TITLE

Population scale mapping of transposable element diversity uncovers novel genetic diversity linked to gene regulation and epigenomic variation

KEYWORDS

Epigenetics, DNA methylation, transposable elements, genomics, Arabidopsis

AUTHORS

Tim Stuart[1], Steven R. Eichten[2], Jonathan Cahn[1], Justin Borevitz[2], Ryan Lister[1]

AFFILIATIONS

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

2 - ARC 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)

ABSTRACT

Variation in the presence or absence of transposable element (TE) insertions in eukaryotic genomes is a significant source of genetic variation between individuals. Here, we identified 23,095 TE presence/absence variants between 216 wild Arabidopsis accessions. TE variants were identified that were associated with altered expression of nearby genes, indicating a possible functional role in the evolution of gene regulation. A majority of previously identified regions of differential DNA methylation between these accessions were associated with nearby TE variants, indicating an important role in facilitating epigenomic variation. Most TE variants are rare, however more than two thirds of the common alleles identified were not in linkage disequilibrium with nearby SNPs, revealing a potential rich source of missing heritability. An examination of TE regulation following intraspecific crosses between parents with differential TE content revealed that the absence of individual TEs from the paternal genome leads to DNA demethylation within maternal copies of these TEs in the embryo, reminiscent of Drosophila hybrid dysgenesis. This suggests that genetic differences between parents may lead to the formation of DNA methylation differences observed between individuals, and might enable activation of previously silent TEs.

INTRODUCTION

Transposable elements (TEs) are mobile genetic elements present in nearly all studied organisms, and make up a large fraction of most eukaryotic genomes. The two classes of TEs are retrotransposons (class I elements), which transpose via an RNA intermediate requiring a reverse transcription reaction, and DNA transposons (class II elements), which transpose via either a cut-paste or, in the case of Helitrons, a rolling circle mechanism with no RNA intermediate (Wicker et al. 2007). TE activity poses significant mutagenic potential because TE

48 insertion may disrupt essential regions of the genome, and so safeguards have evolved in order
49 to suppress this activity. These safeguards include epigenetic transcriptional silencing
50 mechanisms, chiefly involving the methylation of cytosine nucleotides (DNA methylation) to
51 produce 5-methylcytosine (mC), a mark that can signal transcriptional silencing of the
52 methylated locus. In *Arabidopsis thaliana* (Arabidopsis), DNA methylation occurs in three DNA
53 sequence contexts: mCG, mCHG, and mCHH, where H is any base but G. Establishment of
54 DNA methylation marks can be carried out by two distinct pathways – the RNA-directed DNA
55 methylation pathway guided by 24 nucleotide (nt) small RNAs (smRNAs), and the DDM1/CMT2
56 pathway (Zemach et al. 2013; Matzke and Mosher 2014). A major function of DNA methylation
57 in Arabidopsis is in the transcriptional silencing of TEs. Loss of DNA methylation due to
58 mutations in genes essential for DNA methylation establishment and maintenance leads to
59 expression of previously silent TEs, and sometimes transposition (Mirouze et al. 2009; Miura et
60 al. 2001; Saze et al. 2003; Lippman et al. 2004; Jeddeloh et al. 1999; Zemach et al. 2013).
61 TEs are thought to play an important role in evolution, not only because of the disruptive
62 potential of their transposition. The release of transcriptional and posttranscriptional silencing of
63 TEs can lead to bursts of TE activity, quickly generating new genetic diversity on which
64 selection may act (Vitte et al. 2014). TEs may carry regulatory information such as promoters
65 and transcription factor binding sites, and their mobilization may lead to the creation or
66 expansion of gene regulatory networks (Hénaff et al. 2014; Bolger et al. 2014; Ito et al. 2011;
67 Makarevitch et al. 2015). Furthermore, the transposase enzymes required and encoded by TEs
68 have frequently been domesticated and repurposed as endogenous proteins, such as the
69 *DAYSLEEPER* gene in Arabidopsis, derived from a hAT transposase enzyme (Bundock and
70 Hooykaas 2005). Clearly, the activity of TEs can have widespread and unpredictable effects on
71 the host genome. However, the identification of TE presence/absence variants in genomes has
72 remained difficult. It is challenging to identify the structural variants caused by TE mobilization

using current short-read sequencing technologies as these reads are typically mapped to a reference genome, which has the effect of masking structural changes that may be present. However, in terms of the number of base pairs affected, a large fraction of genetic differences between Arabidopsis accessions appear to be due to variation in TE content (Cao et al. 2011). Therefore, identification of TE variants is essential in order to develop a more comprehensive understanding of the genetic variation that exists between genomes, and of the consequences of TE movement upon genome and cellular function.

The tools developed previously for identification of novel TE insertion sites have several limitations. They either require a library of active TE sequences, cannot identify TE absence variants, or are not designed with population studies in mind (Thung et al. 2014; Robb et al. 2013; Hénaff et al. 2015). In order to accurately map the locations of TE presence/absence variants with respect to a reference genome, we have developed a novel algorithm, TEPID (Transposable Element Polymorphism IDentification), which is designed for population studies. We tested our algorithm using both simulated and real Arabidopsis sequencing data, finding that TEPID is able to accurately identify TE presence/absence variants with respect to the Col-0 reference genome. We applied our TE analysis method to existing genome resequencing data for 216 different wild Arabidopsis accessions, and identified widespread TE variation amongst these accessions (Schmitz et al. 2013). The majority of these TE variants arose recently in evolutionary times, represent novel genetic variants, and are associated with a variety of epigenomic and transcriptional variation.

RESULTS

Computational identification of TE presence/absence variation

We developed TEPID, an analysis pipeline capable of detecting TE presence/absence variants from paired end DNA sequencing data. TEPID integrates split and discordant read mapping

