

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

Tim Stuart<sup>1</sup>, Steven R. Eichten<sup>2</sup>, Jonathan Cahn<sup>1</sup>, Yuliya Karpievitch<sup>1</sup>, Justin Borevitz<sup>2</sup> and Ryan Lister<sup>1</sup>

<sup>1</sup>ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, Australia

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

Corresponding author: Ryan Lister [ryan.listner@uwa.edu.au](mailto:ryan.listner@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)

## <sup>1</sup> Abstract

2 Variation in the presence or absence of transposable elements (TEs) is a major source of genetic  
3 variation between individuals. Here, we identified 23,095 TE presence/absence variants between 216  
4 wild *Arabidopsis* accessions. Most variants are rare, and we find a burden of rare variants associated  
5 with extremes of both gene expression and DNA methylation levels within the population. Of  
6 the common alleles identified, two thirds were not in linkage disequilibrium with nearby SNPs,  
7 implicating these TE variants as a source of novel genetic diversity. Nearly 200 common TE variants  
8 were associated with significantly altered expression of nearby genes, with potential functional  
9 consequences including decreased pathogen resistance. A major fraction of inter-accession DNA  
10 methylation differences were associated with nearby TE insertion variants, with DNA, indicating  
11 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  
16 reaction, and DNA transposons (class II elements), which transpose via either a cut-paste or, in  
17 the case of Helitrons, a rolling circle mechanism with no RNA intermediate [1]. TE activity poses  
18 mutagenic potential as a TE insertion may disrupt essential regions of the genome. Consequently,  
19 safeguards have evolved in order to suppress this activity. These safeguards include epigenetic  
20 transcriptional silencing mechanisms, chiefly involving the methylation of cytosine nucleotides (DNA  
21 methylation) to produce 5-methylcytosine (mC), a mark that can signal transcriptional silencing of  
22 the methylated locus. In *Arabidopsis thaliana* (Arabidopsis), DNA methylation occurs in three  
23 DNA sequence contexts: mCG, mCHG, and mCHH, where H is any base but G. Establishment of  
24 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  
27 silencing of TEs. Loss of DNA methylation due to mutations in genes essential for DNA methylation  
28 establishment and maintenance leads to expression of previously silent TEs, and in some cases  
29 transposition [2, 4–8].

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

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

56 and identified widespread TE variation amongst these accessions [20].

## 57 Results

### 58 Computational identification of TE presence/absence variation

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

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

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

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

## 106 Abundant TE positional variation among natural *Arabidopsis* populations

### 107

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

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

139 TE insertions and deletions were distributed throughout chromosome 1 in a pattern that was

similar to the distribution of all Col-0 TEs (Figure 2A). TE deletions and common TE variants were found in similar chromosomal regions, as deletion variants represent the rare loss of common variants. TE deletions and common variants were more highly enriched in the pericentromeric regions than rare variants or TE insertions. The distribution of rare TE variants and TE insertions was somewhat similar 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

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

## 204 TE variants affect gene expression

205 To determine whether the TE variants identified affected nearby gene expression, we compared the  
206 steady state transcript abundance within mature leaf tissue between accessions with and without  
207 TE insertions or deletions, for genes with TE variants located in the 2 kb gene upstream region,  
208 5' UTR, exon, intron, 3' UTR or 2 kb downstream region (Figure 4A). While the steady state  
209 transcript abundance of most genes appeared to be unaffected by the presence of a TE, 196 genes  
210 displayed significant differences in transcript abundance linked with the presence of a TE variant,  
211 indicating a role for these variants in the local regulation of gene expression (1% false discovery  
212 rate; greater than 2-fold change in transcript abundance; Figure 4A, Figure 4 - source data 1). We  
213 did not find any functional category enrichments in this set of differentially expressed genes. As  
214 rare TE variants with a MAF less than 3% may also be associated with a difference in transcript  
215 abundance, but were unable to be statistically tested due to their rarity, we performed a burden  
216 test for enrichment of rare variants in the extremes of expression [28]. Briefly, this method counts  
217 the frequency of rare variants within each gene expression rank in the population, and aggregates  
218 this information over the entire population to determine whether an enrichment of rare variants  
219 exists within any gene expression rank. We found a strong enrichment for gene expression extremes  
220 for TE variants in the all gene features tested (Figure 4B). While TE variants in gene upstream  
221 regions showed a strong enrichment of both high and low gene expression ranks, TE variants in  
222 exons or gene downstream regions seemed to have a stronger enrichment for low expression ranks  
223 than high ranks. Randomization of the accession names removed these enrichments completely  
224 (Figure 4 - figure supplement 1), and there was little difference between TE insertions and TE  
225 deletions in the gene expression rank enrichments found (Figure 4 - figure supplement 2). This rare  
226 variant analysis further indicated that TE variants may alter the transcript abundance of nearby

227 genes.

228 As both increases and decreases in transcript abundance of nearby genes were observed for TE  
229 variants within each gene feature, it appears to be difficult to predict the impact a TE variant may  
230 have on nearby gene expression. Furthermore, gene-level transcript abundance measurements may  
231 fail to identify the potential positional effect of TE variants upon transcription. To more closely  
232 examine changes in transcript abundance associated with TE variants among the accessions, we  
233 inspected a subset of TE variant sites and identified TE variants that appear to have an impact on  
234 transcriptional patterns beyond changes in total transcript abundance from a nearby gene. For  
235 example, the presence of a TE insertion within an exon of *AtRLP18* (AT2G15040) was associated  
236 with truncation of the transcripts at the TE insertion site in accessions possessing the TE variant,  
237 as well as silencing of a downstream gene encoding a leucine-rich repeat protein (AT2G15042)  
238 (Figure 5A, B). Both genes had significantly lower transcript abundance in accessions containing  
239 the TE insertion ( $p < 5.8 \times 10^{-10}$ , Mann-Whitney U test). *AtRLP18* is reported to be involved in  
240 bacterial resistance, with the disruption of this gene by T-DNA insertion mediated mutagenesis  
241 resulting in increased susceptibility to the bacterial plant pathogen *Pseudomonas syringae* [29]. We  
242 examined pathogen resistance phenotype data [30], and found that accessions containing the TE  
243 insertion in the *AtRLP18* exon were more often sensitive to infection by *Pseudomonas syringae*  
244 transformed with *avrPph3* genes (Figure 5C). This suggests that the accessions containing this TE  
245 insertion within *AtRLP18* may have an increased susceptibility to certain bacterial pathogens.

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

## 264 TE variants explain many DNA methylation differences between acces- 265 sions

266 As TEs are frequently highly methylated in *Arabidopsis* [32–35], we next assessed the DNA  
267 methylation state surrounding TE variant sites to determine whether TE variants might be  
268 responsible for some of the differences in DNA methylation patterns previously observed between  
269 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 the presence/absence of the nearest TE variant. We observed a negative correlation between the distance from a C-DMR to the nearest TE insertion and the correlation between the DNA methylation level at the C-DMR with the presence/absence of the TE insertion (Figure 6B). This suggests a distance-dependent effect of TE insertion presence on C-DMR methylation. In contrast, we found no such relationship for TE deletions on C-DMRs, or for insertions or deletions on CG-DMRs (Figure 6B). DNA methylation levels at C-DMRs located within 1 kb of a TE insertion (TE-DMRs) were more often positively correlated with the presence/absence of a TE insertion than the DNA methylation levels at C-DMRs further than 1 kb from a TE insertion (non-TE-DMRs). This was evident from the distribution of correlations for non-TE-DMRs being centred around zero, whereas for TE-DMRs this distribution was skewed to the right (Figure 6C,  $D=0.23$ ). For TE deletions, we did not observe such a difference in the distributions of correlation coefficients between TE-DMRs and non-TE-DMRs, nor for CG-DMRs and their nearby TE insertions or deletions (Figure 6C,  $D=0.07-0.10$ ). Furthermore, DNA methylation levels were often higher in the presence of the nearby TE insertion, while this relationship was generally not observed for C-DMRs further than 1 kb from a TE variant, for TE deletions, or for CG-DMRs (Figure 6 - figure supplement 1).

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

We next examined levels of DNA methylation in regions flanking all TE variants regardless of the presence or absence of a population DMR call. While DNA methylation levels around pericentromeric TE insertions and deletions (<3 Mb from a centromere) seemed to be unaffected by the presence of a TE insertion (Figure 7A), TE insertions in the chromosome arms were associated with an increase in DNA methylation levels in all contexts (Figure 7A, B). In contrast, TE deletions in the chromosome arms did not affect patterns of DNA methylation, as the flanking methylation

level in all contexts seemed to remain high following deletion of the TE (Figure 7A, C). As the change in DNA methylation levels around TE variant sites appeared to be restricted to regions <200 bp from the insertion site, we correlated DNA methylation levels in 200 bp regions flanking TE variants with the presence/absence of TE the variants. DNA methylation levels were often positively correlated with the presence of a TE insertion when the insertion was distant from a centromere (Figure 7D). TE deletions were more variably correlated with local DNA methylation levels, but also showed a slight bias towards positive correlations for TE deletions distant from the centromeres. As methylome data was available for both leaf and bud tissue for 12 accessions, we repeated this analysis comparing between tissue types, but did not observe any difference in the patterns of methylation surrounding TE variant sites between the two tissues (Figure 7 - figure supplement 1).

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

## Genome-wide analysis of TE variant association with methylation levels highlights distant and local control

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

359 checking for TE insertions near these C-DMRs that were present in the same accessions as the  
360 *trans* associated TE, as these could lead to a false *trans* association. These were extremely rare,  
361 with only 4 such cases for TE insertions, and 38 cases for TE deletions, and so were unable to  
362 explain the high degree of *trans* associations found.

## 363 Discussion

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

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

389 Perhaps most importantly, we provide evidence that differential TE content between genomes of  
390 *Arabidopsis* accessions underlies a large fraction of the previously reported population C-DMRs, in  
391 agreement with recent similar findings [16]. Thus, the frequency of pure epialleles, independent of  
392 underlying genetic variation, may be even more rare than previously anticipated [36]. We did not  
393 find evidence of CG-DMR methylation level being altered by the presence of common TE variants,  
394 but rather rare TE variants may be more important in shaping patterns of DNA methylation at  
395 some CG-DMRs, though the reasons for this distinction remain unclear. The level of local DNA  
396 methylation changes associated with TE variants was also related to the distance from a TE variant  
397 to the centromere, with variants in the chromosome arms being more strongly correlated with DNA  
398 methylation levels. This seems to be due to a higher baseline level of DNA methylation at the  
399 pericentromeric regions, which prevent any further increase in DNA methylation level following  
400 insertion of a TE. Furthermore, we found an important distinction between TE insertions and

401 TE deletions in the effect that these variants have on nearby DNA methylation levels. While  
402 flanking DNA methylation levels appeared to increase following a TE insertion, the deletion of an  
403 ancestral TE was often not associated with a corresponding decrease in flanking DNA methylation  
404 levels (Figure 7). This indicates that high levels of DNA methylation, once established, may be  
405 maintained in the absence of the TE insertion that presumably triggered the original change in  
406 DNA methylation level. It is then possible that TE variants explain more of the variation in DNA  
407 methylation patterns than we find direct evidence for, if some C-DMRs were formed by the insertion  
408 of an ancestral TE that is now absent in all the accessions analysed here. These DMRs would then  
409 represent the epigenetic “scars” of past TE insertions.

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

## 421 Methods

### 422 TEPID development

#### 423 Mapping

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

#### 430 TE variant discovery

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

436 To identify TE insertions with respect to the reference genome, split read alignments are first  
437 filtered to remove reads where the distance between split mapping loci is less than 5 kb, to remove  
438 split reads due to small indels, or split reads with a mapping quality (MAPQ) less than 5. Split  
439 and discordant read mapping coordinates are then intersected using pybedtools [41, 42] with the

440 Col-0 reference TE annotation, requiring 80% overlap between TE and read mapping coordinates.  
441 To determine putative TE insertion sites, regions are then identified that contain independent  
442 discordant read pairs aligned in an orientation facing one another at the insertion site, with their  
443 mate pairs intersecting with the same TE (Figure 1). The total number of split and discordant reads  
444 intersecting the insertion site and the TE is then calculated, and a TE insertion predicted where  
445 the combined number of reads is greater than a threshold determined by the average sequencing  
446 depth over the whole genome (1/10 coverage if coverage is greater than 10, otherwise a minimum  
447 of 2 reads). Alternatively, in the absence of discordant reads mapped in orientations facing one  
448 another, the required total number of split and discordant reads at the insertion site linked to the  
449 inserted TE is set higher, requiring twice as many reads.

