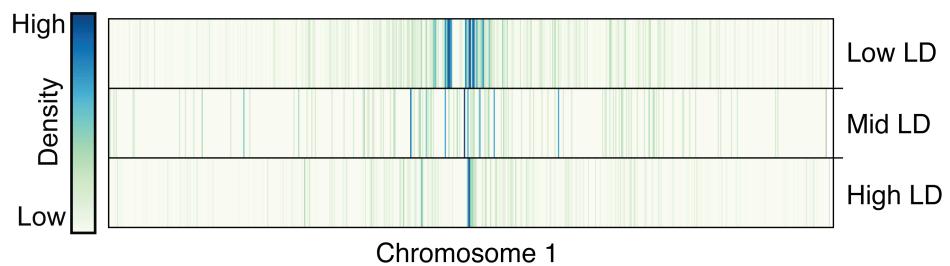


Figure 3: figure supplement 1

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

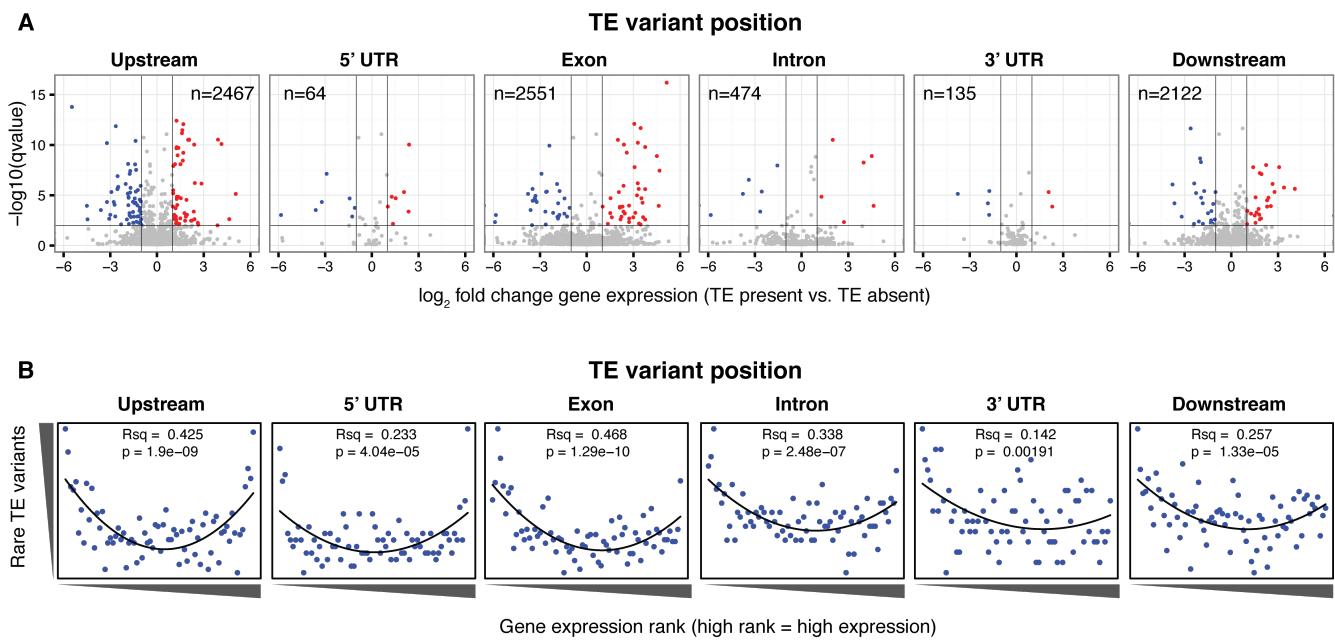


Figure 4: Differential transcript abundance associated with TE variant presence/absence

- 808 (A) Volcano plots showing transcript abundance differences for genes associated with TE insertion
 809 variants at different positions, indicated in the plot titles. Genes with significantly different
 810 transcript abundance in accessions with a TE insertion compared to accessions without a TE
 811 insertion are colored blue (lower transcript abundance in accessions containing TE insertion) or
 812 red (higher transcript abundance in accessions containing TE insertion). Vertical lines indicate
 813 ± 2 fold change in FPKM. Horizontal line indicates the 1% FDR.
- 814 (B) Relationship between TE rare variant counts and gene expression rank. Plot shows the
 815 cumulative number of rare TE variants in equal-sized bins for gene expression ranks, from the
 816 lowest-ranked accession (left) to the highest-ranked accession (right). Lines indicate the fit of a
 817 quadratic model.