98 information, read mapping quality, sequencing breakpoints, as well as local variations in
99 sequencing coverage to identify novel TE presence/absence variants with respect to a
100 reference TE annotation (Fig. 1A; see methods). After TE variant discovery has been
101 performed, TEPID then includes a second step designed for population studies. This examines
102 each region of the genome where there was a TE insertion identified in any member of the
103 group, and checks for evidence of this insertion in each member of the population. In this way,
104 TEPID leverages TE variant information for a group of related samples to correct false negative
105 calls within the group. This feature sets TEPID apart from previous similar methods for TE
106 variant discovery using short read data (Hénaff et al. 2015). Testing of TEPID using simulated
107 TE variants in the Arabidopsis genome showed that it was able to reliably detect simulated TE
108 variants (Fig. 1B). In order to assess further the accuracy of TE variant discovery using TEPID,
109 we compared our predicted TE variants identified in the Landsberg *erecta* (L*er*) accession with
110 the *de novo* assembly reference genome created using long sequencing reads. Previously
111 published 100 bp paired-end L*er* genome resequencing reads (Schneeberger et al. 2011) were
112 first analyzed using TEPID, enabling identification of 446 TE insertions and 758 TE absence
113 variants with respect to the Col-0 reference TE annotation (Supplementary File 1). Reads
114 providing evidence for these variants were then mapped to the L*er* reference genome that was
115 generated by *de novo* assembly using Pacific Biosciences P5-C3 chemistry with a 20 kb insert
116 library (Chin et al. 2013), using the same alignment parameters as was used to map reads to
117 the Col-0 reference genome. This resulted in 98.7% of reads being aligned concordantly to the
118 L*er* reference, whereas 100% aligned discordantly or as split reads to the Col-0 reference
119 genome (Table 1). To find whether reads mapped to homologous regions in both the Col-0 and
120 L*er* reference genomes, we conducted a blast search (Camacho et al. 2009) using the DNA
121 sequence between read pair mapping locations in the L*er* genome against the Col-0 genome,
122 and found the top blast result for 80% of reads providing evidence for TE insertions, and 89% of

123 reads providing evidence for TE absence variants in L*er*, to be located within 200 bp of the TE
124 variant reported by TEPID. We conclude that reads providing evidence for TE variants map
125 discordantly or as split reads when mapped to the Col-0 reference genome, but map
126 concordantly to homologous regions of the L*er de novo* assembled reference genome,
127 indicating that structural variation is present at the sites identified by TEPID, and that this is
128 resolved in the *de novo* assembled genome.
129
130 **Abundant TE positional variation among natural Arabidopsis populations**
131 We used TEPID to analyze previously published 100 bp paired-end genome resequencing data
132 for 216 different wild Arabidopsis accessions (Schmitz et al. 2013), and identified 15,007 TE
133 insertions and 8,088 TE absence variants, totaling 23,095 unique TE variants. In most wild
134 accessions we identified 300-500 TE insertions (mean = 378; Fig. S1A) and 1,000-1,500 TE
135 absence variants (mean = 1,279; Fig. S1B), the majority of which were shared by two or more
136 accessions (Fig. S2). PCR validations were performed for a random subset of 10 insertions and
137 10 absence variants in 14 accessions, and confirmed the high accuracy of TE variant discovery
138 using the TEPID package, with results similar to that observed using simulated data (Fig. 1C,
139 D). TE variants were distributed throughout chromosome 1 in a pattern that is similar to the
140 distribution of all Col-0 TEs, and were enriched in the pericentromeric regions (Fig. 2A, S3). This
141 distribution was also similar to that observed for regions of the genome previously identified as
142 being differentially methylated in all DNA methylation contexts (mCG, mCHG, mCHH) between
143 the wild accessions (population DMRs) (Schmitz et al. 2013). Furthermore, TE variants were
144 depleted within genes and DNase I hypersensitivity sites (Sullivan et al. 2014), while they were
145 enriched in gene flanking regions and within other annotated TEs or pseudogenes (Fig. 2B).
146 Among the identified TE variants, several TE superfamilies were over- or under-
147 represented compared to the number expected by chance given the overall genomic frequency

148 of different TE types (Fig. 2C; Table S1, S2). In particular, TE variants in the RC/Helitron
149 superfamily were less numerous than expected, with an 11.5% depletion of RC/Helitron
150 elements in the set of TE variants. TEs belonging to the LTR/Gypsy superfamily were more
151 variable than expected, with a 7.5% enrichment in the set of TE variants. This was unlikely to be
152 due to a differing ability of our detection methods to identify TE variants of different lengths, as
153 the TE variants identified had a similar distribution of lengths as all Arabidopsis TEs annotated
154 in the Col-0 reference genome (Fig. S4A-C). These enrichments suggest that the RC/Helitron
155 TEs have been relatively dormant in recent evolutionary history, while the LTR/Gypsy TEs have
156 been more active. At the family level, we observed similar patterns of TE variant enrichment or
157 depletion (Fig. S5). As expected, this TE diversity tended to reflect the known genetic
158 relationships between accessions. We further examined Arabidopsis (Col-0) DNA sequencing
159 data from a transgenerational stress experiment to investigate the possible minimum number of
160 generations required for TE variants to arise (Jiang et al. 2014). We identified a single potential
161 TE insertion in a sample following 10 generations of single-seed descent under high salinity
162 stress conditions, and no TE variants in the control single-seed descent set. However, without
163 experimental validation it remains unclear if this represents a true variant. Therefore, we
164 conclude that TE variants likely arise at a rate less than 1 insertion in 30 generations.

165 Although thousands of TE variants were identified, it is important to classify their
166 relationship to other commonly identified sources of genetic variation such as single nucleotide
167 polymorphisms (SNPs). This comparison can determine if TE variants are often in linkage
168 disequilibrium with nearby SNPs, or if they are a previously unassessed source of genetic
169 variation between accessions unlinked to the underlying SNP-based haplotypes. To investigate
170 this relationship, SNPs previously identified between the wild accessions (Schmitz et al. 2013)
171 were compared to the presence/absence of individual TE variants. For the 7,300 testable TE
172 variants in the sample set with a minor allele frequency above 3%, the nearest 300 flanking

173 SNPs upstream and downstream of the TE insertion site were analyzed for linkage
174 disequilibrium (LD, r^2; Fig. 2D-F; see methods). TE variants were classified as being either
175 'young', 'mid', or 'old' variants by comparing ranked r^2 values to flanking SNPs against the
176 median ranked r^2 value for all SNP-SNP comparisons to account for regional variation in LD
177 (Fig. 2D, E). This analysis identified TE variants that were unlinked (young) or linked (old) to
178 SNPs forming local haplotypes, or were in an intermediate linkage state (mid) with their
179 surrounding haplotype. The majority (61%) of testable TE variants were largely unlinked to
180 nearby SNPs, and are therefore predicted to be young variants present in a few divergent
181 accessions (Fig. 2F). In contrast, 29% of TE variants displayed high levels of linkage with
182 nearby SNPs, and are likely old insertions. TE variants displayed a similar chromosomal
183 distribution regardless of age classification (Fig. 2G). Overall, this analysis revealed an
184 abundance of previously unexplored genetic variation that exists amongst Arabidopsis
185 accessions caused by the presence or absence of TEs, and illustrates the importance of
186 identifying TE variants to capture missing heritability alongside other genetic diversity such as
187 SNPs.
188
189 **TE variants affect gene expression**
190 To determine whether the TE variants identified affected nearby gene expression, we compared
191 the steady state transcript abundance within mature leaf tissue, between accessions with and
192 without TE insertions, for genes with TE variants located in the 2 kb gene upstream region, 5'
193 UTR, exon, intron, 3' UTR or 2 kb downstream region (Fig. 3). While the steady state transcript
194 abundance of most genes appeared to be unaffected by the presence of a TE, 45 genes
195 displayed significant differences in transcript abundance linked with the presence of a TE
196 variant, indicating a functional role for these variants in the local regulation of gene expression
197 (q-value < 0.001 with greater than 2-fold change in transcript abundance; Fig. 3, Table 2). We

198 did not find any functional category enrichments in this set of differentially expressed genes. It
199 should be noted that rare TE variants, with a minor allele frequency less than 3%, may also be
200 associated with a difference in transcript abundance, but were unable to be statistically tested
201 due to their rarity. Future studies using larger sample sizes may be able to further examine the
202 frequency at which TE variants impact gene expression.