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

#### 469 *TE variant refinement*

470 Once TE insertions are identified using the ‘tepid-map’ and ‘tepid-discover’ algorithms, these  
471 variants can be refined if multiple related samples are analysed. The ‘tepid-refine’ algorithm is  
472 designed to interrogate regions of the genome in which a TE insertion was discovered in other  
473 samples but not the sample in question, and check for evidence of that TE insertion in the sample  
474 using lower read count thresholds compared to the ‘tepid-discover’ step. In this way, the refine  
475 step leverages TE variant information for a group of related samples to reduce false negative calls  
476 within the group. This distinguishes TEPID from other similar methods for TE variant discovery  
477 utilizing short sequencing reads. A file containing the coordinates of each insertion, and a list of  
478 sample names containing the TE insertion must be provided to the ‘tepid-refine’ algorithm, which  
479 this can be generated using the ‘merge\_insertions.py’ script included in the TEPID package. Each  
480 sample is examined in regions where there was a TE insertion identified in another sample in the  
481 group. If there is a sequencing breakpoint within this region (no continuous read coverage spanning  
482 the region), split reads mapped to this region will be extracted from the alignment file and their  
483 coordinates intersected with the TE reference annotation. If there are split reads present at the  
484 variant site that are linked to the same TE as was identified as an insertion at that location, this

485 TE insertion is recorded in a new file as being present in the sample in question. If there is no  
486 sequencing coverage in the queried region for a sample, an “NA” call is made indicating that it is  
487 unknown whether the particular sample contains the TE insertion or not.

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

## 492 TE variant simulation

493 To test the sensitivity and specificity of TEPID, 100 TE insertions (50 copy-paste transpositions, 50  
494 cut-paste transpositions) and 100 TE absence variants were simulated in the Arabidopsis genome  
495 using the RSVSim R package, version 1.7.2 [44], and synthetic reads generated from the modified  
496 genome at various levels of sequencing coverage using wgsim [40] (<https://github.com/lh3/wgsim>).  
497 These reads were then used to calculate the true positive, false positive, and false negative TE  
498 variant discovery rates for TEPID at various sequencing depths, by running ‘tepid-map’ and  
499 ‘tepid-discover’ using the simulated reads with the default parameters (Figure 1 - figure supplement  
500 1).

## 501 Estimation of sensitivity

502 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  
503 with the TEPID package to identify TE variants. Reads providing evidence for TE variants  
504 were then mapped to the *de novo* assembled *Ler* genome [21]. To determine whether reads  
505 mapped to homologous regions of the *Ler* and Col-0 reference genome, the *de novo* assembled *Ler*  
506 genome sequence between mate pair mapping locations in *Ler* were extracted, with repeats masked  
507 using RepeatMasker with RepBase-derived libraries and the default parameters (version 4.0.5,  
508 <http://www.repeatmasker.org>). A blastn search was then conducted against the Col-0 genome  
509 using the following parameters: ‘-max-target-seqs 1’, ‘-evalue 1e-6’ [23]. Coordinates of the top  
510 blast hit for each read location were then compared with the TE variant sites identified using those  
511 reads. To estimate false negative rates for TEPID TE absence calls, *Ler* TE absence calls were  
512 compared with a known set of Col-0-specific TE insertions, absent in *Ler* [16]. For TEPID TE  
513 insertion calls, we mapped Col-0 DNA sequencing reads [24] to the *Ler* PacBio assembly, and  
514 identified sites with read evidence reaching the TEPID threshold for a TE insertion call to be made.

## 516 Arabidopsis TE variant discovery

517 We ran the TEPID, including the insertion refinement step, on previously published sequencing  
518 data for 216 different Arabidopsis populations (NCBI SRA SRA012474; Schmitz et al. 2013),  
519 mapping to the TAIR10 reference genome and using the TAIR9 TE annotation. The ‘-mask’ option  
520 was also used to mask the mitochondrial and plastid genomes. We also ran TEPID using previously  
521 published transgenerational data for salt stress and control conditions (NCBI SRA SRP045804;

522 [24]), using the ‘–mask’ option to mask mitochondrial and plastid genomes, and the ‘–strict’ option  
523 for highly related samples.

## 524 TE variant / SNP comparison

525 SNP information for 216 *Arabidopsis* accessions was obtained from the 1001 genomes data center  
526 ([http://1001genomes.org/data/Salk/releases/2013\\_24\\_01/](http://1001genomes.org/data/Salk/releases/2013_24_01/); [20]). This was formatted into reference  
527 (Col-0 state), alternate, or NA calls for each SNP. Accessions with both TE variant information  
528 and SNP data were selected for analysis. Hierarchical clustering of accessions by SNPs as well as  
529 TE variants were used to identify essentially clonal accessions, as these would skew minor allele  
530 frequency calculations. A single representative from each cluster of similar accessions was kept,  
531 leading to a total of 187 accessions for comparison. For each TE variant with minor allele frequency  
532 greater than 3%, the nearest 300 upstream and 300 downstream SNPs with a minor allele frequency  
533 greater than 3% were selected. Pairwise genotype correlations ( $r^2$  values) for all complete cases  
534 were obtained for SNP-SNP and SNP-TE variant states.  $r^2$  values were then ordered by decreasing  
535 rank and a median SNP-SNP rank value was calculated. For each of the 600 ranked surrounding  
536 positions, the number of times the TE rank was greater than the SNP-SNP median rank was  
537 calculated as a relative LD metric of TE to SNP. TE variants with less than 200 ranks over the  
538 SNP-SNP median were classified as low-LD insertions. TE variants with ranks between 200 and  
539 400 were classified as mid-LD, while TE variants with greater than 400 ranks above their respective  
540 SNP-SNP median value were classified as variants in high LD with flanking SNPs.

## 541 PCR validations

### 542 Selection of accessions to be genotyped

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

### 548 Selection of TE variants for validation and primer design

549 Ten TE insertion sites and 10 TE absence sites were randomly selected for validation by PCR  
550 amplification. Only insertions and absence variants that were variable in at least two of the  
551 fourteen accessions selected to be genotyped were considered. For insertion sites, primers were  
552 designed to span the predicted TE insertion site. For TE absence sites, two primer sets were  
553 designed; one primer set to span the TE, and another primer set with one primer annealing  
554 within the TE sequence predicted to be absent, and the other primer annealing in the flanking  
555 sequence (Figure 2 - figure supplement 3). Primer sequences were designed that did not anneal  
556 to regions of the genome containing previously identified SNPs in any of the 216 accessions [20]  
557 or small insertions and deletions, identified using lumpy-sv with the default settings [46] (<https://github.com/arq5x/lumpy-sv>), had an annealing temperature close to 52°C calculated based on  
558 nearest neighbor thermodynamics (MeltingTemp submodule in the SeqUtils python module; [47]),  
559 GC content between 40% and 60%, and contained the same base repeated not more than four  
560

561 times in a row. Primers were aligned to the TAIR10 reference genome using bowtie2 [37] with the  
562 ‘-a’ flag set to report all alignments, and those with more than 5 mapping locations in the genome  
563 were then removed.

#### 564 PCR

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

#### 568 mRNA analysis

569 Processed mRNA data for 144 wild Arabidopsis accessions were downloaded from NCBI GEO  
570 GSE43858 [20]. To find differential gene expression dependent on TE presence/absence variation,  
571 we first filtered TE variants to include only those where the TE variant was shared by at least 5  
572 accessions with RNA data available. We then grouped accessions based on TE presence/absence  
573 variants, and performed a Mann-Whitney U test to determine differences in RNA transcript  
574 abundance levels between the groups. We used q-value estimation to correct for multiple testing,  
575 using the R qvalue package v2.2.2 with the following parameters: lambda = seq(0, 0.6, 0.05),  
576 smooth.df = 4 [48]. Genes were defined as differentially expressed where there was a greater than  
577 2 fold difference in expression between the groups, with a q-value less than 0.01. Gene ontology  
578 enrichment analysis was performed using PANTHER (<http://pantherdb.org>).

#### 579 DNA methylation data analysis

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

#### 582 Rare variant analysis

583 To assess the effect of rare TE variants on gene expression or DMR DNA methylation levels, we  
584 tested for a burden of rare variants in the population extremes, essentially as described previously  
585 [28]. For each rare TE variant near a gene or DMR, we ranked the gene expression level or DMR  
586 DNA methylation level for all accessions in the population, and tallied the ranks of accessions  
587 containing a rare variant. These rank counts were then binned to produce a histogram of the  
588 distribution of ranks. We then fit a quadratic model to the counts data, and calculated the R<sup>2</sup> and  
589 p-value for the fit of the model.

#### 590 TE variant and DMR genome-wide association analysis

591 Accessions were subset to those with both leaf DNA methylation data and TEPIID calls. Pairwise  
592 correlations were performed for observed data pairs for each TE variant and a filtered set of  
593 population C-DMRs, with those C-DMRs removed where more than 15% of the accessions had no  
594 coverage. This amounted to a final set of 9,777 C-DMRs. Accession names were then permuted

595 to produce a randomized dataset, and pairwise correlations again calculated. This was repeated  
596 500 times to produce a distribution of expected Pearson correlation coefficients for each pairwise  
597 comparison. Correlation values more extreme than any of the 500 permutations were deemed  
598 significant.

## 599 Data access

600 TEPID source code can be accessed at <https://github.com/ListerLab/TEPID>. Code and  
601 data needed to reproduce this analysis can be found at <https://github.com/timoast/Arabidopsis-TE-variants>. *Ler* TE variants are available in Figure 1 - source data 1 and  
602 2. TE variants identified among the 216 wild Arabidopsis accessions resequenced by Schmitz et al.  
603 (2013) are available in Figure 2 - source data 1, 2 and 3.