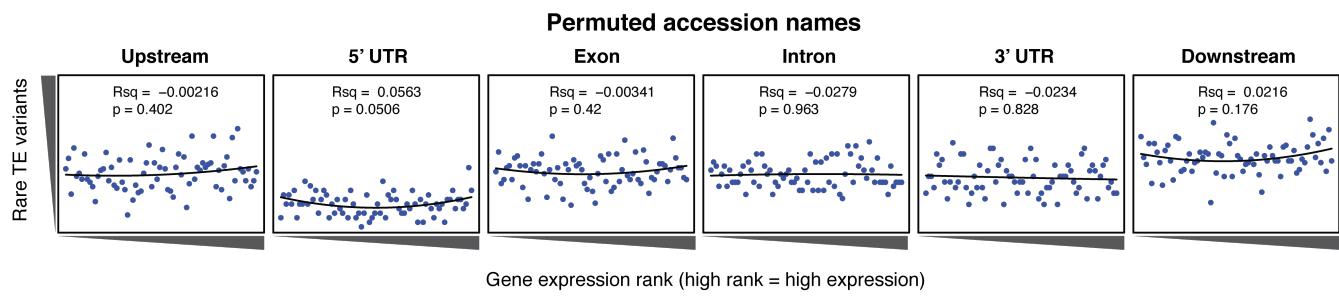


Figure 4: figure supplement 1

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

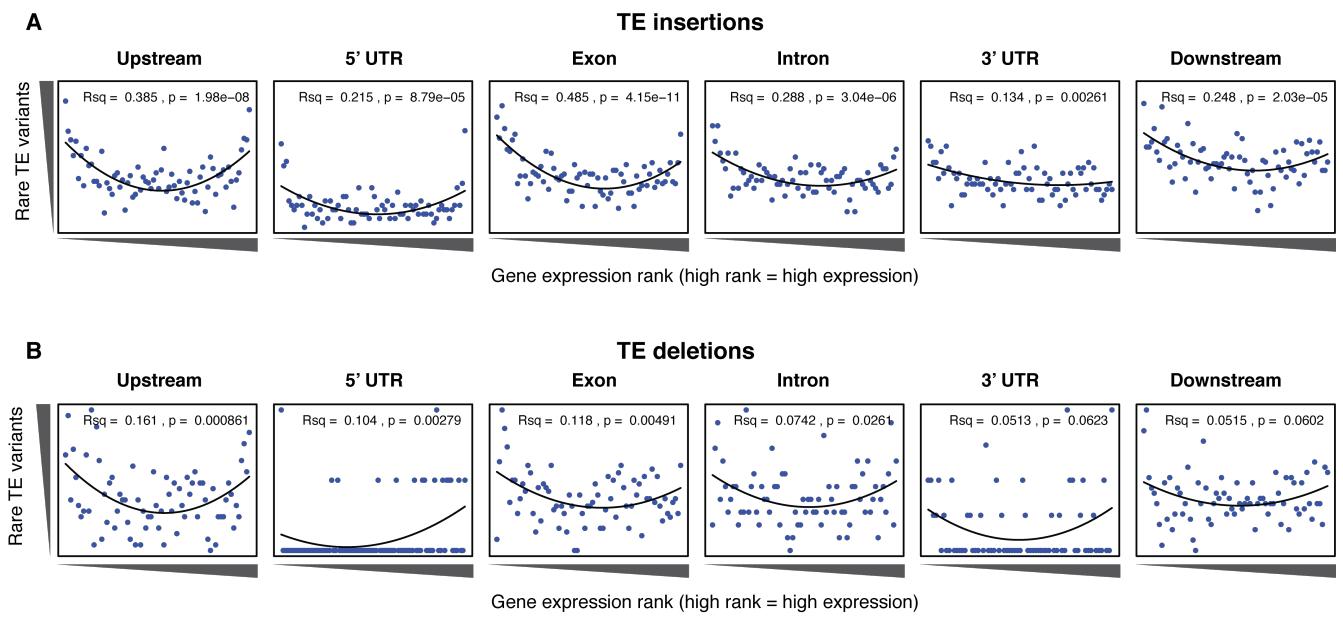


Figure 4: figure supplement 2

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