203 As both increases and decreases in transcript abundance of nearby genes were
204 observed for TE variants within each gene feature, it appears to be difficult to predict the impact
205 a TE variant may have on nearby gene expression. However, gene-level transcript abundance
206 measurements may fail to identify the potential positional effect of TE variants upon
207 transcription. To more closely examine changes in transcript abundance associated with TE
208 variants among the accessions, we inspected a subset of TE variant sites and identified TE
209 variants that appear to have an impact on transcriptional patterns beyond changes in total
210 transcript abundance. For example, the presence of a TE within an exon of *AtRLP18*
211 (AT2G15040) was associated with truncation of the transcripts at the TE insertion site in
212 accessions possessing the TE variant, as well as silencing of a downstream gene encoding a
213 leucine-rich repeat protein (AT2G15042) (Fig. 4A-D). Both genes have significantly lower
214 transcript abundance in accessions containing the TE insertion ($p < 5.8 \times 10^{-10}$, Mann-Whitney U
215 test; Fig. 4C). *AtRLP18* is reported to be involved in bacterial resistance, with the disruption of
216 this gene by T-DNA insertion mediated mutagenesis resulting in increased susceptibility to the
217 bacterial plant pathogen *Pseudomonas syringae* (Wang et al. 2008). We examined pathogen
218 resistance phenotype data (Aranzana et al. 2005) for accessions with and without a TE insertion
219 in the *AtRLP18* exon, and found that accessions containing the TE insertion were sensitive to
220 infection by *Pseudomonas syringae* transformed with *avrPpH3* genes at a much higher
221 frequency (Fig. 4E). This suggests that the wild accessions identified here to contain a TE
222 insertion within *AtRLP18* may have an increased susceptibility to certain bacterial pathogens.

223 We also observed some TE variants associated with increased expression of nearby
224 genes. For example, presence of a TE within the upstream region of a gene encoding a
225 pentatricopeptide repeat (PPR) protein (AT2G01360) was associated with higher steady state
226 transcript abundance of this gene (Fig. 4F-H). Interestingly, transcription appeared to begin at
227 the TE insertion point, rather than the transcriptional start site of the gene (Fig. 4F). Accessions
228 containing the TE insertion had significantly higher AT2G01360 transcript abundance than the
229 accessions without the TE insertion ($p < 1.8 \times 10^{-7}$, Mann-Whitney U test; Fig. 4H). The
230 apparent transcriptional activation, linked with presence of a TE belonging to the *HELITRON1*
231 family, indicates that this element may carry sequences or other regulatory information that has
232 altered the expression of genes downstream of the TE insertion site. Importantly, this variant
233 was classified as a young TE insertion, as it is not in linkage disequilibrium with surrounding
234 SNPs, and therefore the associated changes in gene transcript abundance would not be
235 identified using only SNP data. This TE variant was also upstream of *QPT* (AT2G01350),
236 involved in NAD biosynthesis (Katoh et al. 2006), which did not show alterations in steady state
237 transcript abundance associated with the presence of the TE variant, indicating a potential
238 directionality of regulatory elements carried by the TE (Fig. 4G, H). Overall, these examples
239 demonstrate that TE variants can have unpredictable, yet important, effects of the expression of
240 nearby genes, and these effects may be missed by studies focused on genetic variation at the
241 level of SNPs.
242
243 **TE variants drive DNA methylation differences between accessions**
244 As TEs are frequently highly methylated in Arabidopsis (Lister et al. 2008; Cokus et al. 2008;
245 Zhang et al. 2006; Zilberman et al. 2007), we next assessed the DNA methylation state
246 surrounding TE variant sites to determine whether TE variants might be responsible for some of
247 the differences in DNA methylation patterns previously observed between the wild accessions

248 (Schmitz et al. 2013). We found that 61% of the 13,485 previously reported population DMRs
249 were located within 1 kb of a TE variant, significantly more than expected by chance ($p < 1 \times 10^{-}$
250 ^{4}, determined by resampling 10,000 times; Fig. 5A). Old TE variants were more often located
251 close to population DMRs, with 41.5% of TE variants within 1 kb of a population DMR classified
252 as old, 11.5% higher than the total genomic frequency of old elements (Fig. 5B). There was a
253 corresponding depletion of young TE variants in this set, with 48.5% being classified as young,
254 11.5% lower than the total genomic frequency of young TE variants. This indicates that
255 established TE insertions may be more important than recent TE insertions in defining DNA
256 methylation landscapes. Alternatively, unmethylated TEs may be preferentially lost from the
257 population due to selection, resulting in old TE variants being more highly methylated. DNA
258 methylation levels at population DMRs located within 1 kb a TE variant, henceforth termed TE-
259 DMRs, were positively correlated with the presence of the TE variant (Fig. 5C), while DNA
260 methylation levels at population DMRs further than 1 kb from a TE variant did not have a
261 significant association with the presence/absence of the nearest TE variant. TE-DMRs were
262 significantly more highly methylated in accessions containing the TE, suggesting that TE
263 variants may facilitate changes in DNA methylation patterns between accessions (Welch's t-
264 test, $p < 2.2 \times 10^{-16}$; Fig. 5D). Overall, this indicates that a large fraction of the population DMRs
265 previously identified between these accessions are associated with the presence of local TE
266 insertions.