## 605 Acknowledgments

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

## 614 Author contributions

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

## 619 Competing financial interests

620 The authors declare no competing financial interests.

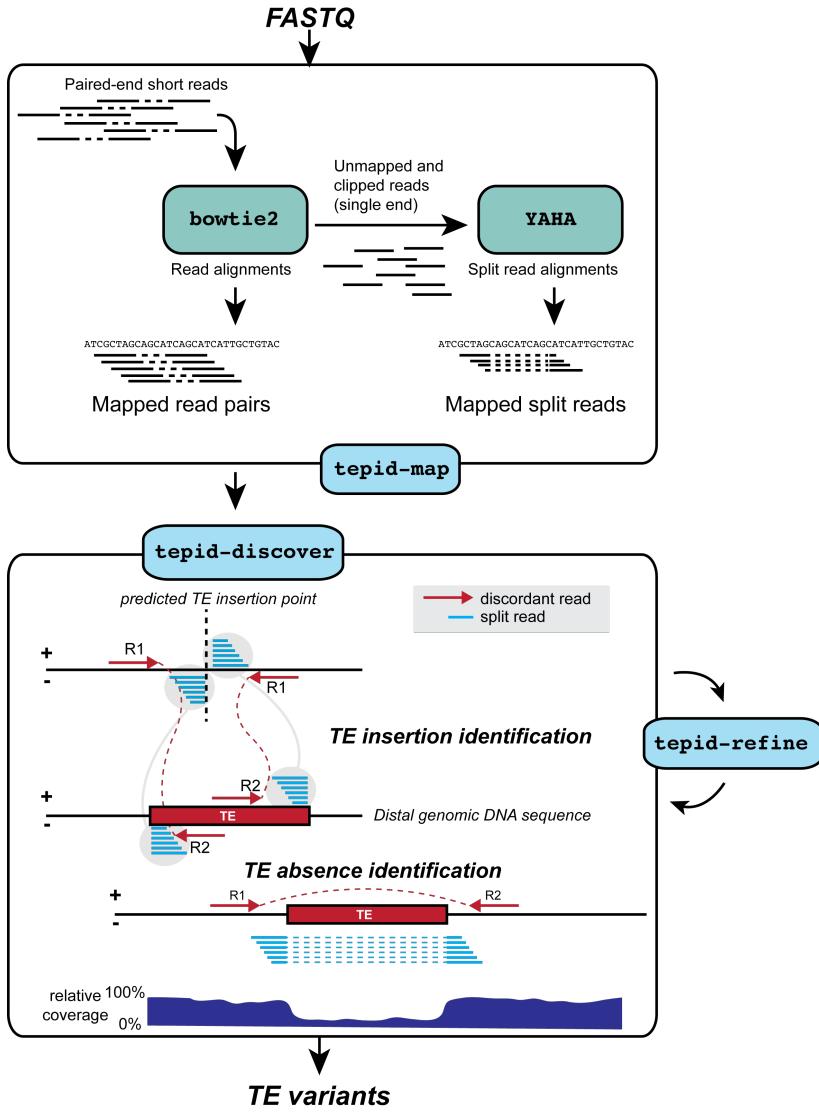
## 621 References

- 622 [1] Thomas Wicker et al. “A unified classification system for eukaryotic transposable elements.”  
623 In: *Nature Reviews Genetics* 8.12 (Dec. 2007), pp. 973–982. DOI: [10.1038/nrg2165](https://doi.org/10.1038/nrg2165).
- 624 [2] Assaf Zemach et al. “The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyl-  
625 transferases to Access H1-Containing Heterochromatin”. In: *Cell* 153.1 (Mar. 2013), pp. 193–  
626 205. DOI: [10.1016/j.cell.2013.02.033](https://doi.org/10.1016/j.cell.2013.02.033).
- 627 [3] Marjori A Matzke and Rebecca A Mosher. “RNA-directed DNA methylation: an epigenetic  
628 pathway of increasing complexity”. In: *Nature Reviews Genetics* 15.6 (May 2014), pp. 394–408.  
629 DOI: [10.1038/nrg3683](https://doi.org/10.1038/nrg3683).
- 630 [4] Marie Mirouze et al. “Selective epigenetic control of retrotransposition in Arabidopsis.” In:  
631 *Nature* 461.7262 (Sept. 2009), pp. 427–430. DOI: [10.1038/nature08328](https://doi.org/10.1038/nature08328).
- 632 [5] Asuka Miura et al. “Mobilization of transposons by a mutation abolishing full DNA methyla-  
633 tion in Arabidopsis”. In: *Nature* 411.6834 (2001), pp. 212–214. DOI: [10.1038/35075612](https://doi.org/10.1038/35075612).
- 634 [6] Hidetoshi Saze, Ortrun Mittelsten Scheid, and Jerzy Paszkowski. “Maintenance of CpG  
635 methylation is essential for epigenetic inheritance during plant gametogenesis”. In: *Nature  
636 Genetics* 34.1 (Mar. 2003), pp. 65–69. DOI: [10.1038/ng1138](https://doi.org/10.1038/ng1138).
- 637 [7] Zachary Lippman et al. “Role of transposable elements in heterochromatin and epigenetic  
638 control.” In: *Nature* 430.6998 (July 2004), pp. 471–476. DOI: [10.1038/nature02651](https://doi.org/10.1038/nature02651).
- 639 [8] Jeffrey A Jeddelloh, Trevor L Stokes, and Eric J Richards. “Maintenance of genomic methyla-  
640 tion requires a SWI2/SNF2-like protein”. In: *Nature Genetics* 22.1 (1999), pp. 94–97. DOI:  
641 [10.1038/8803](https://doi.org/10.1038/8803).
- 642 [9] Clémentine Vitte et al. “The bright side of transposons in crop evolution.” In: *Briefings in  
643 Functional Genomics* 13.4 (July 2014), pp. 276–295. DOI: [10.1093/bfgp/elu002](https://doi.org/10.1093/bfgp/elu002).
- 644 [10] Elizabeth Hénaff et al. “Extensive amplification of the E2F transcription factor binding sites  
645 by transposons during evolution of Brassica species.” In: *The Plant Journal* 77.6 (Mar. 2014),  
646 pp. 852–862. DOI: [10.1111/tpj.12434](https://doi.org/10.1111/tpj.12434).
- 647 [11] Anthony Bolger et al. “The genome of the stress-tolerant wild tomato species”. In: *Nature  
648 Genetics* 46.9 (July 2014), pp. 1034–1038. DOI: [10.1038/ng.3046](https://doi.org/10.1038/ng.3046).
- 649 [12] Hidetaka Ito et al. “An siRNA pathway prevents transgenerational retrotransposition in plants  
650 subjected to stress”. In: *Nature* 472.7341 (Mar. 2011), pp. 115–119. DOI: [10.1038/nature09861](https://doi.org/10.1038/nature09861).
- 651 [13] Irina Makarevitch et al. “Transposable Elements Contribute to Activation of Maize Genes in  
652 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).

- 654 [14] Paul Bundock and Paul Hooykaas. “An Arabidopsis hAT-like transposase is essential for  
655 plant development.” In: *Nature* 436.7048 (July 2005), pp. 282–284. DOI: [10.1038/nature03667](https://doi.org/10.1038/nature03667).
- 656 [15] Jun Cao et al. “Whole-genome sequencing of multiple *Arabidopsis thaliana* populations.” In:  
657 *Nature Publishing Group* 43.10 (Oct. 2011), pp. 956–963. DOI: [10.1038/ng.911](https://doi.org/10.1038/ng.911).
- 658 [16] Leandro Quadrana et al. “The *Arabidopsis thaliana* mobilome and its impact at the species  
659 level.” In: *eLife* 5 (2016), p. 6919. DOI: [10.7554/eLife.15716](https://doi.org/10.7554/eLife.15716).
- 660 [17] Djie Tjwan Thung et al. “Mobster: accurate detection of mobile element insertions in next  
661 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).
- 662 [18] Sofia M. C. Robb et al. “The use of RelocaTE and unassembled short reads to produce  
663 high-resolution snapshots of transposable element generated diversity in rice”. In: *G3: Genes  
664 / Genomes / Genetics* (2013). DOI: [10.1534/g3.112.005348/-/DC1](https://doi.org/10.1534/g3.112.005348/-/DC1).
- 665 [19] Elizabeth Hénaff et al. “Jitterbug: somatic and germline transposon insertion detection at  
666 single-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).
- 668 [20] Robert J Schmitz et al. “Patterns of population epigenomic diversity”. In: *Nature* 495.7440  
669 (Mar. 2013), pp. 193–198. DOI: [10.1038/nature11968](https://doi.org/10.1038/nature11968).
- 670 [21] Chen-Shan Chin et al. “Nonhybrid, finished microbial genome assemblies from long-read SMRT  
671 sequencing data”. In: *Nature Methods* 10.6 (May 2013), pp. 563–569. DOI: [10.1038/nmeth.2474](https://doi.org/10.1038/nmeth.2474).
- 672 [22] Korbinian Schneeberger et al. “Reference-guided assembly of four diverse *Arabidopsis thaliana*  
673 genomes”. In: *Proceedings of the National Academy of Sciences of the United States of America*  
674 108.25 (2011), pp. 10249–10254. DOI: [10.1073/pnas.1107739108](https://doi.org/10.1073/pnas.1107739108).
- 675 [23] Christiam Camacho et al. “BLAST+: architecture and applications.” In: *BMC Bioinformatics*  
676 10.1 (2009), p. 421. DOI: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421).
- 677 [24] Caifu Jiang et al. “Environmentally responsive genome-wide accumulation of de novo Ara-  
678 bidopsis thaliana mutations and epimutations.” In: *Genome Research* 24.11 (Nov. 2014),  
679 pp. 1821–1829. DOI: [10.1101/gr.177659.114](https://doi.org/10.1101/gr.177659.114).
- 680 [25] Alessandra M Sullivan et al. “Mapping and dynamics of regulatory DNA and transcription  
681 factor 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).
- 683 [26] Stefan Grob, Marc W Schmid, and Ueli Grossniklaus. “Hi-C Analysis in *Arabidopsis* Identifies  
684 the KNOT, a Structure with Similarities to the flamenco Locus of *Drosophila*”. In: *Molecular  
685 Cell* (Aug. 2014), pp. 1–16. DOI: [10.1016/j.molcel.2014.07.009](https://doi.org/10.1016/j.molcel.2014.07.009).

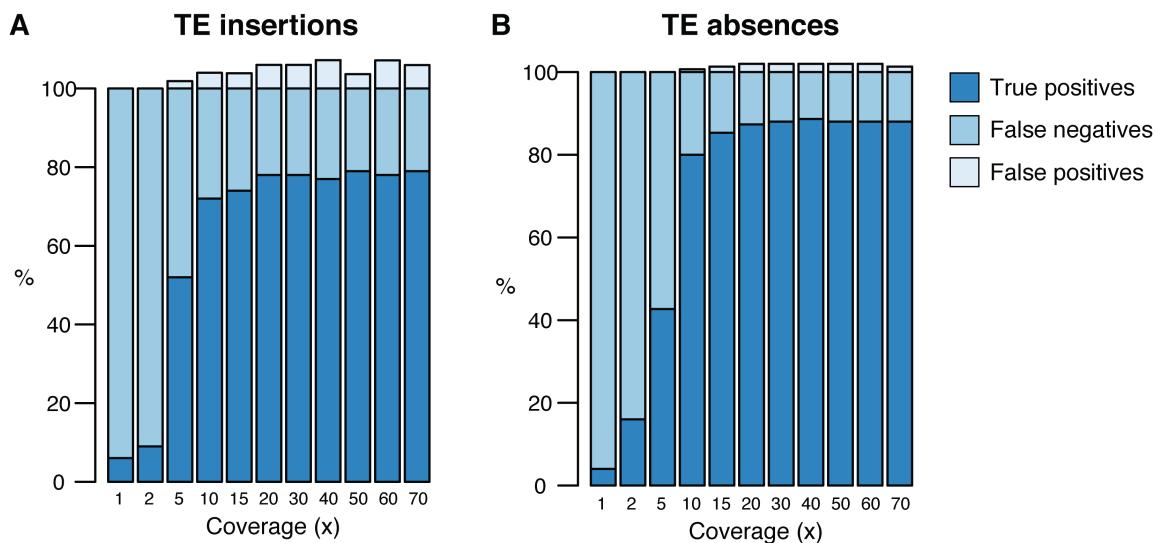
- 686 [27] Matthew W Horton et al. “Genome-wide patterns of genetic variation in worldwide Arabidopsis  
687 thaliana accessions from the RegMap panel”. In: *Nature Publishing Group* 44.2 (Feb. 2012),  
688 pp. 212–216. DOI: [10.1038/ng.1042](https://doi.org/10.1038/ng.1042).
- 689 [28] Jing Zhao et al. “A Burden of Rare Variants Associated with Extremes of Gene Expression in  
690 Human Peripheral Blood”. In: *The American Journal of Human Genetics* 98.2 (Feb. 2016),  
691 pp. 299–309. DOI: [10.1016/j.ajhg.2015.12.023](https://doi.org/10.1016/j.ajhg.2015.12.023).
- 692 [29] Guodong Wang et al. “A genome-wide functional investigation into the roles of receptor-  
693 like proteins in Arabidopsis.” In: *Plant Physiolosy* 147.2 (June 2008), pp. 503–517. DOI:  
694 [10.1104/pp.108.119487](https://doi.org/10.1104/pp.108.119487).
- 695 [30] María José Aranzana et al. “Genome-Wide Association Mapping in Arabidopsis Identifies  
696 Previously Known Flowering Time and Pathogen Resistance Genes”. In: *PLoS Genetics* 1.5  
697 (2005), e60–9. DOI: [10.1371/journal.pgen.0010060](https://doi.org/10.1371/journal.pgen.0010060).
- 698 [31] Akira Katoh et al. “Early steps in the biosynthesis of NAD in Arabidopsis start with  
699 aspartate and occur in the plastid.” In: *Plant Physiolosy* 141.3 (July 2006), pp. 851–857. DOI:  
700 [10.1104/pp.106.081091](https://doi.org/10.1104/pp.106.081091).
- 701 [32] Ryan Lister et al. “Highly integrated single-base resolution maps of the epigenome in Ara-  
702 bidopsis.” 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).
- 703 [33] Shawn J Cokus et al. “Shotgun bisulphite sequencing of the Arabidopsis genome reveals  
704 DNA methylation patterning”. In: *Nature* 452.7184 (Feb. 2008), pp. 215–219. DOI: [10.1038/nature06745](https://doi.org/10.1038/nature06745).
- 705 [34] Xiaoyu Zhang et al. “Genome-wide High-Resolution Mapping and Functional Analysis  
706 of DNA Methylation in Arabidopsis”. In: *Cell* 126.6 (Sept. 2006), pp. 1189–1201. DOI:  
707 [10.1016/j.cell.2006.08.003](https://doi.org/10.1016/j.cell.2006.08.003).
- 709 [35] Daniel Zilberman et al. “Genome-wide analysis of Arabidopsis thaliana DNA methylation  
710 uncovers an interdependence between methylation and transcription.” In: *Nature Genetics*  
711 39.1 (Jan. 2007), pp. 61–69. DOI: [10.1038/ng1929](https://doi.org/10.1038/ng1929).
- 712 [36] Eric J Richards. “Inherited epigenetic variation–revisiting soft inheritance.” In: *Nature Reviews  
713 Genetics* 7.5 (May 2006), pp. 395–401. DOI: [10.1038/nrg1834](https://doi.org/10.1038/nrg1834).
- 714 [37] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In:  
715 *Nature Methods* 9.4 (Mar. 2012), pp. 357–359. DOI: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923).
- 716 [38] Gregory G Faust and Ira M Hall. “SAMBLASTER: fast duplicate marking and structural  
717 variant read extraction.” In: *Bioinformatics* 30.17 (Sept. 2014), pp. 2503–2505. DOI: [10.1093/bioinformatics/btu314](https://doi.org/10.1093/bioinformatics/btu314).