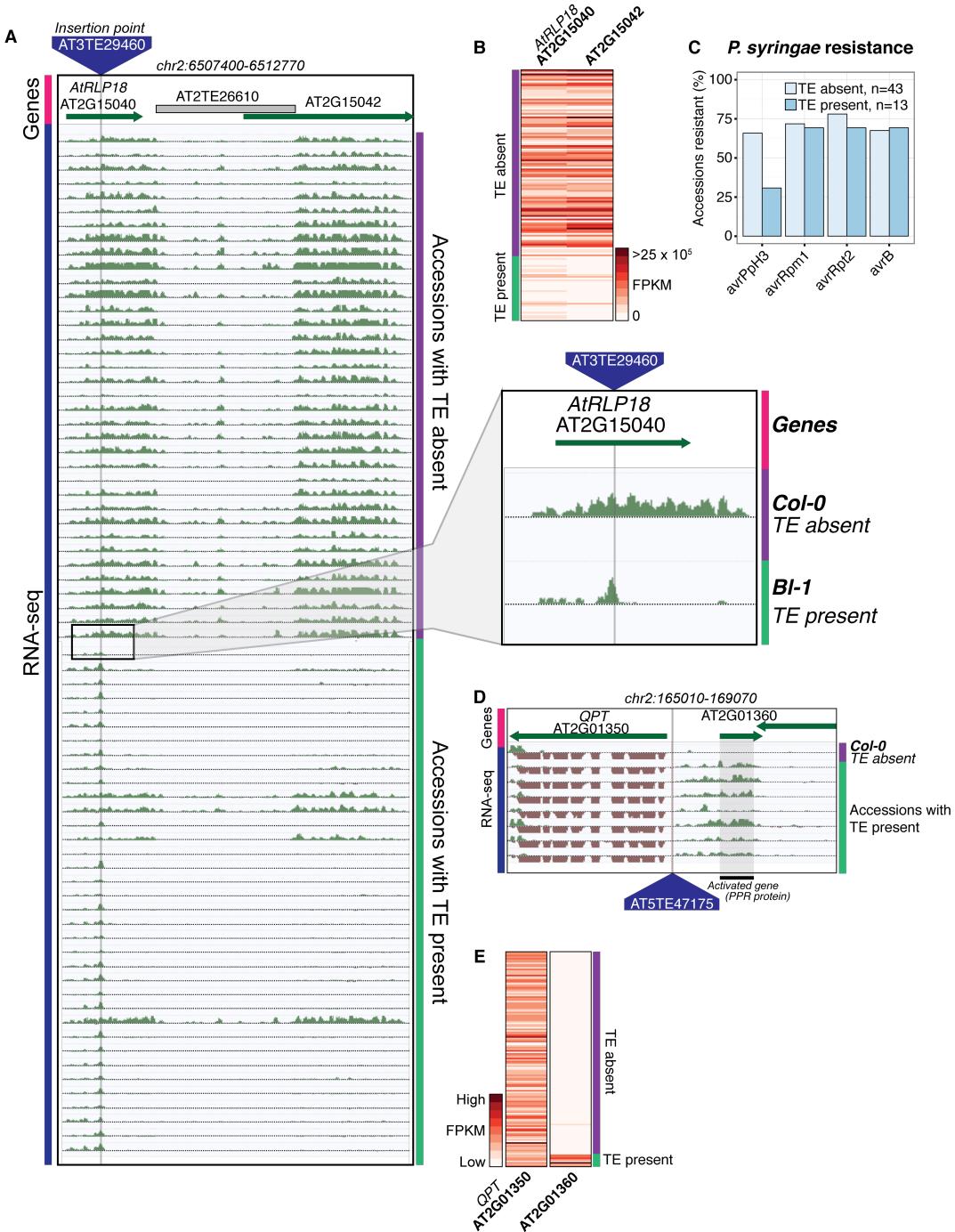


Figure 5: Effects of TE variants on local gene expression

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

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

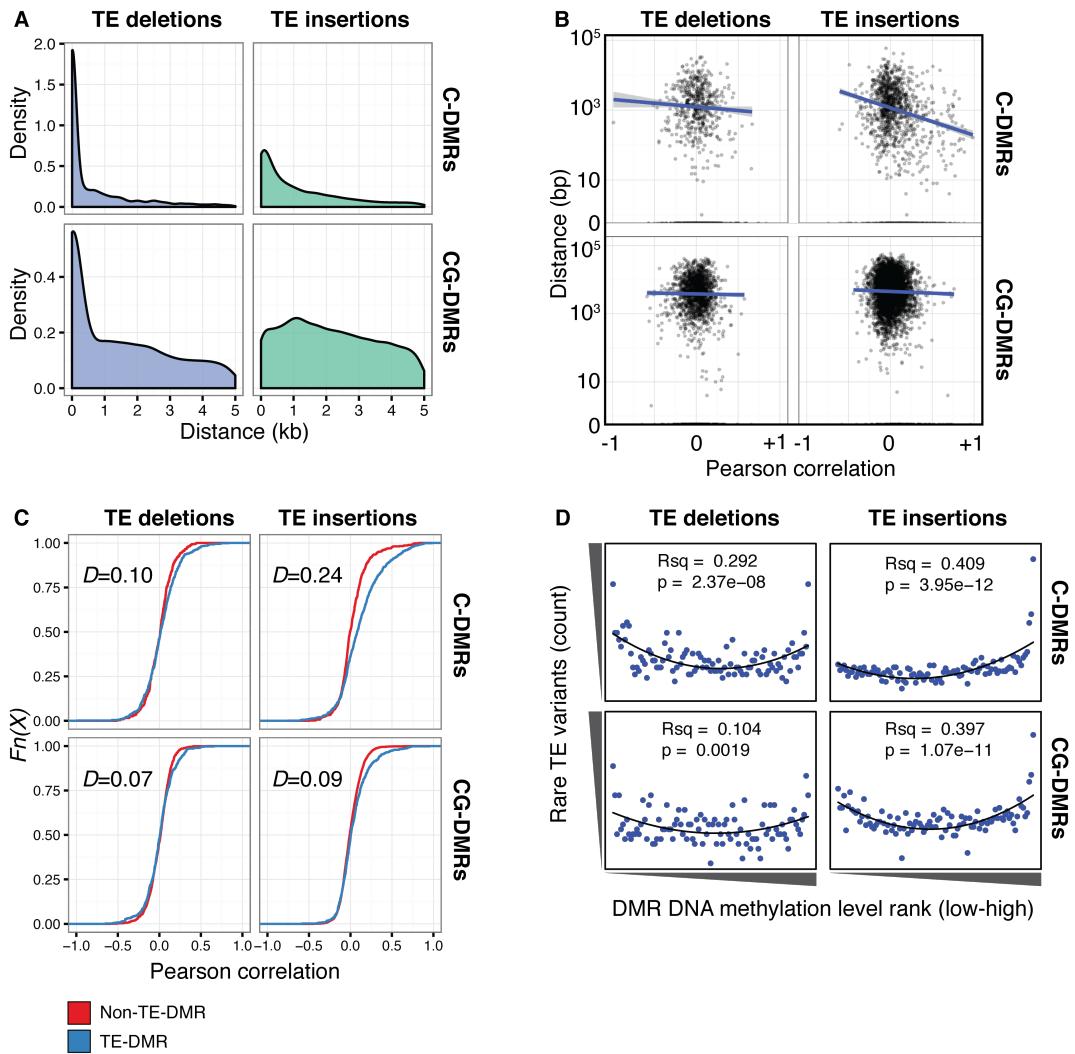


Figure 6: TE variants are associated with nearby DMR methylation levels

- (A) Distribution of distances from TE variants to the nearest population DMR, for TE deletions and TE insertions, 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 log₁₀ 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.