267 We next examined levels of DNA methylation in regions flanking all TE insertions
268 regardless of the presence or absence of a population DMR call, and similarly found that
269 accessions containing a TE insertion had a highly localized enrichment in DNA methylation in all
270 sequence contexts (Fig. 5E, F), again indicating that TE variants may play a role in shaping
271 DNA methylation landscapes between Arabidopsis accessions. As the increase in DNA
272 methylation around TE insertion sites appeared to be restricted to regions >200 bp from the

273 insertion site, we correlated DNA methylation levels in 200 bp regions flanking TE variants with
274 the presence/absence of TE variants. DNA methylation levels were positively correlated with the
275 presence of a TE (Fig. 5G). Furthermore, DNA methylation level was more strongly correlated
276 with the presence of old TE variants. These results indicate that DNA methylation patterns are
277 influenced by the differential TE content of individual genomes, and that the DNA methylation-
278 dependent silencing of TE variants may lead to formation of DMRs between wild Arabidopsis
279 accessions. The age of TE variants also appears to be related to the DNA methylation state
280 surrounding the TE insertion site, with older variants being more highly methylated, suggesting
281 that many generations may be required before a new TE insertion reaches a highly methylated
282 state.
283
284 **TE regulation in the germline**
285 To explore when during development germline changes in TE content may occur, we sought to
286 associate TE DNA methylation dynamics during germline development and early
287 embryogenesis with the differential TE content we observe between the different Arabidopsis
288 accessions. Complex DNA methylation changes occur in the Arabidopsis germline, particularly
289 within TE sequences. During male germline development, the haploid microspore undergoes
290 asymmetric division to produce the larger vegetative cell and a smaller cell that divides once
291 more to produce two identical sperm cells (Fig. 6A) (Kawashima and Berger 2014). Some TEs
292 have previously been found to become transcriptionally active and to transpose in the
293 vegetative cell nucleus (VN). This TE activation has been linked to an increase in 21 nt smRNAs
294 derived from activated TE sequences, that are thought to then be transported to the sperm cells
295 (Slotkin et al. 2009). In contrast, the sperm and post-meiotic microspore genomes lose DNA
296 methylation in the mCHH context within many TEs due to decreased DRM2 abundance but are
297 able to maintain TE silencing (Slotkin et al. 2009; Calarco et al. 2012). A largely distinct set of

298 TEs are demethylated in the mCG context in the VN compared to the microspore and sperm
299 cells, and are thought to be involved in epigenetic imprinting rather than TE silencing (Calarco et
300 al. 2012). In the female germline, the ovule contains a diploid central cell and haploid egg cell.
301 As in the VN, the non-generative central cell acts as a companion to the egg and expresses
302 *DEMETER*, which encodes a DNA demethylase, resulting in global DNA demethylation (Ibarra
303 et al. 2012). During fertilization, one sperm fertilizes the egg to form the diploid embryo, while
304 the other sperm fertilizes the diploid central cell to form the triploid endosperm, which supplies
305 nutrients to the developing embryo (Fig. 6A) (Kawashima and Berger 2014). The endosperm
306 remains globally demethylated, while DNA methylation is gradually restored in the embryo
307 through RdDM (Jullien et al. 2012; Hsieh et al. 2009).

308 In order to determine whether these TEs that undergo DNA methylation changes in the
309 male germline or in the embryo and endosperm are especially active elements, we examined
310 both the number of unique insertions caused by these differentially methylated elements
311 (transposition frequency), as well as the presence/absence variability of these TEs among the
312 wild accessions (Fig. 6B, C) (Hsieh et al. 2009; Calarco et al. 2012). We found that 68% of TEs
313 previously found to be demethylated in the mCHH context in Col-0 sperm and microspore were
314 absent from non-Col-0 accessions (Fig. 6B), a much larger fraction than expected by chance (p
315 $< 1 \times 10^{-4}$, determined by resampling 10,000 times) (Calarco et al. 2012). In contrast, TEs
316 differentially methylated in the mCG context in the sperm or microspore, thought to be involved
317 in epigenetic imprinting, or in the mCHH context in the developing embryo (Hsieh et al. 2009),
318 were not absent from non-Col-0 accessions significantly more often than expected ($p > 0.09$,
319 determined by resampling 10,000 times; Fig. 6B). While TEs mCG-hypomethylated in the
320 endosperm were absent from non-Col-0 genomes significantly more often than expected ($p < 1$
321 $\times 10^{-4}$, determined by resampling 10,000 times), the scale of this difference was small, with only
322 30% of the TEs absent in non-Col-0 accessions. mCHH-demethylated TEs in the sperm and

323 microspore also had a significantly higher transposition frequency, suggesting that they have
324 been more active than most other TEs in the Arabidopsis genome (Fig. 6C; Welch's t-test, $p < 1$
325 $\times 10^{-5}$). While mCHH-hypermethylated TEs in the embryo were not found to be frequently
326 absent from non-Col-0 accessions, these elements did show a significantly higher transposition
327 frequency, as did elements mCG-hypomethylated in the endosperm, indicating that
328 demethylation of active TEs in the germline companion cells (the VN and endosperm) may be
329 important in driving DNA methylation changes in the generative cells (the sperm and embryo),
330 perhaps acting to suppress transposition of active elements in the germline.

331 TEs that are demethylated in the mCHH context in the developing sperm may rely on
332 smRNAs in the seed for the restoration of proper DNA methylation of paternal TE sequences in
333 the embryo following fertilization, through RdDM (Jullien et al. 2012). These smRNAs are
334 thought to be maternally-derived, and are likely produced in the endosperm (Mosher et al. 2009;
335 Calarco et al. 2012). We sought to determine whether TEs present in only one parental genome
336 in an intraspecific cross would show altered levels of smRNAs targeted to these TEs, as the
337 absence of a TE in the maternal genome may lead to the loss of a maternally-derived smRNA
338 signal in the seed. Using previously published whole seed 21-24 nt smRNA data for reciprocal
339 crosses between Col-0 and Cvi or L*er* (Pignatta et al. 2014; Gehring et al. 2014), we examined
340 seed smRNA levels in TEs present in the Col-0 genome but absent from the Cvi genome for
341 crosses between Col-0 and Cvi, or absent from the L*er* genome for crosses between Col-0 and
342 L*er* (Fig. 6D, direction of each cross is depicted as *female x male* on axis labels). Consistent
343 with previous reports, we found that smRNA levels in TEs were dependent on maternal
344 genotype but independent of paternal genotype, indicating that they are maternally-derived.
345 Furthermore, we confirmed that TEs present in only the paternal genome in a cross had
346 significantly lower 21-24 nt smRNA levels compared with crosses where the TE was present in
347 the maternal genome ($p < 1.4 \times 10^{-10}$, Mann-Whitney U test; Fig. 6E). As these smRNAs are

348 thought to be required for TE silencing in the embryo (Jullien et al. 2012; Matzke and Mosher
349 2014), this may then lead to TE activation and may explain the higher transposition frequencies
350 we observe for TEs demethylated in the mCHH context in the sperm genome.