- 719 [39] Gregory G Faust and Ira M Hall. “YAHA: fast and flexible long-read alignment with optimal  
720 breakpoint detection.” In: *Bioinformatics* 28.19 (Oct. 2012), pp. 2417–2424. DOI: [10.1093/bioinformatics/bts456](https://doi.org/10.1093/bioinformatics/bts456).
- 722 [40] Heng Li et al. “The Sequence Alignment/Map format and SAMtools.” In: *Bioinformatics*  
723 25.16 (Aug. 2009), pp. 2078–2079. DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
- 724 [41] Ryan K Dale, Brent S Pedersen, and Aaron R Quinlan. “Pybedtools: a flexible Python library  
725 for manipulating genomic datasets and annotations.” In: *Bioinformatics* 27.24 (Dec. 2011),  
726 pp. 3423–3424. DOI: [10.1093/bioinformatics/btr539](https://doi.org/10.1093/bioinformatics/btr539).
- 727 [42] Aaron R Quinlan and Ira M Hall. “BEDTools: a flexible suite of utilities for comparing genomic  
728 features.” In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. DOI: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).
- 730 [43] Ronald H Plasterk. “The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*.”  
731 In: *The EMBO Journal* 10.7 (July 1991), pp. 1919–1925.
- 732 [44] Christoph Bartenhagen and Martin Dugas. “RSVSim: an R/Bioconductor package for the  
733 simulation of structural variations.” In: *Bioinformatics* 29.13 (July 2013), pp. 1679–1681. DOI:  
734 [10.1093/bioinformatics/btt198](https://doi.org/10.1093/bioinformatics/btt198).
- 735 [45] K Edwards, C Johnstone, and C Thompson. “A simple and rapid method for the preparation  
736 of plant genomic DNA for PCR analysis.” In: *Nucleic Acids Research* 19.6 (Mar. 1991),  
737 p. 1349.
- 738 [46] Ryan M Layer et al. “LUMPY: a probabilistic framework for structural variant discovery.”  
739 In: *Genome Biology* 15.6 (2014), R84. DOI: [10.1186/gb-2014-15-6-r84](https://doi.org/10.1186/gb-2014-15-6-r84).
- 740 [47] Peter J A Cock et al. “Biopython: freely available Python tools for computational molecular  
741 biology and bioinformatics.” In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. DOI:  
742 [10.1093/bioinformatics/btp163](https://doi.org/10.1093/bioinformatics/btp163).
- 743 [48] John D Storey and Robert Tibshirani. “Statistical significance for genomewide studies.” In:  
744 *Proceedings of the National Academy of Sciences of the United States of America* 100.16 (Aug.  
745 2003), pp. 9440–9445. DOI: [10.1073/pnas.1530509100](https://doi.org/10.1073/pnas.1530509100).



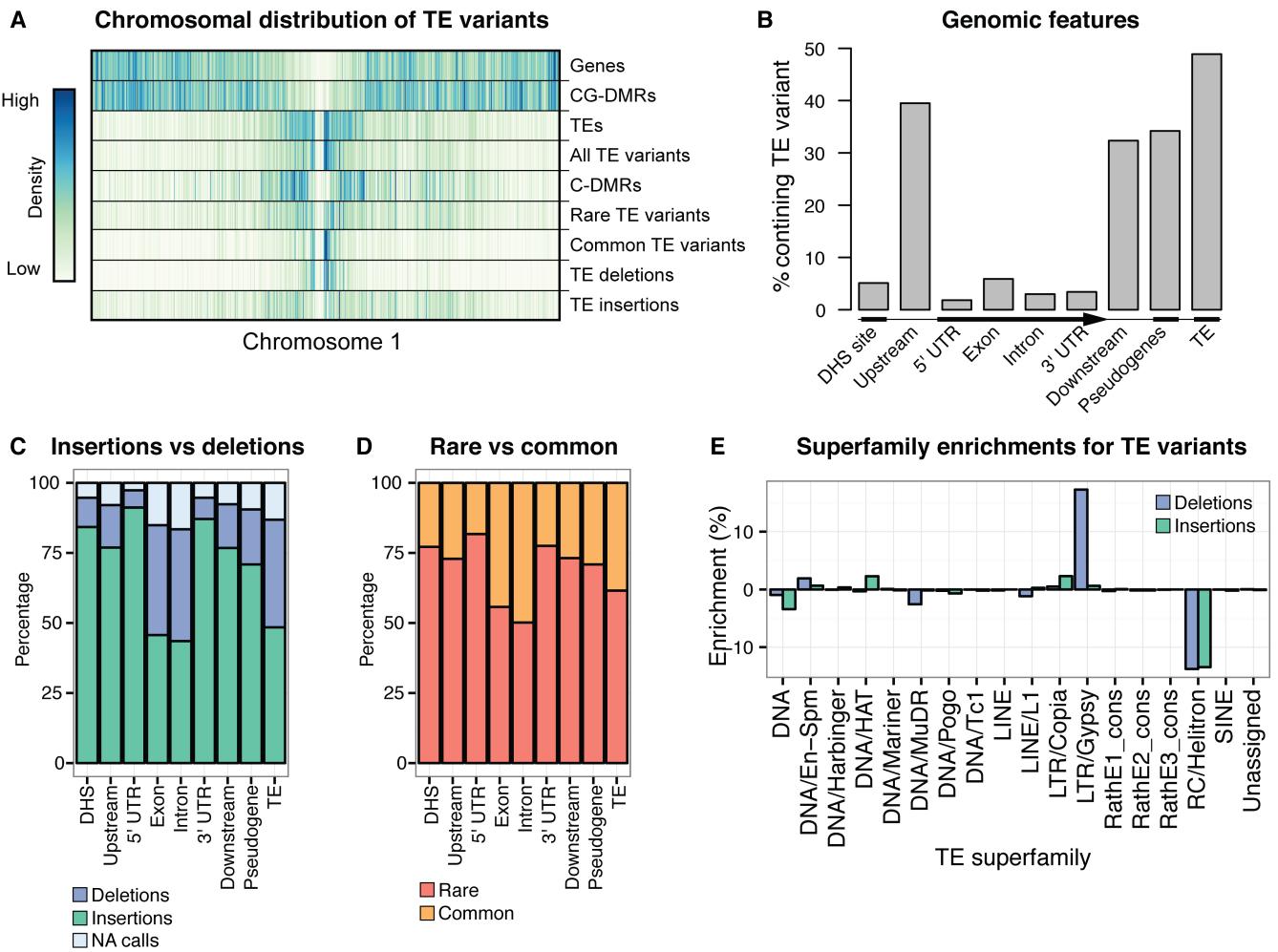
**Figure 1:** TE variant discovery pipeline

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



**Figure 1:** figure supplement 1

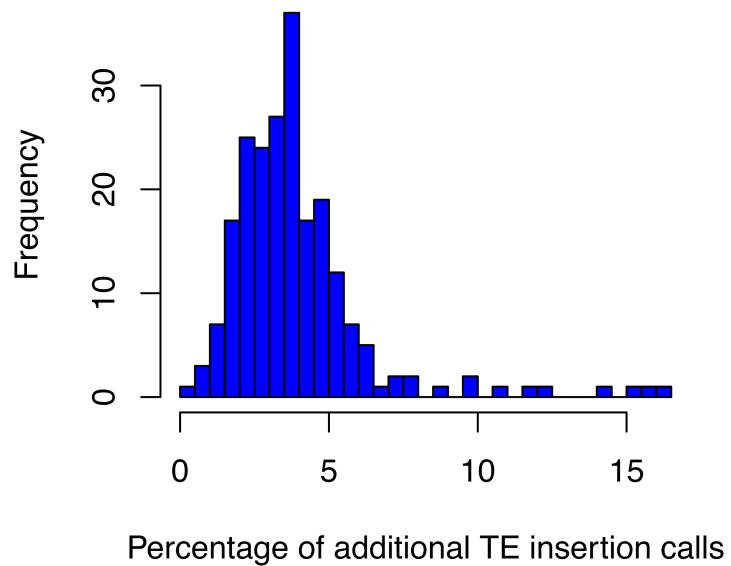
755 Testing of the TEPID pipeline using simulated TE variants in the Arabidopsis Col-0 genome  
 756 (TAIR10), for a range of sequencing coverage levels. TE insertions (A) and TE absence calls (B).



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

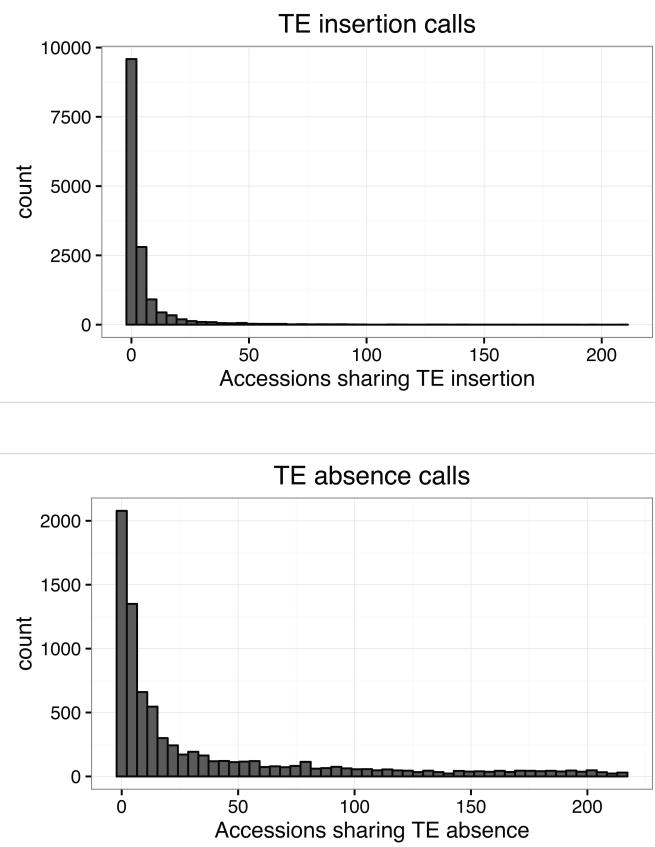
- 757 (A) Distribution of identified TE variants on chromosome 1, with distributions of all Col-0 genes,  
758 Col-0 TEs, and population DMRs.
- 759 (B) Frequency of TE variants at different genomic features.
- 760 (C) Proportion of TE variants within each genomic feature classified as deletions or insertions.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