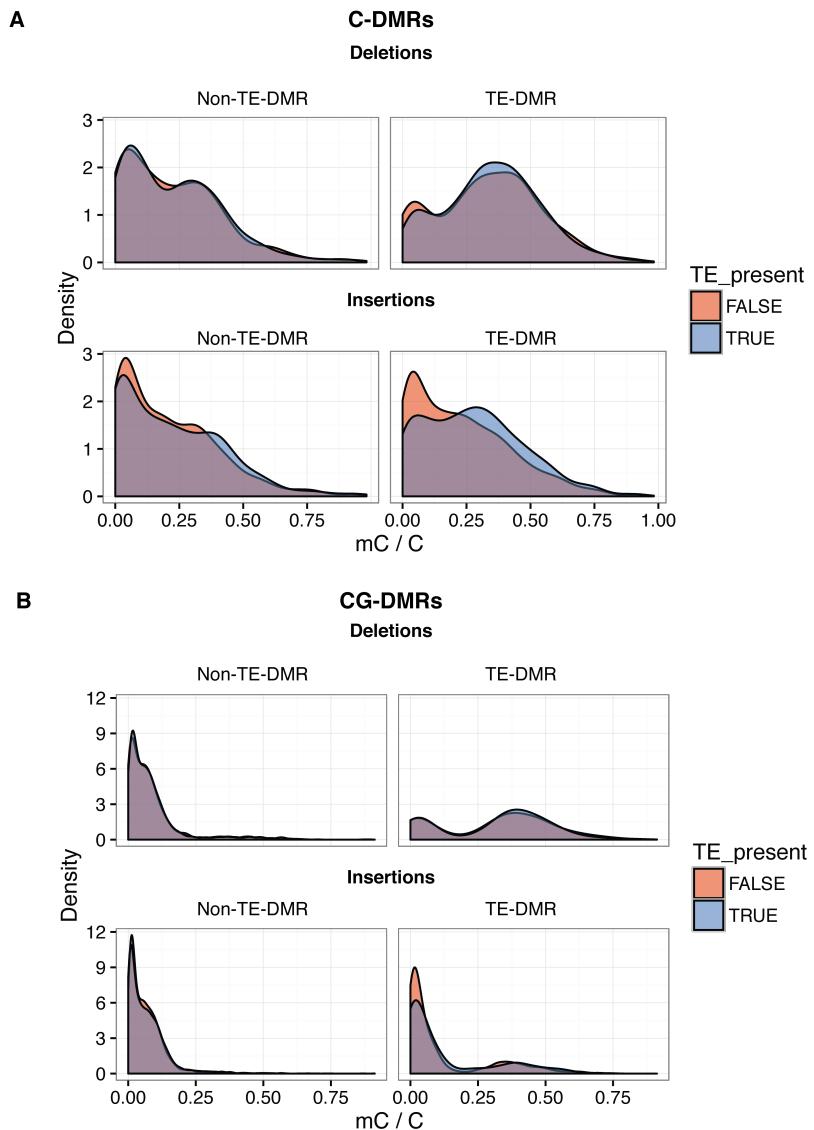


Figure 6: figure supplement 1

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

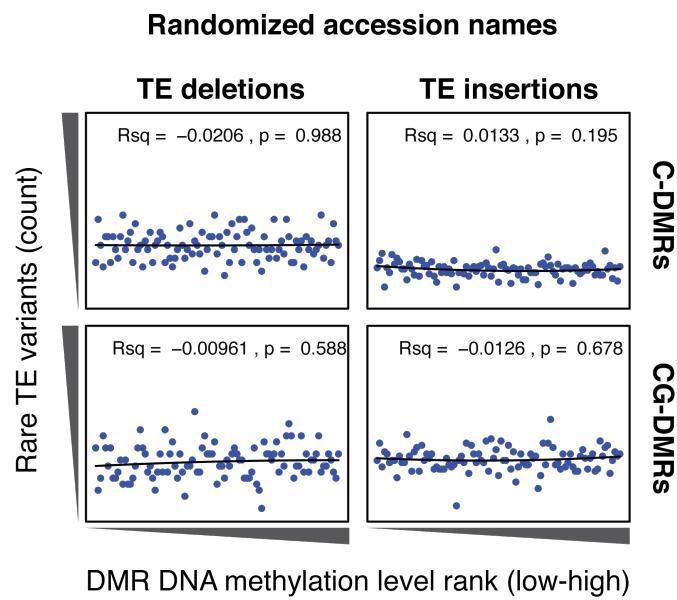


Figure 6: figure supplement 2