351 To test whether the absence of a TE in the maternal genome was sufficient to prevent
352 the re-methylation of the paternal copy of that TE following fertilization, we examined previously
353 published embryonic DNA methylation data within Cvi-absent TEs following reciprocal crosses
354 between Col-0 and Cvi (Pignatta et al. 2014) (Fig. 6F). For all crosses between Col-0 and Cvi,
355 Cvi-absent TEs showed significantly lower mCG levels, and this is likely due to the global
356 reduction in mCG levels that exists in the Cvi genome rather than the absence of the TE in one
357 parent (Pignatta et al. 2014). Surprisingly, we found for TEs that lose mCHH in the sperm
358 (Calarco et al. 2012), when the TE was absent from the maternal genome (Cvi x Col-0),
359 embryonic mCHH levels within these TEs were not significantly different from DNA methylation
360 levels in the Col-0 x Col-0 cross, where the TE was present in both parents ($p > 0.04$, Mann-
361 Whitney U test; Fig. 6F). This was contrary to the anticipated pattern, where the loss of maternal
362 smRNAs observed for this cross (Fig. 6E) would prevent these sperm-demethylated TEs from
363 being re-methylated in the embryo, and may indicate that maternally-derived smRNAs are not
364 essential for establishing DNA methylation patterns in early embryonic development. However,
365 when the parents were reversed and the TE variants were present in the maternal genome but
366 absent from the paternal genome (Col-0 x Cvi), we found a significant reduction in mCHH
367 methylation for these TEs, indicating that paternal signals may be required for DNA methylation
368 of maternal TE sequences ($p < 5 \times 10^{-7}$, Mann-Whitney U test). To further investigate this result,
369 we analyzed an independent dataset containing embryonic DNA methylation data for crosses
370 between Col-0 and L*er*, generated at a slightly later developmental stage (7-8 days after
371 pollination) than the data for crosses between Col-0 and Cvi (6 days after pollination). Within
372 TEs absent from L*er* and demethylated in the Col-0 sperm (Ibarra et al. 2012) we found no

373 significant difference in DNA methylation levels in any context dependent on the parental
374 genotypes (Fig. S6). This could indicate that embryonic DNA methylation changes within
375 maternal TEs, dependent upon paternal genotype, may be restricted to early developmental
376 time points, less than 6 days after pollination, or may be dependent on the intraspecific cross
377 performed.

378 Considering the high proportion of sperm mCHH-demethylated TEs that were found to
379 be absent in non-Col-0 accessions (Fig. 6B), crosses between plants with differential
380 presence/absence of these TEs may be possible in the wild, and such crosses may lead to the
381 demethylation of the maternal copies of these TEs in the embryo when a paternal copy is
382 absent. These findings support a model for TE silencing escape facilitated by genetic
383 differences between parents, as the loss of DNA methylation within TEs has been shown
384 previously to be sufficient for transcriptional activation and transposition of demethylated TEs
385 (Gendrel et al. 2002; Hirochika et al. 2000; Stroud et al. 2013). Such mating situations may also
386 lead to the formation of stable epialleles depending of the length of time the TE sequences are
387 able to escape remethylation, and this may play a role in formation of the patterns of differential
388 DNA methylation previously observed between wild accessions (Schmitz et al. 2013).

389

390 **DISCUSSION**

391 Here, we discovered widespread differential TE content between wild Arabidopsis accessions. A
392 subset (32%) of TE variants with a minor allele frequency above 3% were able to be tested for
393 linkage with nearby SNPs. The majority of these TE variants were unlinked to surrounding
394 SNPs, indicating that they represent genetic variants currently overlooked in genomic studies.
395 We found a marked depletion of TE variants within gene bodies and DNase I hypersensitivity
396 sites, indicating that the more deleterious TE insertions have likely been removed from this
397 population through selection. Importantly, we were able to identify examples where TE variants

398 appear to have an effect upon gene expression, both in the disruption of transcription and in the
399 spreading or disruption of regulatory information leading to the transcriptional activation of
400 genes, indicating that these TE variants can have important consequences upon expression of
401 protein coding genes. Furthermore, we provide evidence that differential TE content between
402 genomes of wild Arabidopsis accessions underlies a large fraction of the previously reported
403 DNA methylation differences between accessions. Thus, the frequency of pure epialleles,
404 independent of underlying genetic variation, may be even more rare than previously anticipated
405 (Richards 2006). The level of DNA methylation changes associated with TE variants is related
406 to TE age, with old variants being more strongly correlated with increased DNA methylation
407 levels. This suggests that the methylation of new TE insertions is a gradual process that occurs
408 incrementally over many generations, or that unmethylated TEs are preferentially lost from the
409 population over time.

410 Identification of TE variants between Arabidopsis accessions has also enabled a closer
411 examination of the changes in TE smRNA and DNA methylation levels following fertilization of
412 intraspecific hybrids. smRNA levels in the seed appear to be dependent on maternal genome
413 content, as the presence of a TE in only the paternal genome of a cross is associated with
414 decreased levels of corresponding 21-24 nt smRNAs derived from those TEs in the seed. This
415 loss of smRNAs, dependent on maternal genotype, is strikingly similar to findings from studies
416 performed in *Drosophila melanogaster* over 3 decades ago, where maternal absence of
417 paternal *P* elements in a cross was found to lead to activation and frequent transposition of
418 these *P* elements, due to absence of maternally-derived smRNA signals needed for TE
419 silencing, while TE activation was not observed when the *P* elements were present in the
420 maternal genome (Bingham et al. 1982; Blumenstiel and Hartl 2005). This is thought to be the
421 underlying cause of hybrid dysgenesis, a phenotype characterized by sterility and high rates of
422 germline TE activity, and where the transpositions caused by active *P* elements are often lethal

423 to the hybrid offspring. It has been hypothesized previously that a process similar to Drosophila
424 hybrid dysgenesis may occur in Arabidopsis, as smRNAs in the Arabidopsis seed also appear
425 to be maternally-derived, and there is some prior evidence that this may be true (Martienssen
426 2010). Interspecific crosses between *Arabidopsis thaliana* females and *Arabidopsis arenosa*
427 males was observed to lead to the expression of previously silent paternal *ATHILA* TEs, thought
428 to be due to the presence of *ATHILA* elements in the *A. arenosa* genome that are absent from
429 the *A. thaliana* genome (Josefsson et al. 2006). However, the reciprocal cross is impossible to
430 generate due to pollination failure, constituting a fundamental limitation of this system for
431 studying hybrid dysgenesis. The use of wild *A. thaliana* accessions may prove more fruitful in
432 future experiments aiming to elucidate the processes of germline TE regulation in plants.