## TE calls due to TEPID refinement step



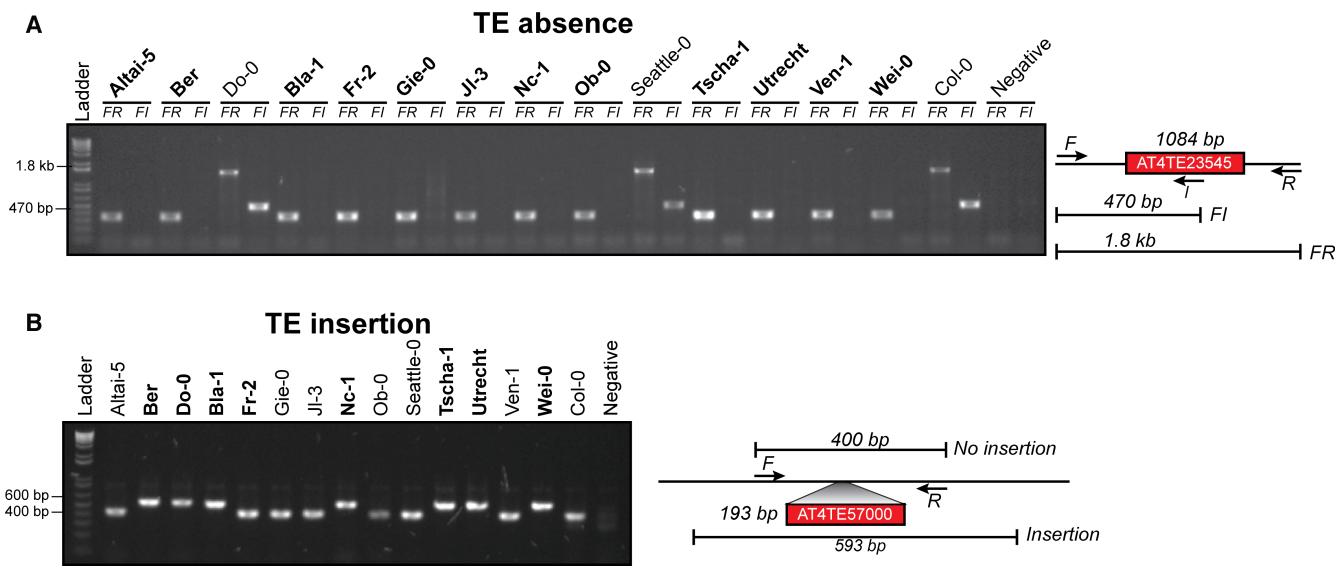
**Figure 2:** figure supplement 1

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



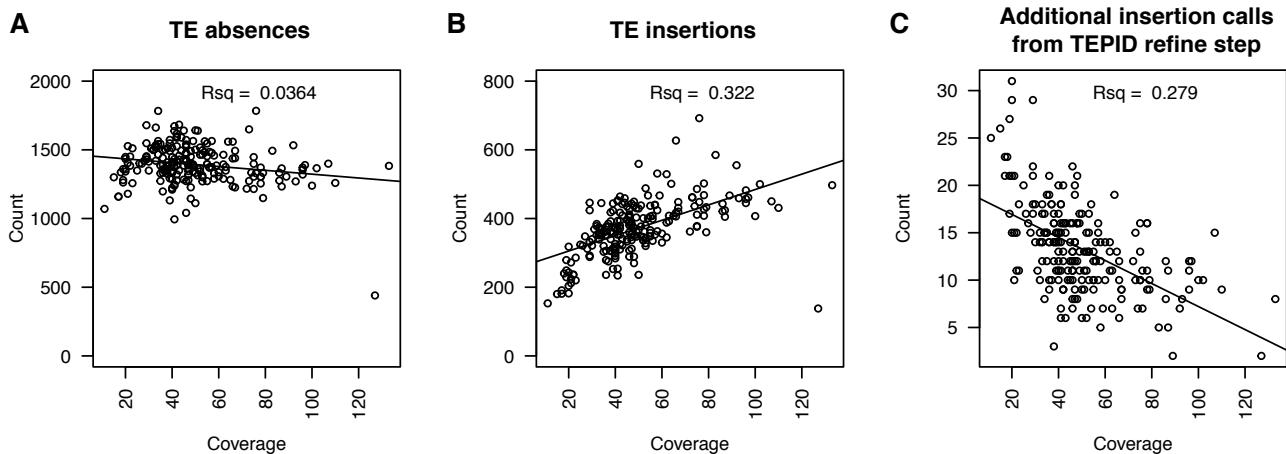
**Figure 2:** figure supplement 2

766 Minor allele frequency distribution for TE variants.



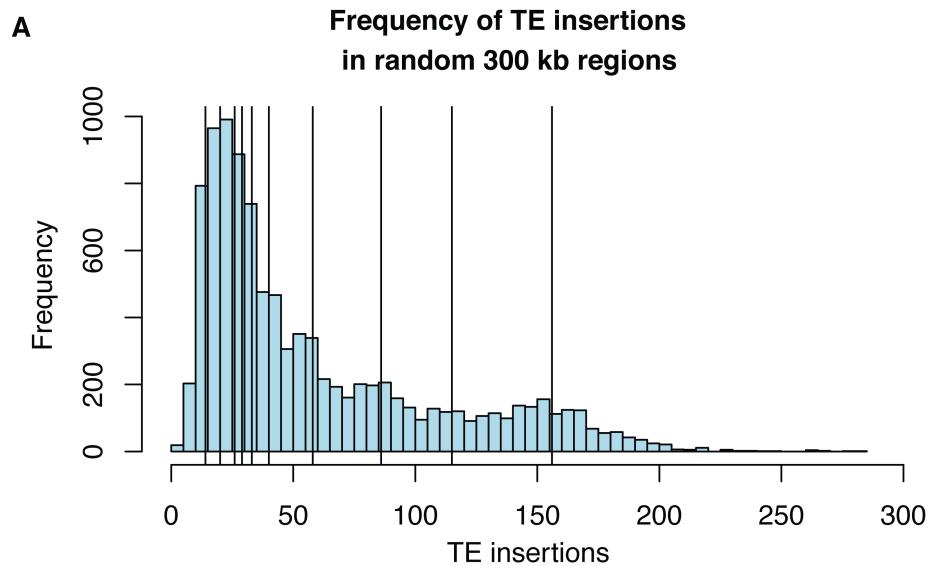
**Figure 2:** figure supplement 3

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



**Figure 2:** figure supplement 4

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

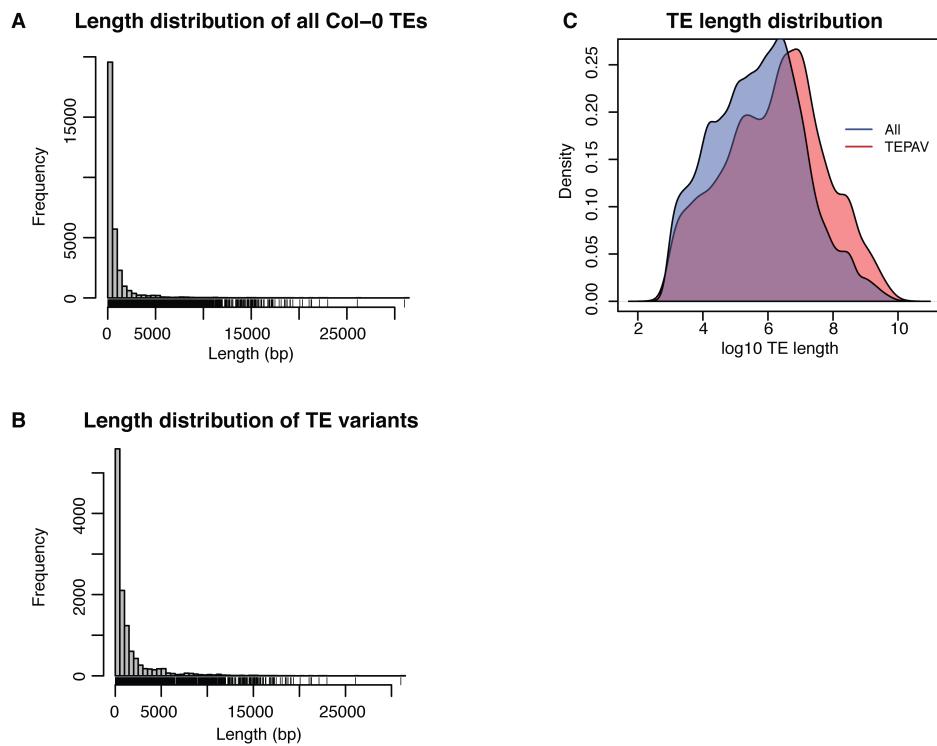


**B**

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

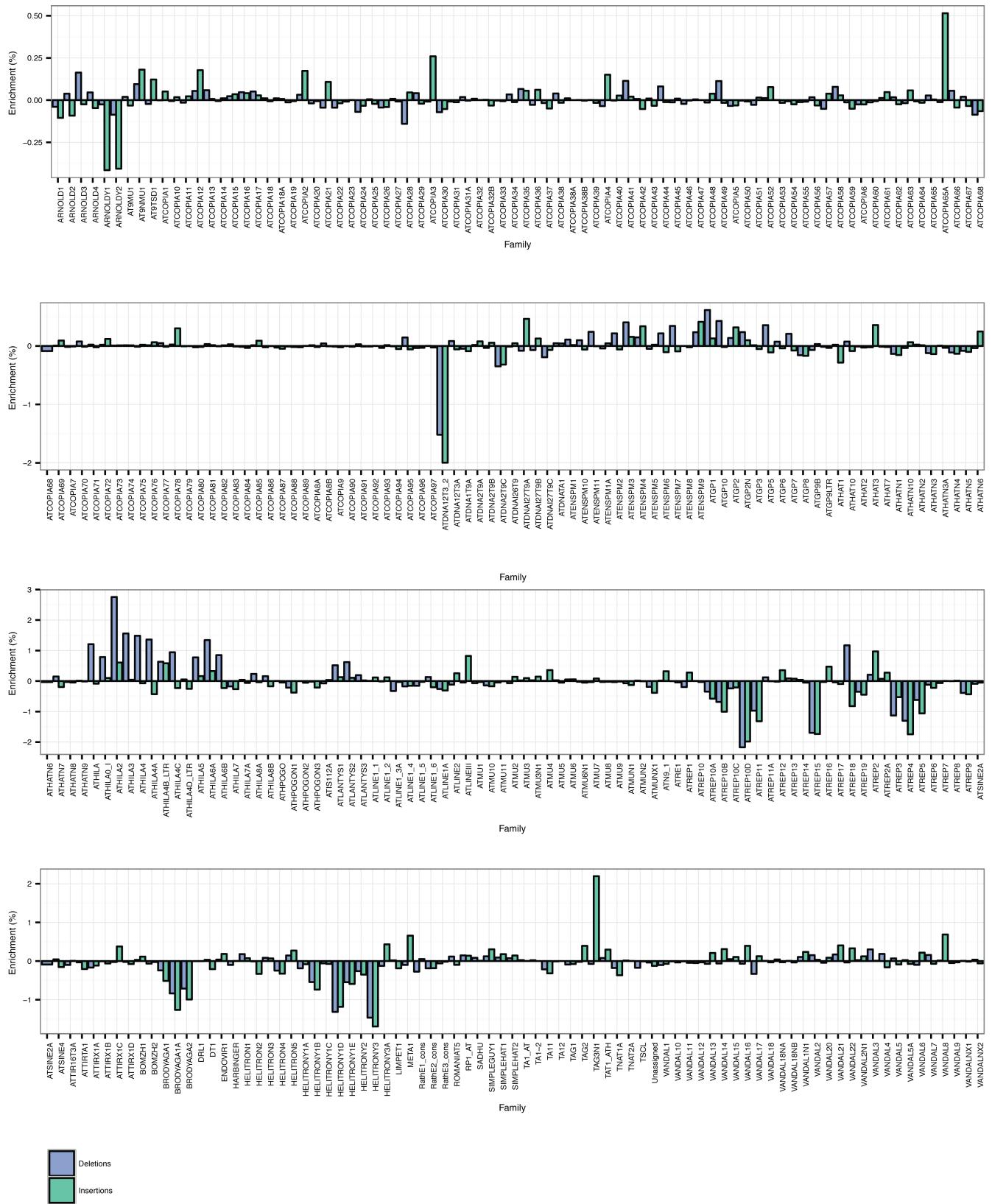
**Figure 2:** figure supplement 5. Frequency of TE insertion in the *KNOT* region