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

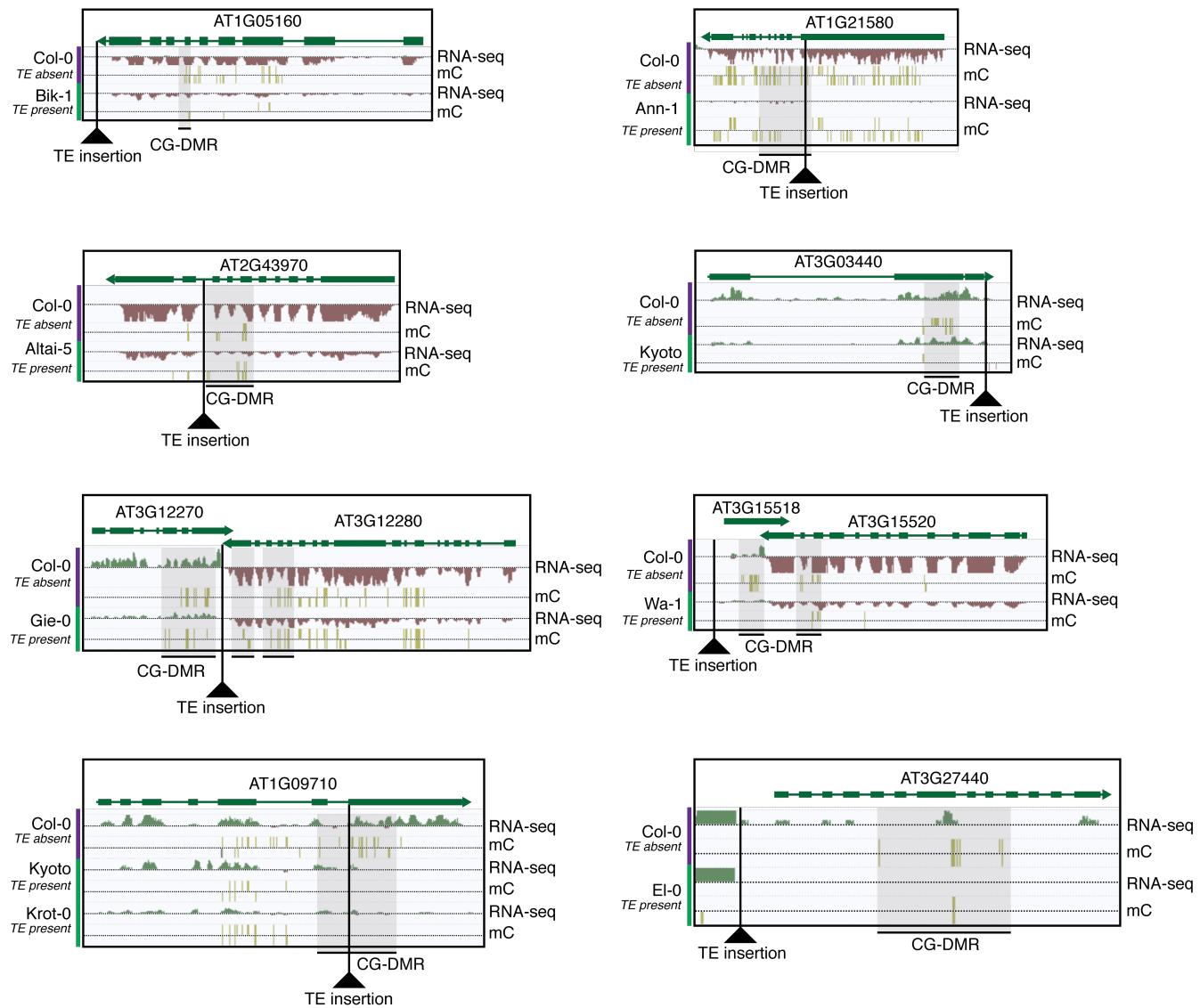


Figure 6: figure supplement 3

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