433 Surprisingly, we found that this decrease in smRNA levels targeting paternal TE
434 sequences was not linked to DNA demethylation of the corresponding paternal TEs in the
435 embryo, but instead observed an inverse relationship, where TEs absent from the paternal
436 genome in a cross were linked with embryonic demethylation of the maternal copy. This
437 indicates that maternal smRNAs are not essential for restoring the paternal patterns of DNA
438 methylation that are erased in the sperm, nor are they sufficient to maintain DNA methylation
439 within maternal TEs in the absence of paternal TE copies. As 21 nt smRNA production is greatly
440 increased within the pollen VN, and these smRNAs may be transported to the sperm cells
441 (Slotkin et al. 2009), it is possible that these paternal smRNAs remain present in the sperm
442 cytoplasm during fertilization. These smRNAs may play an important role in establishing early
443 patterns of DNA methylation in the embryo, perhaps explaining the non-reliance of paternal TE
444 sequences upon maternal smRNAs. This issue is somewhat complicated by the lack of embryo-
445 specific smRNA data, as all existing data have been generated from whole seed. If maternal
446 smRNAs are produced and remain in the endosperm rather than the embryo, this could further
447 explain the apparent reliance upon paternal silencing signals for proper establishment of DNA

448 methylation patterns following fertilization. Our data provides the first evidence that embryonic
449 TE silencing in Arabidopsis may be dependent on paternal, rather than maternal, silencing
450 factors.

451 DNA methylation changes triggered by genetic differences between parents clearly
452 occur in Arabidopsis, although perhaps via a different mechanism as is responsible for causing
453 hybrid dysgenesis in Drosophila. These DNA methylation changes that occur in the embryo may
454 play an important role in the formation of DNA methylation differences between wild Arabidopsis
455 accessions, depending on the length of time these changes are able to persist in hybrid
456 progeny. If the loss of DNA methylation within maternal TEs that we observe is able to persist in
457 hybrid plants, this may lead to the formation of stable epialleles as the demethylated TE copy is
458 propagated through the population. Further experiments will be required to determine the
459 stability of these embryonic changes in DNA methylation that occur in hybrid plants.
460 Alternatively, if these DNA methylation changes are limited to a small developmental window,
461 perhaps less than 6 days after pollination, there may only be a short period of time where TE
462 silencing is lost and TE activation can occur, leading to new TE insertions in the early embryo.
463 Overall, our results show that TE presence/absence variants between wild Arabidopsis
464 accessions can be linked to many DNA methylation changes previously observed in the
465 population, and can have important consequences upon nearby gene expression. Furthermore,
466 the differential TE content between parents can lead to DNA methylation changes in the early
467 embryo, and could lead to activation of these elements.

468

469 # METHODS

470 **TEPID development**

471 *Mapping*

472 FASTQ files are mapped to the reference genome using the 'tepid-map' algorithm (Fig. 1A).
473 This first calls bowtie2 (Langmead and Salzberg 2012) with the following options: '--local', '--
474 dovetail', '--fr', '-R5', '-N1'. Soft-clipped and unmapped reads are extracted using Samblaster
475 (Faust and Hall 2014), and remapped using the split read mapper Yaha (Faust and Hall 2012),
476 with the following options: '-L 11', '-H 2000', '-M 15', '-osh'. Split reads are extracted from the
477 Yaha alignment using Samblaster (Faust and Hall 2014). Alignments are then converted to bam
478 format, sorted, and indexed using samtools (Li et al. 2009).
479
480 *TE variant discovery*
481 The 'tepid-discover' algorithm examines mapped bam files generated by the 'tepid-map' step to
482 identify TE presence/absence variants with respect to the reference genome. Firstly, mean
483 sequencing coverage, mean library insert size, and standard deviation of the library insert size
484 is estimated. Discordant read pairs are then extracted, defined as mate pairs that map more
485 than 4 standard deviations from the mean insert size from one another, or on separate
486 chromosomes.
487 To identify TE insertions with respect to the reference genome, split read alignments are
488 first filtered to remove reads where the distance between split mapping loci is less than 5 kb, to
489 remove split reads due to small indels, or split reads with a mapping quality (MAPQ) less than 5.
490 Split and discordant read mapping coordinates are then intersected using pybedtools (Dale et
491 al. 2011; Quinlan and Hall 2010) with the Col-0 reference TE annotation, requiring 80% overlap
492 between TE and read mapping coordinates. To determine putative TE insertion sites, regions
493 are then identified that contain independent discordant read pairs aligned in an orientation
494 facing one another at the insertion site, with their mate pairs intersecting with the same TE (Fig.
495 1A). The total number of split and discordant reads intersecting the insertion site and the TE is
496 then calculated, and a TE insertion predicted where the combined number of reads is greater

497 than a threshold determined by the average sequencing depth over the whole genome (1/10
498 coverage if coverage is greater than 10, otherwise a minimum of 2 reads). Alternatively, in the
499 absence of discordant reads mapped in orientations facing one another, the required total
500 number of split and discordant reads at the insertion site linked to the inserted TE is set higher,
501 requiring twice as many reads.

502 To identify TE absence variants with respect to the reference genome, split and
503 discordant reads separated >20 kb from one another are first removed, as 99.9% of Arabidopsis
504 TEs are shorter than 20 kb, and this removes split reads due to larger structural variants not
505 related to TE diversity (Fig. S4A). Col-0 reference annotation TEs that are located within the
506 genomic region spanned by the split and discordant reads are then identified. TE absence
507 variants are predicted where at least 80% of the TE sequence is spanned by a split or
508 discordant read, and the sequencing depth within the spanned region is <10% the sequencing
509 depth of the 2 kb flanking sequence, and there are a minimum number of split and discordant
510 reads present, determined by the sequencing depth (1/10 coverage; Fig. 1A). A threshold of
511 80% TE sequence spanned by split or discordant reads is used, as opposed to 100%, to
512 account for misannotation of TE sequence boundaries in the Col-0 reference TE annotation, as
513 well as TE fragments left behind by DNA TEs during cut-paste transposition (TE footprints) that
514 may affect the mapping of reads around annotated TE borders (Plasterk 1991). This was found
515 to improve TE absence detection using simulated data. Furthermore, the coverage within the
516 spanned region may be more than 10% that of the flanking sequence, but in such cases twice
517 as many split and discordant reads are required. If multiple TEs are spanned by the split and
518 discordant reads, and the above requirements are met, multiple TEs in the same region can be
519 identified as absent with respect to the reference genome. Absence variants in non-Col-0
520 accessions are subsequently recategorized as TE insertions present in the Col-0 genome but
521 absent from a given wild accession.