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



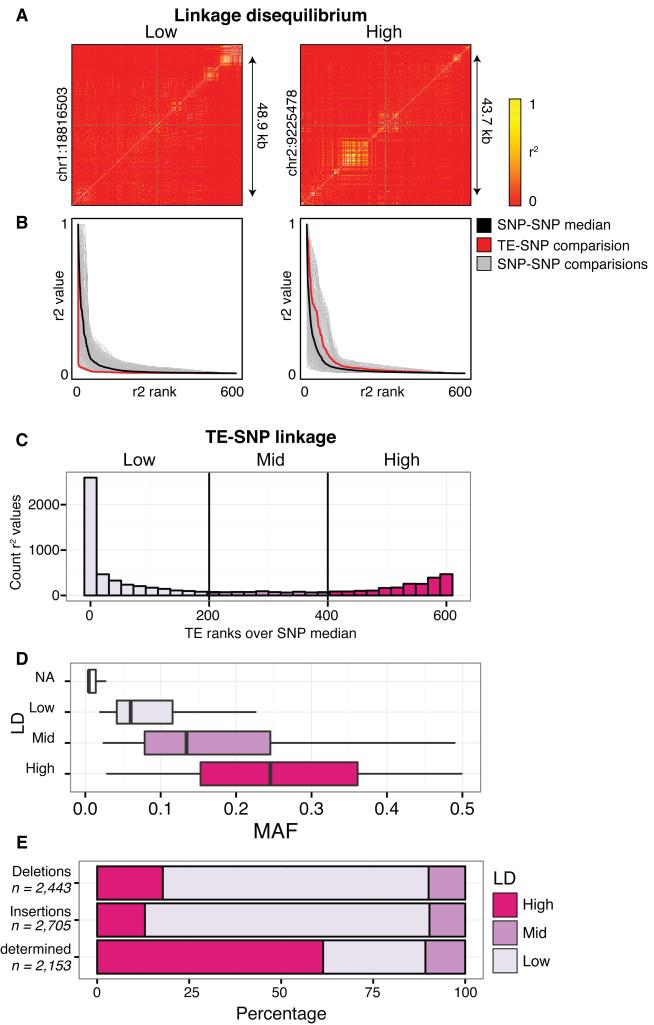
**Figure 2:** figure supplement 6. Length distribution for all Col-0 TEs and all TE variants

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



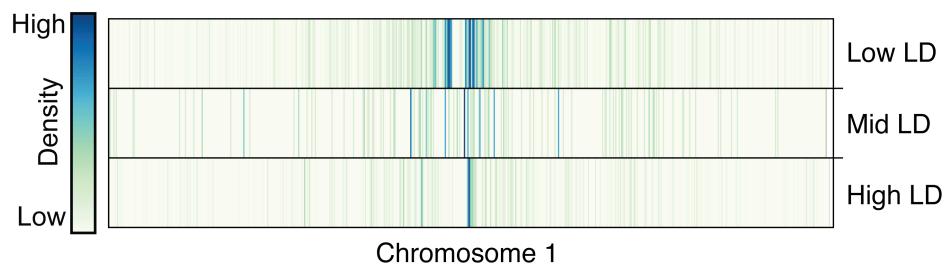
**Figure 2:** figure supplement 7

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



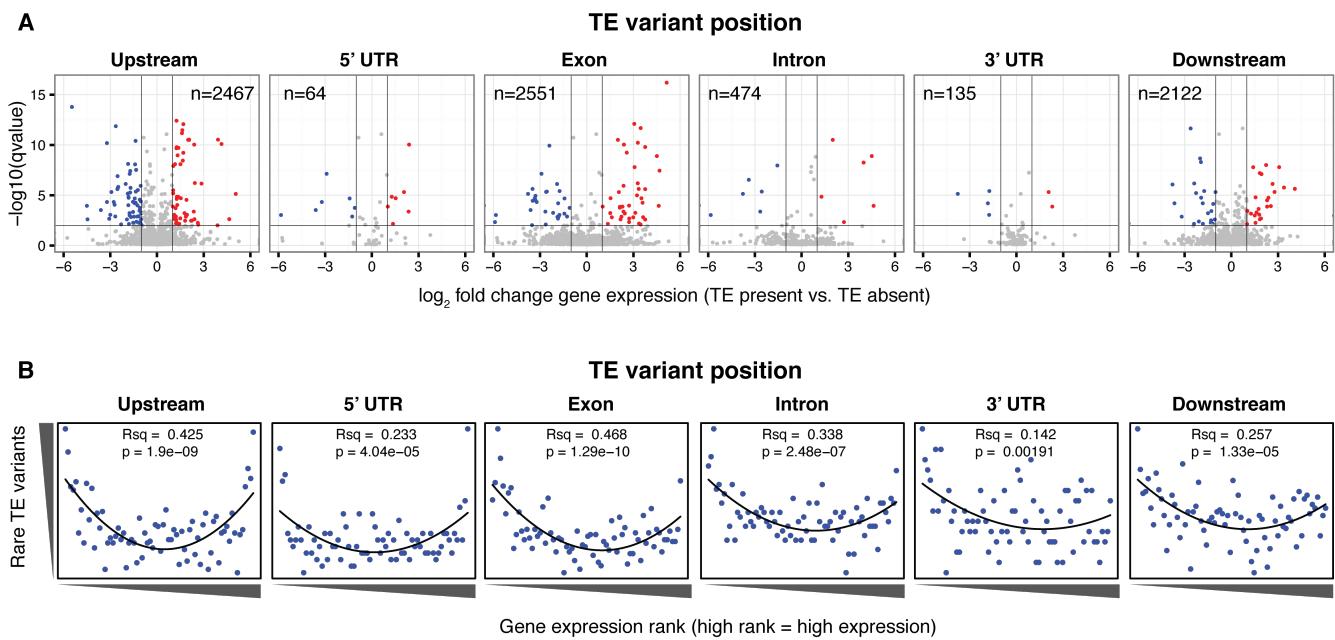
**Figure 3:** Patterns of TE-SNP linkage

- (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 undetermined TE variants in each LD category.



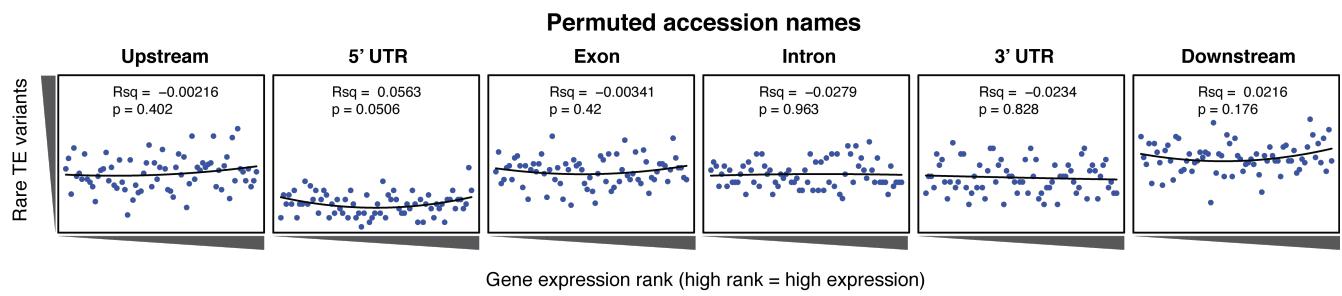
**Figure 3:** figure supplement 1

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



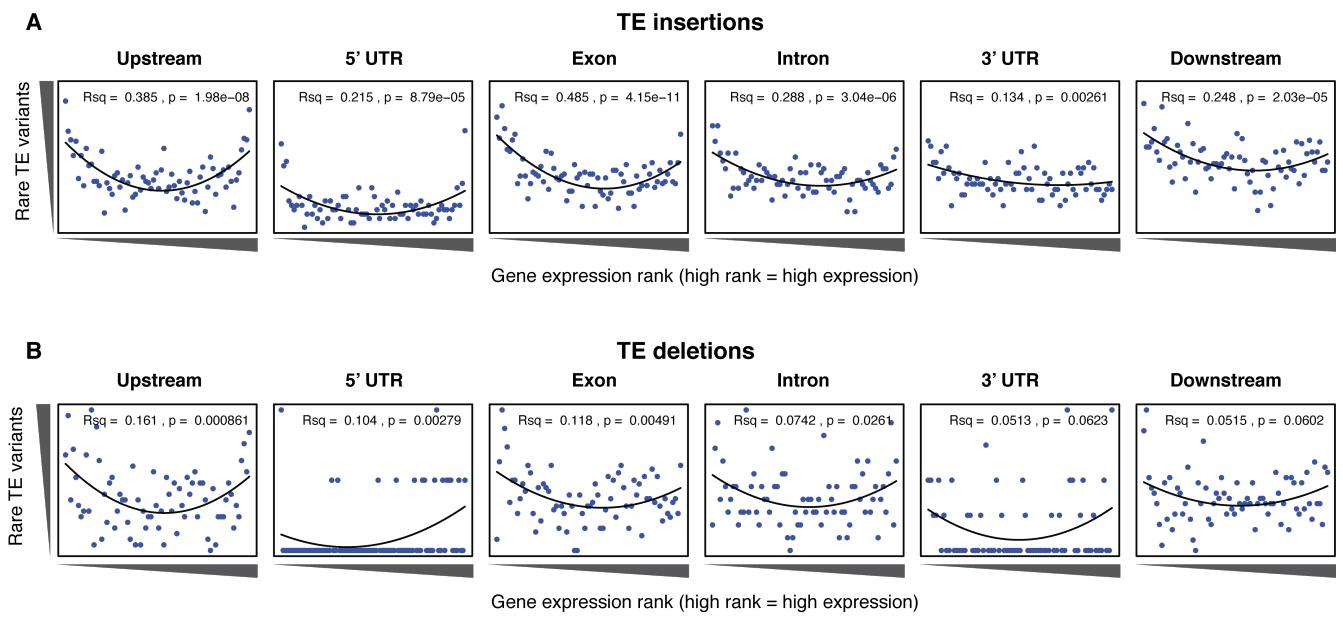
**Figure 4:** Differential transcript abundance associated with TE variant presence/absence