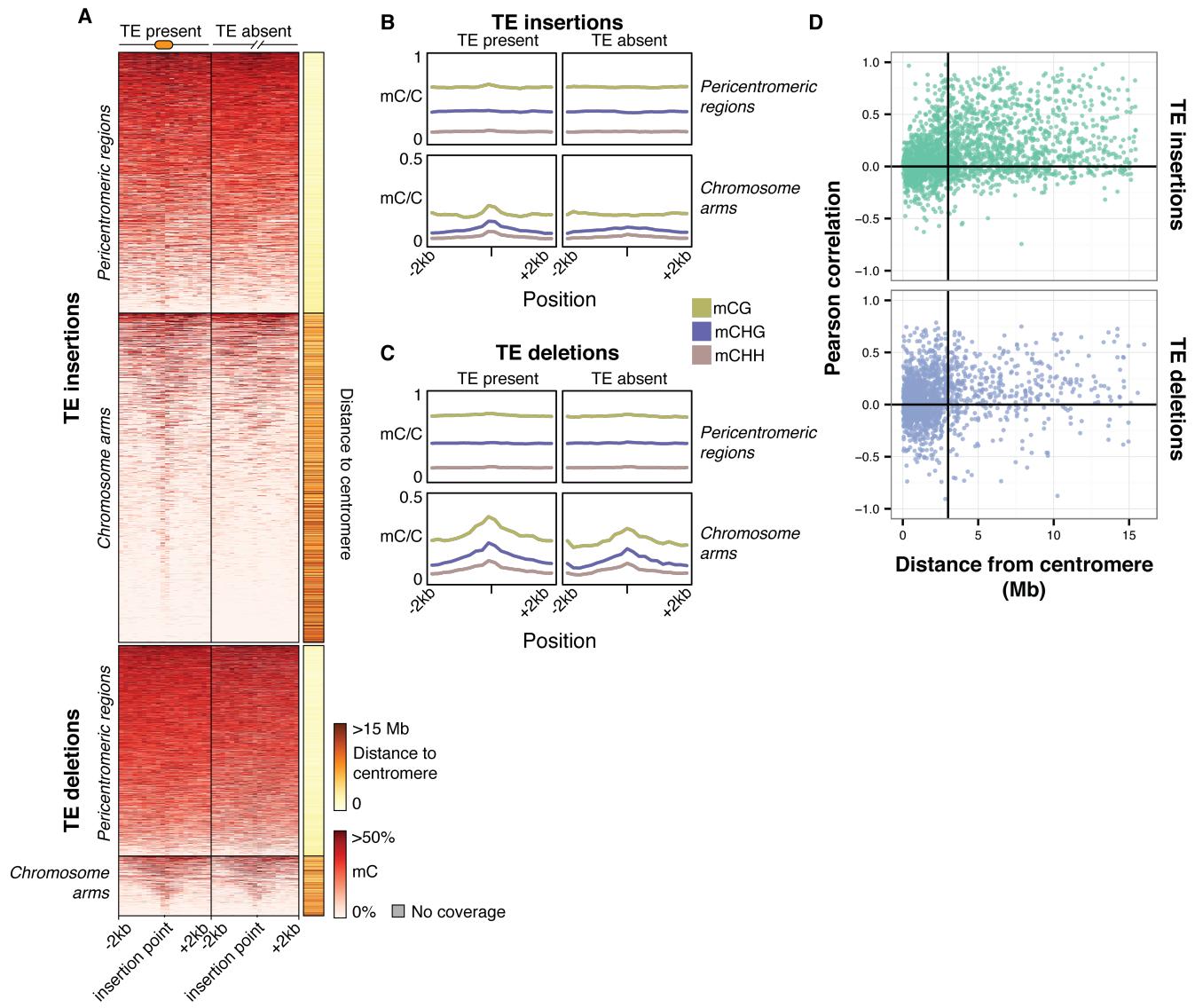


Figure 7: Local patterns of DNA methylation surrounding TE variant sites

- (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.