522

523 *TE variant refinement*

524 Once TE insertions are identified using the 'tepid-map' and 'tepid-discover' algorithms, these
525 variants can be refined if multiple related samples are analyzed. The 'tepid-refine' algorithm is
526 designed to interrogate regions of the genome in which a TE insertion was discovered in other
527 samples but not the sample in question, and check for evidence of that TE insertion in the
528 sample using lower read count thresholds compared to the 'tepid-discover' step. In this way, the
529 refine step leverages TE variant information for a group of related samples to reduce false
530 negative calls within the group. This distinguishes TEPID from other similar methods for TE
531 variant discovery utilizing short sequencing reads. A file containing the coordinates of each
532 insertion, and a list of sample names containing the TE insertion must be provided to the 'tepid-
533 refine' algorithm, which this can be generated using the 'merge_insertions.py' script included in
534 the TEPID package. Each sample is examined in regions where there was a TE insertion
535 identified in another sample in the group. If there is a sequencing breakpoint within this region
536 (no continuous read coverage spanning the region), split reads mapped to this region will be
537 extracted from the alignment file and their coordinates intersected with the TE reference
538 annotation. If there are split reads present at the variant site that are linked to the same TE as
539 was identified as an insertion at that location, this TE insertion is recorded in a new file as being
540 present in the sample in question. If there is no sequencing coverage in the queried region for a
541 sample, an "NA" call is made indicating that it is unknown whether the particular sample
542 contains the TE insertion or not.

543

544 While the above description relates specifically to use of TEPID for identification of TE variants
545 in Arabidopsis in this study, this method can be also applied to other species, with the only

546 prerequisite being the annotation of TEs in a reference genome and the availability of paired-
547 end DNA sequencing data.
548
549 **TE variant simulation**
550 To test the sensitivity and specificity of TEPID, 100 TE insertions (50 copy-paste transpositions,
551 50 cut-paste transpositions) and 100 TE absence variants were simulated in the Arabidopsis
552 genome using the RSVSim R package, version 1.7.2 (Bartenhagen and Dugas 2013), and
553 synthetic reads generated from the modified genome at various levels of sequencing coverage
554 using wgsim (Li et al. 2009) (https://github.com/lh3/wgsim). These reads were then used to
555 calculate the true positive, false positive, and false negative TE variant discovery rates for
556 TEPID at various sequencing depths, by running 'tepid-map' and 'tepid-discover' using the
557 simulated reads with the default parameters (Fig. 1B).
558
559 **L*er* TE analysis**
560 Previously published 100 bp paired end sequencing data for L*er*
561 (http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/current/Ler-1/Reads/;
562 (Schneeberger et al. 2011)) was downloaded and analyzed with the TEPID package to identify
563 TE variants. Reads providing evidence for TE variants were then mapped to the *de novo*
564 assembled L*er* genome (Chin et al. 2013). To determine whether reads mapped to homologous
565 regions of the L*er* and Col-0 reference genome, the *de novo* assembled L*er* genome sequence
566 between mate pair mapping locations in L*er* were extracted, with repeats masked using
567 RepeatMasker with RepBase-derived libraries and the default parameters (version 4.0.5,
568 http://www.repeatmasker.org). A blastn search was then conducted against the Col-0 genome
569 using the following parameters: '-max-target-seqs 1', '-evalue 1e-6' (Camacho et al. 2009).

570 Coordinates of the top blast hit for each read location were then compared with the TE variant
571 sites identified using those reads.
572
573 **Arabidopsis TE variant discovery**
574 We ran the TEPID, including the insertion refinement step, on previously published sequencing
575 data for 216 different Arabidopsis populations (NCBI SRA SRA012474; Schmitz et al. 2013),
576 mapping to the TAIR10 reference genome and using the TAIR9 TE annotation. The '--mask'
577 option was also used to mask the mitochondrial and plastid genomes. We also ran TEPID using
578 previously published transgenerational data for salt stress and control conditions (NCBI SRA
579 SRP045804; Jiang et al. 2014), using the '--mask' option to mask mitochondrial and plastid
580 genomes, and the '--strict' option for highly related samples.
581
582 **TE variant / SNP comparison**
583 SNP information for 216 Arabidopsis accessions was obtained from the 1001 genomes data
584 center (http://1001genomes.org/data/Salk/releases/2013_24_01/; Schmitz et al. 2013). This was
585 formatted into reference (Col-0 state), alternate, or NA calls for each SNP. Accessions with both
586 TE variant information and SNP data were selected for analysis. Hierarchical clustering of
587 accessions by SNPs as well as TE variants were used to identify essentially clonal accessions,
588 as these would skew minor allele frequency calculations. A single representative from each
589 cluster of similar accessions was kept, leading to a total of 187 accessions for comparison. For
590 each TE variant with minor-allele-frequency greater than 3%, the nearest 300 upstream and 300
591 downstream SNPs with a minor-allele-frequency greater than 3% were selected. Pairwise
592 genotype correlations (r^2 values) for all complete cases were obtained for SNP-SNP and SNP-
593 TE variant states. r^2 values were then ordered by decreasing rank and a median SNP-SNP rank
594 value was calculated. For each of the 600 ranked surrounding positions, the number of times

595 the TE rank was greater than the SNP-SNP median rank was calculated as a relative 'age'
596 metric of TE to SNP. TE variants with less than 200 ranks over the SNP-SNP median were
597 classified as 'young' insertions. Mid-age TE variants had ranks between 200 and 400, while TE
598 variants with greater than 400 ranks above their respective SNP-SNP median value were
599 classified as 'old' variants.
600
601 **PCR validations**
602 *Selection of accessions to be genotyped*
603 To assess the accuracy of TE variant calls in accessions with a range of sequencing depths of
604 coverage, we grouped accessions into quartiles based on sequencing depth of coverage and
605 randomly selected a total of 14 accessions for PCR validations from these quartiles. DNA was
606 extracted for these accessions using Edward's extraction protocol (Edwards et al. 1991), and
607 purified prior to PCR using AMPure beads.
608
609 *Selection of TE variants for validation and primer design*
610 Ten TE insertion sites and 10 TE absence sites were randomly selected for validation by PCR
611 amplification. Only insertions and absence variants that were variable in at least two of the
612 fourteen accessions selected to be genotyped were considered. For insertion sites, primers
613 were designed to span the predicted TE insertion site. For TE absence sites, two primer sets
614 were designed; one primer set to span the TE, and another primer set with one primer
615 annealing within the TE sequence predicted to be absent, and the other primer annealing in the
616 flanking sequence (Fig. 1C). Primer sequences were designed that did not anneal to regions of
617 the genome containing previously identified SNPs in any of the 216 accessions (Schmitz et al.
618 2013) or small insertions and deletions, identified using lumpy-sv with the default settings (Layer
619 et al. 2014)(https://github.com/arq5x/lumpy-sv), had an annealing temperature close to 52ºC