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



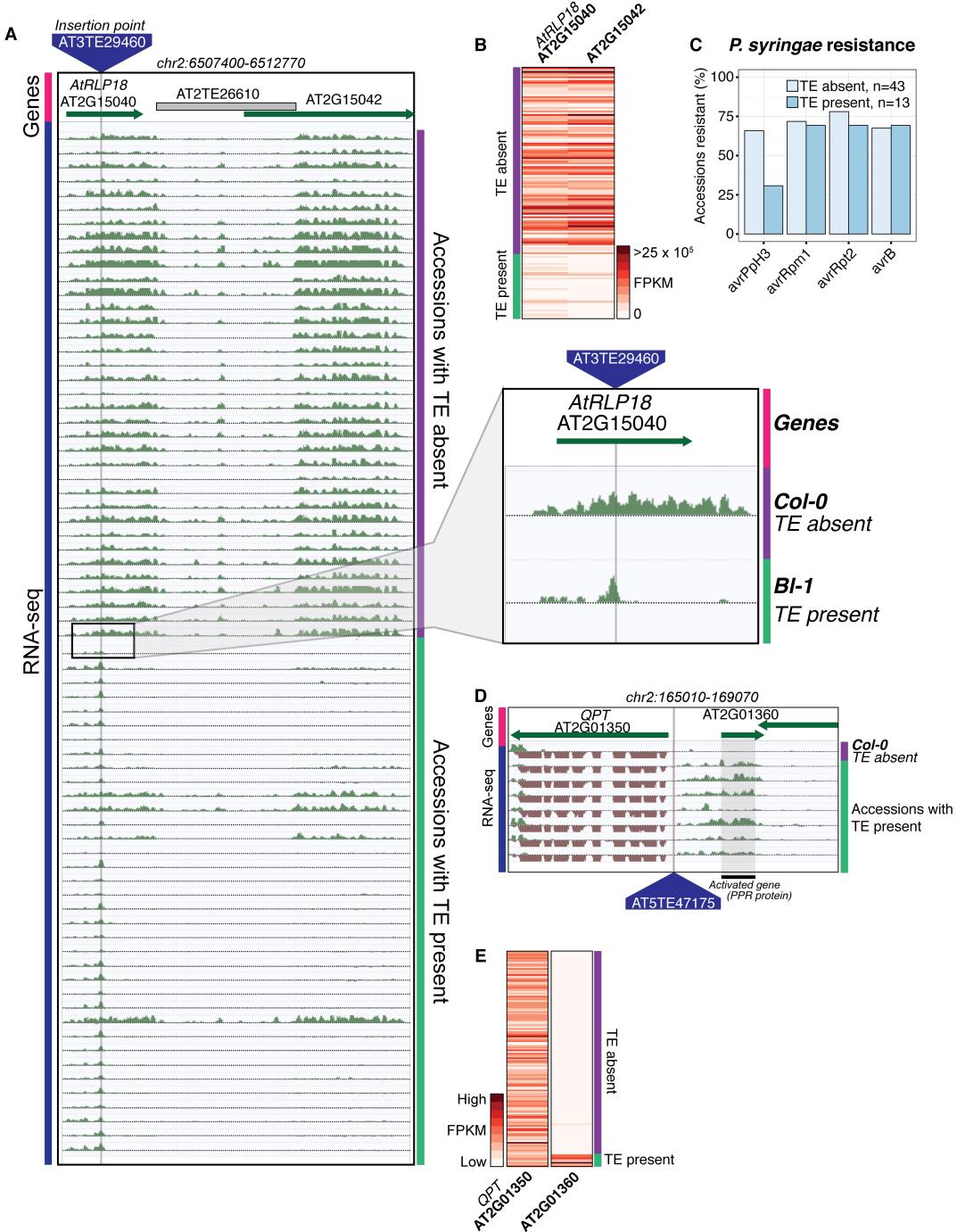
**Figure 4:** figure supplement 1

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



**Figure 4:** figure supplement 2

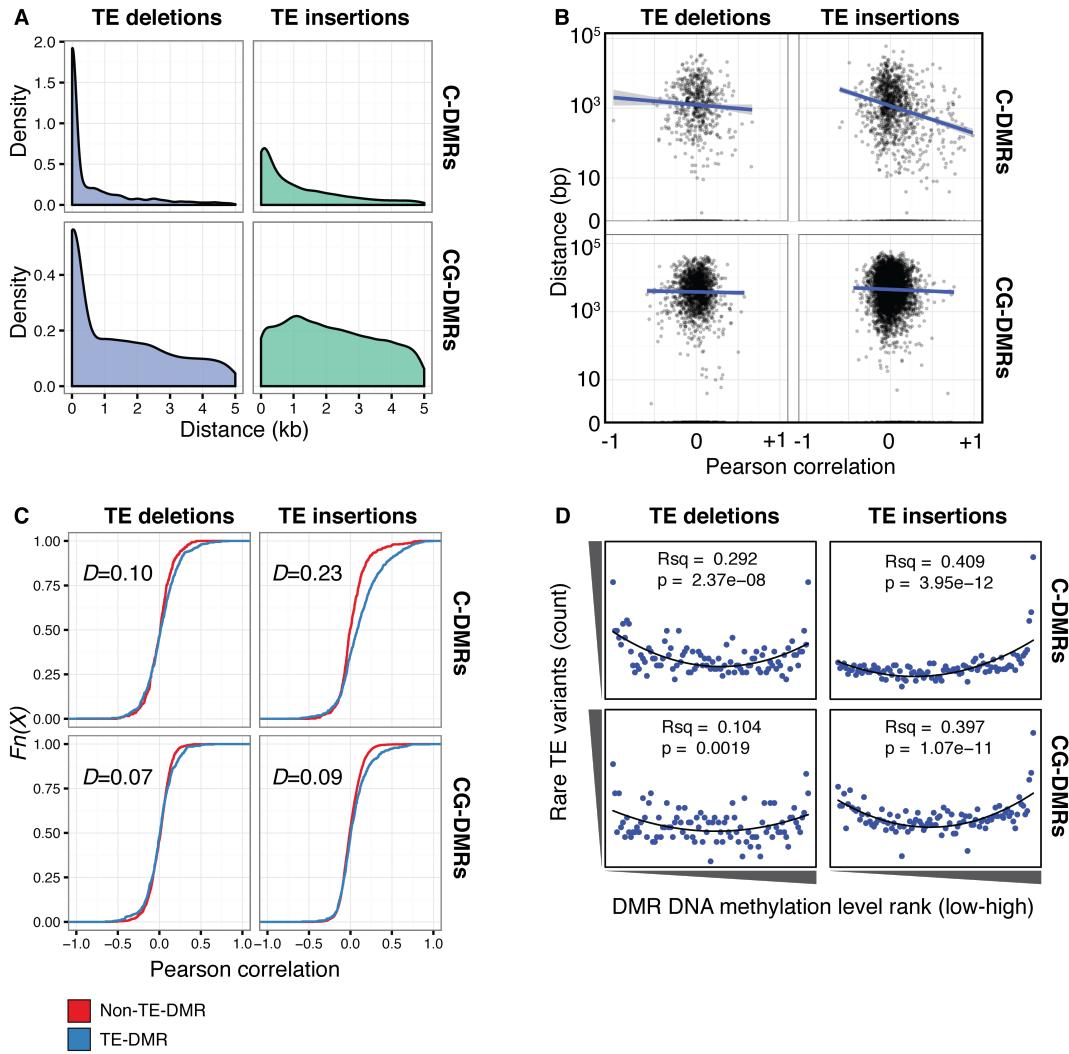
816 Relationship between rare TE variants and gene expression rank as for Figure 4B, for TE insertions  
 817 (A) and TE deletions (B) separately.



**Figure 5:** Effects of TE variants on local gene expression

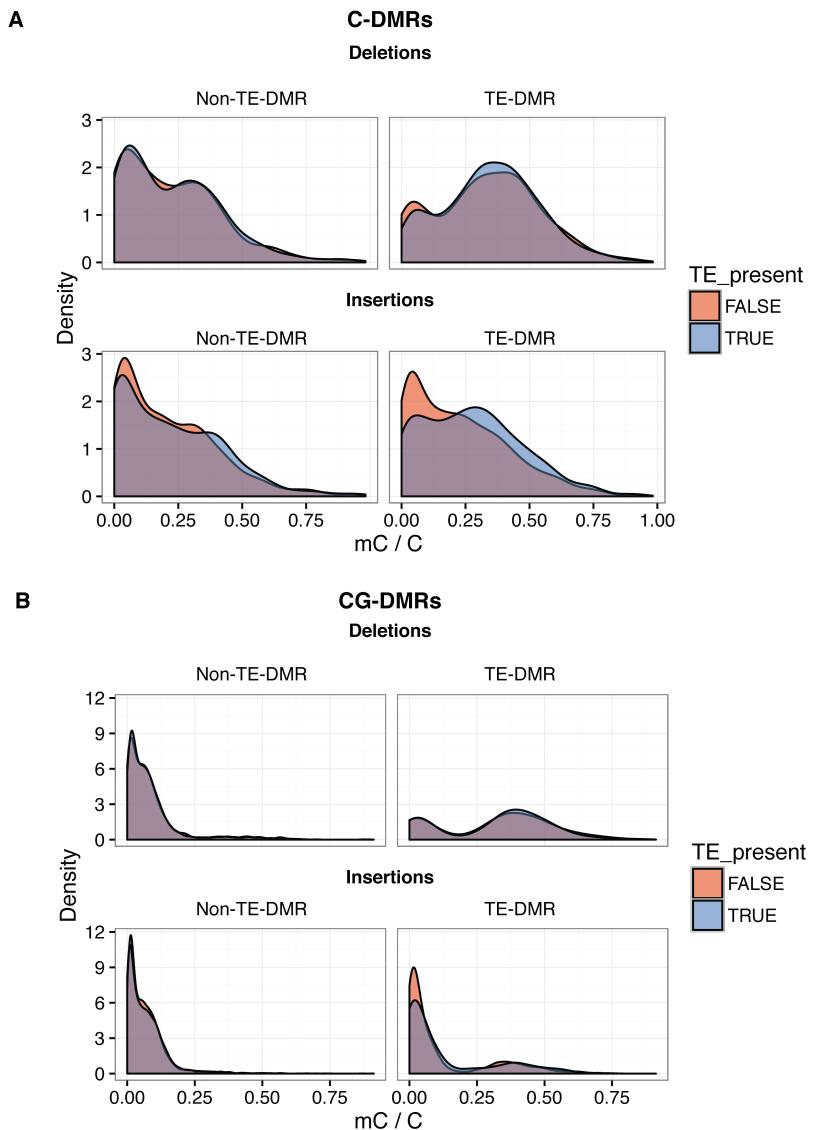
- (A) Genome browser representation of RNA-seq data for genes *AtRLP18* (AT2G15040) and a leucine-rich repeat family protein (AT2G15042) for Db-1, containing a TE insertion within the exon of the gene *AtRLP18*, and for a Col-0 (not containing the TE insertion within the exon of *AtRLP18*). Inset shows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

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



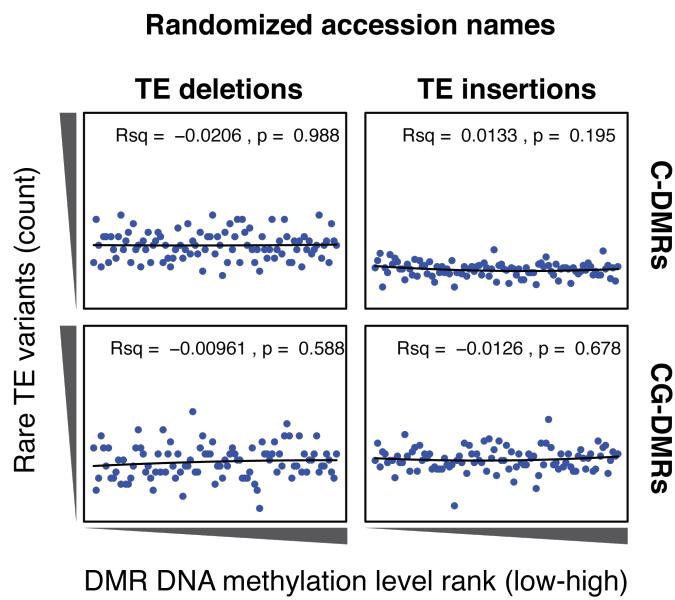
**Figure 6:** TE variants are associated with nearby DMR methylation levels

- (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 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.