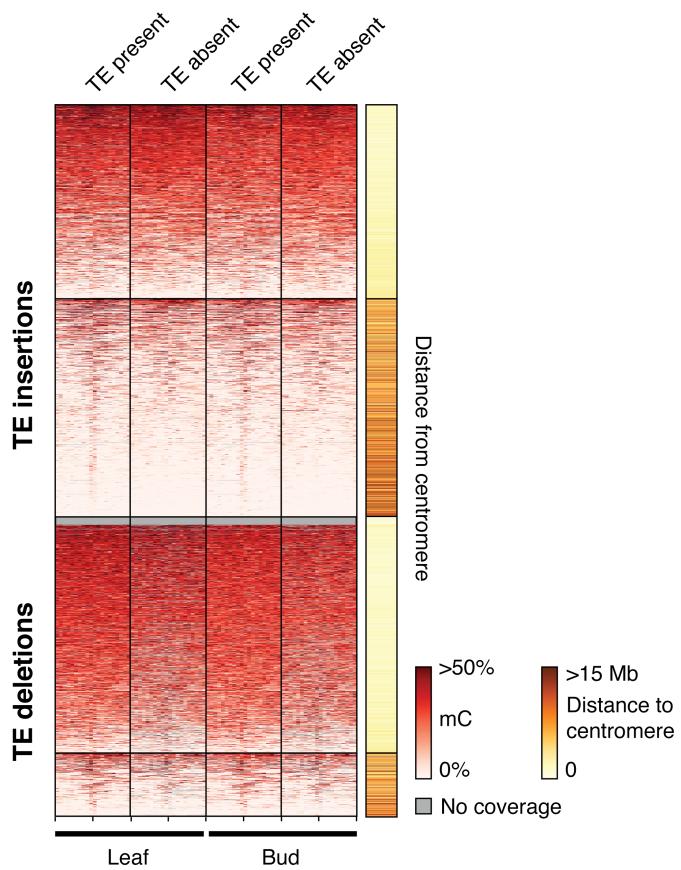


Figure 7: figure supplement 1

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

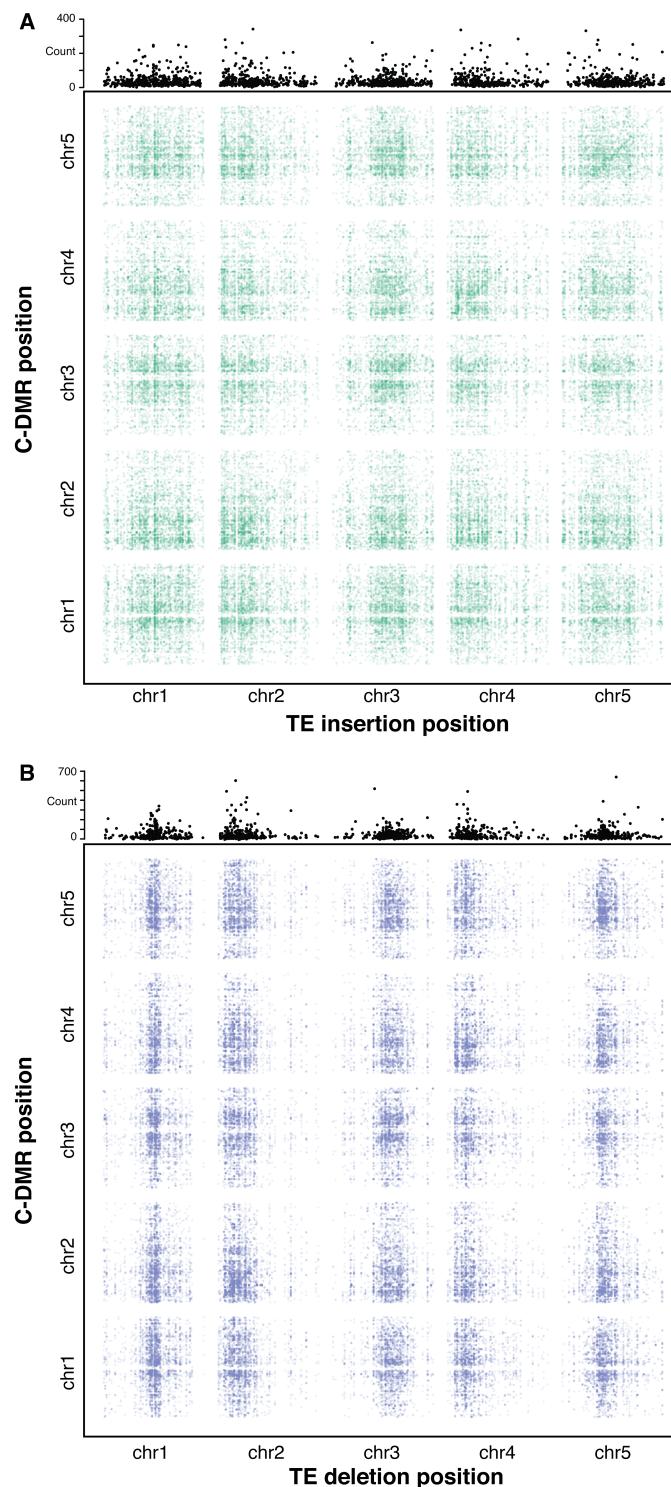


Figure 8: Association scan between TE variants and C-DMR methylation variation

- 871 (A) Significant correlations between TE insertions and C-DMR DNA methylation level. Points
 872 show correlations between individual TE-DMR pairs that were more extreme than any of 500
 873 permutations of the DMR data. Top plots show the total number of significant correlations for
 874 each TE insertion across the whole genome.
- 875 (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
Ler 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	17	16	0.0079	4.1	16	0.0041
TE insertions	28	26	0.0089	9.1	26	0.0047
NA calls	8.7	6.2	0.0053	1.6	6.2	0.0027
Total	54	48	0.01	15	48	0.0054