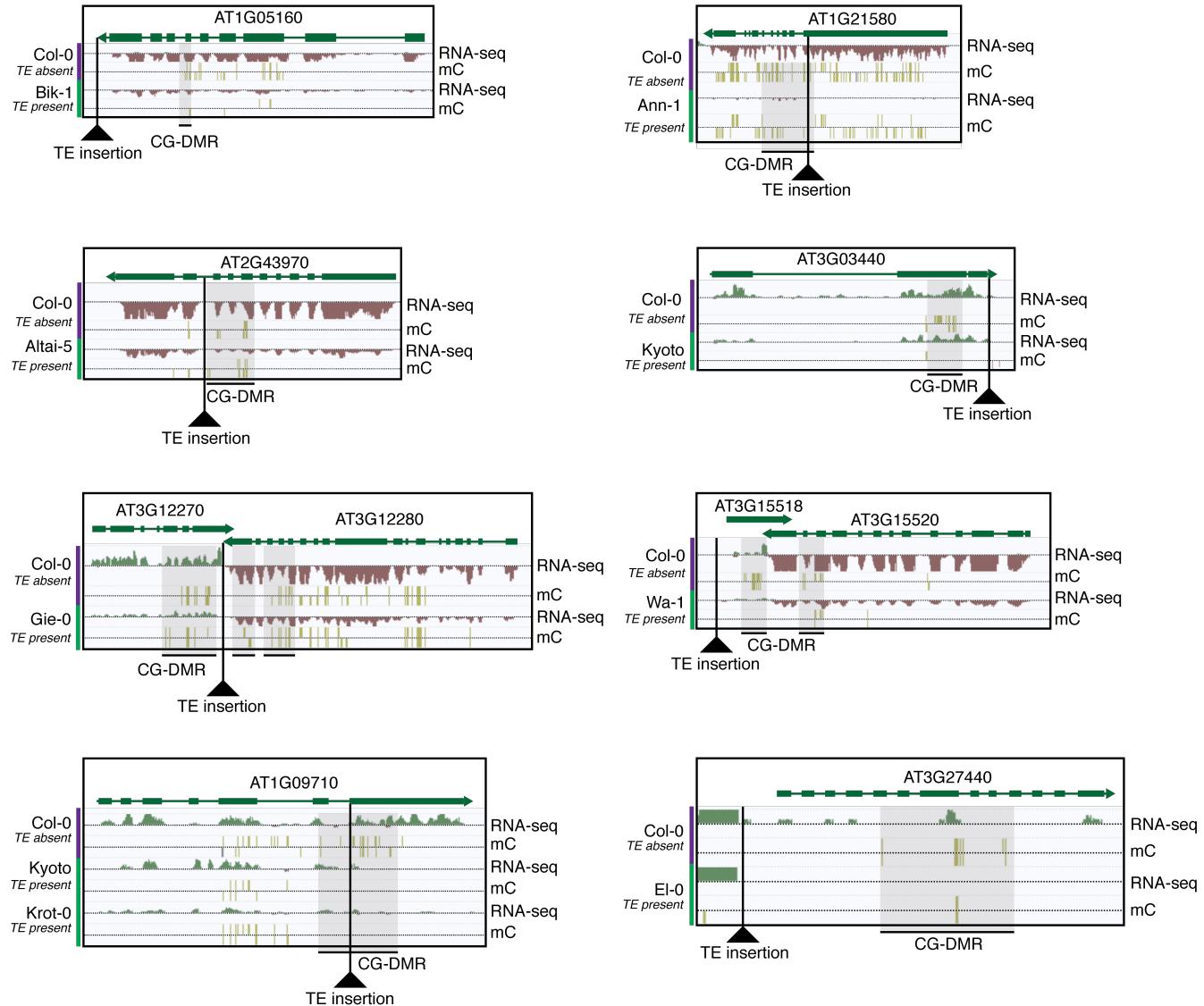
**Figure 6:** figure supplement 1

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



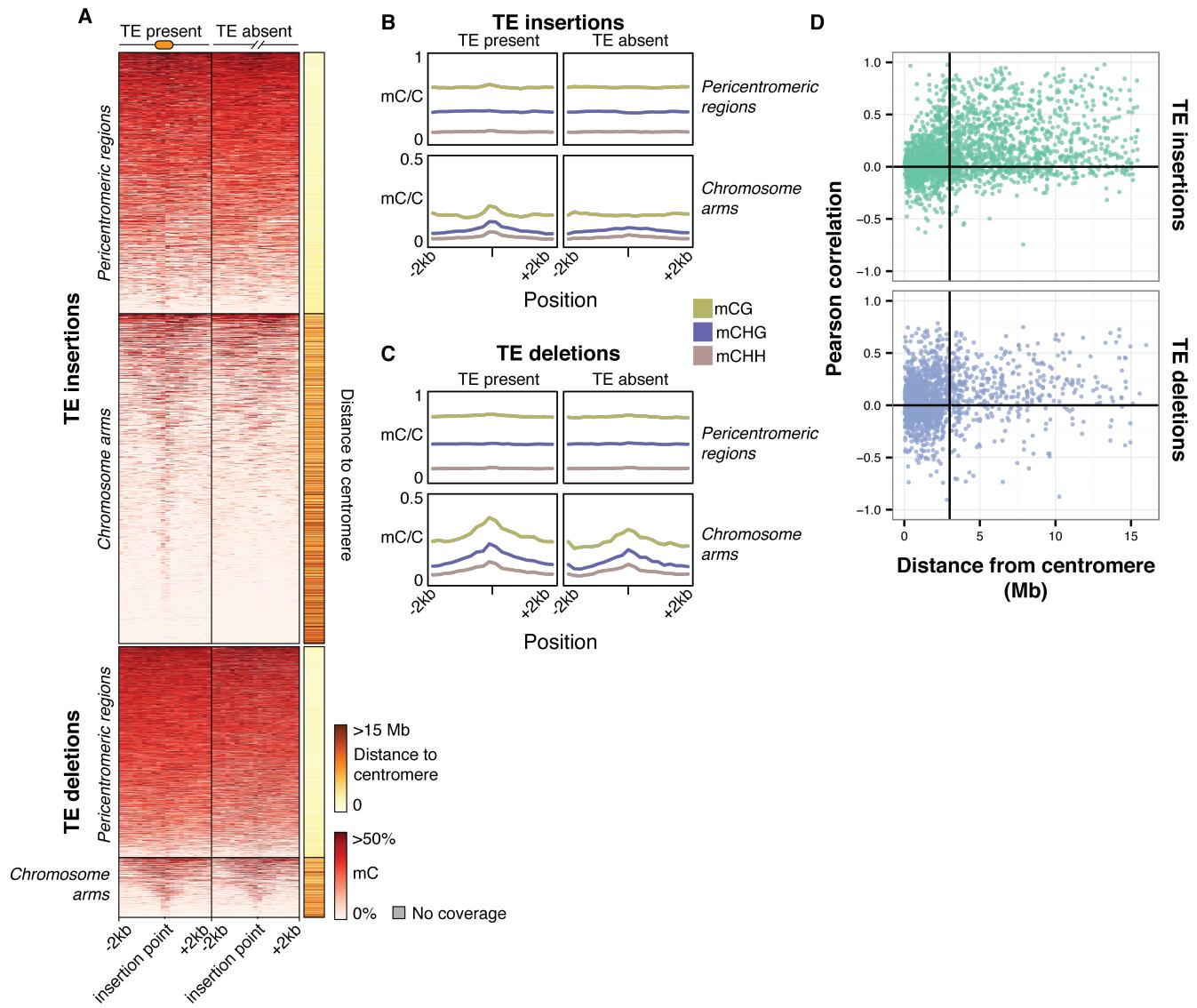
**Figure 6:** figure supplement 2

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



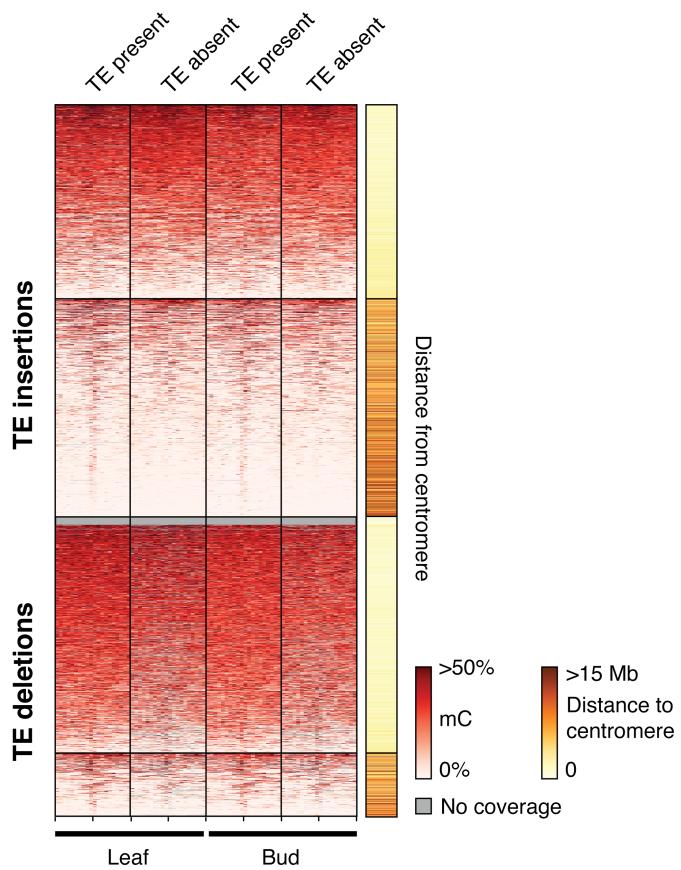
**Figure 6:** figure supplement 3

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



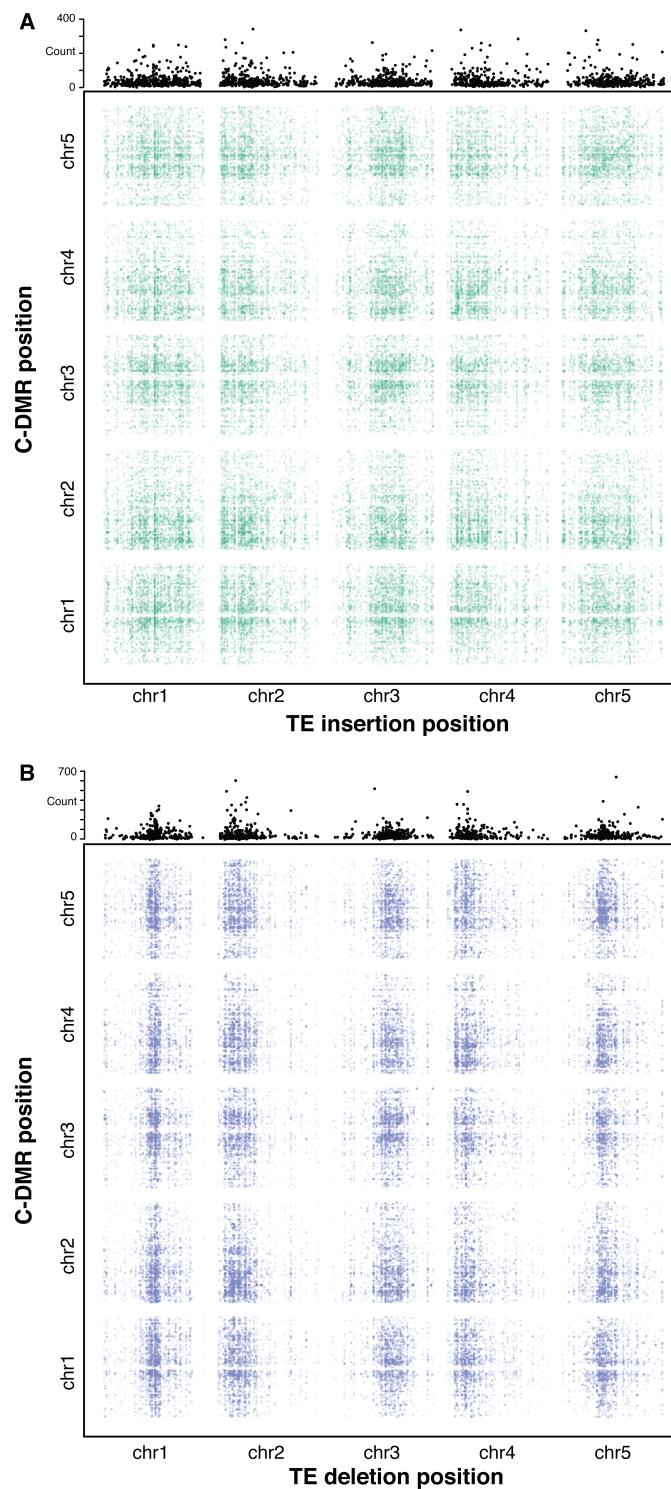
**Figure 7:** Local patterns of DNA methylation surrounding TE variant sites

- (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

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



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

- (A) Significant correlations between TE insertions and C-DMR DNA methylation level. Points show correlations between individual TE-DMR pairs that were more extreme than any of 500 permutations of the DMR data. Top plots show the total number of significant correlations for each TE insertion across the whole genome.
- (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
<b>TE deletions</b>	8.7	16	0.0078	4.3	16	0.0041
<b>TE insertions</b>	36	26	0.0089	9.4	26	0.0047
<b>NA calls</b>	3.4	6.2	0.0052	1.7	6.2	0.0027
<b>Total</b>	48	41	0.01	15	42	0.0054