620 calculated based on nearest neighbor thermodynamics (MeltingTemp submodule in the
621 SeqUtils python module; (Cock et al. 2009)), a GC content between 40% and 60%, and
622 contained the same base repeated not more than four times in a row. Primers were aligned to
623 the TAIR10 reference genome using bowtie2 (Langmead and Salzberg 2012) with the '-a' flag
624 set to report all alignments, and those with more than 5 mapping locations in the genome were
625 then removed.
626
627 *PCR*
628 PCR was performed with 10 ng of extracted, purified Arabidopsis DNA using Taq polymerase.
629 PCR products were analyzed by agarose gel electrophoresis. Col-0 was used as a positive
630 control, water was added to reactions as a negative control.
631
632 **mRNA analysis**
633 Processed mRNA data for 144 wild Arabidopsis accessions were downloaded from NCBI GEO
634 GSE43858 (Schmitz et al. 2013). To find differential gene expression dependent on TE
635 presence/absence variation, we first filtered TE variants to include only those where the TE
636 variant was shared by at least 5 accessions with RNA data available, corresponding to a minor
637 allele frequency above 3%. We then grouped accessions based on TE presence/absence
638 variants, and performed a Mann-Whitney U test to determine differences in RNA transcript
639 abundance levels between the groups. We used q-value estimation to correct for multiple
640 testing, using the R qvalue package v2.2.2 with the following parameters: lambda = seq(0, 0.6,
641 0.05), smooth.df = 4 (Storey and Tibshirani 2003). Genes were defined as differentially
642 expressed where there was a greater than 2 fold difference in expression between the groups,
643 with a q-value less than 0.001. Gene ontology enrichment analysis was performed using DAVID
644 (https://david.ncifcrf.gov/).

645

646 **DNA methylation data analysis**

647 Processed DNA methylation data for wild Arabidopsis accessions were downloaded from NCBI
648 GEO GSE43857 (Schmitz et al. 2013). Weighted embryo DNA methylation data in 300 bp
649 windows for Col-0 crosses with Cvi 6 days after pollination were downloaded from the Dryad
650 Digital Repository; http://dx.doi.org/10.5061/dryad.gv536.2 (Gehring et al. 2014). Processed
651 embryo DNA methylation data in 50 bp windows for crosses between Col-0 and L*er* 7-8 days
652 after pollination were downloaded from NCBI GEO GSE38935 (Ibarra et al. 2012).

653

654 **Small RNA data analysis**

655 Normalized 21-24 nt smRNA read counts (reads per million reads mapped; RPM) in 300 bp
656 windows for whole seed 6 days after pollination, for reciprocal crosses between Col-0, Cvi, and
657 L*er* were downloaded from the Dryad Digital Repository;
658 http://dx.doi.org/10.5061/dryad.gv536.2 (Gehring et al. 2014).

659

660 ## DATA ACCESS

661 TEPID source code can be accessed at https://github.com/ListerLab/TEPID. L*er* TE variants are
662 available in Supplementary File 1. TE variants identified among the 216 wild Arabidopsis
663 accessions resequenced by Schmitz et al. (2013) are available in Supplementary File 2.

664

665 ## ACKNOWLEDGMENTS

666 This work was supported by the Australian Research Council (ARC) Centre of Excellence
667 program in Plant Energy Biology CE140100008 (J.B., R.L.). R.L. was supported by an ARC
668 Future Fellowship (FT120100862) and Sylvia and Charles Viertel Senior Medical Research

669 Fellowship, and work in the laboratory of R.L. was funded by the Australian Research Council.
670 T.S. was supported by the Jean Rogerson Postgraduate Scholarship. S.R.E. was supported by
671 an Australian Research Council Discovery Early Career Research Award (DE150101206). We
672 thank Robert J. Schmitz, Mathew G. Lewsey, Ronan C. O'Malley, and Ian Small for their critical
673 reading of the manuscript. We also thank Kevin Murray for his helpful comments regarding the
674 development of TEPID.
675

676 AUTHOR CONTRIBUTIONS

677 R.L. and T.S. designed the research project. R.L. and J.B. supervised research. T.S. developed
678 and tested TEPID. J.C. performed PCR validations of TE variants. T.S. and S.R.E. performed
679 bioinformatic analysis. R.L., T.S., J.B. and S.R.E. prepared the manuscript.
680

681 COMPETING FINANCIAL INTERESTS

682 The authors declare no competing financial interests.

683 # REFERENCES

684 Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang
685 C, et al. 2005. Genome-Wide Association Mapping in Arabidopsis Identifies Previously
686 Known Flowering Time and Pathogen Resistance Genes. *PLoS Genetics* **1**: e60–9.

687 Bartenhagen C, Dugas M. 2013. RSVSim: an R/Bioconductor package for the simulation of
688 structural variations. *Bioinformatics* **29**: 1679–1681.

689 Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of P-M hybrid dysgenesis: the
690 role of the P element, a P-strain-specific transposon family. *Cell* **29**: 995–1004.

691 Blumenstiel JP, Hartl DL. 2005. Evidence for maternally transmitted small interfering RNA in the
692 repression of transposition in Drosophila virilis. *Proc Natl Acad Sci USA* **102**: 15965–15970.

693 Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S,
694 Sørensen I, Lichtenstein G, et al. 2014. The genome of the stress-tolerant wild tomato
695 species. *Nat Genet* **46**: 1034–1038.

696 Bundock P, Hooykaas P. 2005. An Arabidopsis hAT-like transposase is essential for plant
697 development. *Nature* **436**: 282–284.

698 Calarco JP, Borges F, Donoghue MTA, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F,
699 FeijO JA, Becker JD, et al. 2012. Reprogramming of DNA Methylation in Pollen Guides
700 Epigenetic Inheritance via Small RNA. *Cell* **151**: 194–205.

701 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
702 BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

703 Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O,

704 Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana
705 populations. *Nat Genet* **43**: 956–963.

706 Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A,
707 Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies
708 from long-read SMRT sequencing data. *Nat Meth* **10**: 563–569.

709 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff
710 F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational
711 molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

712 Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF,
713 Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis
714 genome reveals DNA methylation patterning. *Nature* **452**: 215–219.

715 Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating
716 genomic datasets and annotations. *Bioinformatics* **27**: 3423–3424.

717 Edwards K, Johnstone C, Thompson C. 1991. A simple and rapid method for the preparation of
718 plant genomic DNA for PCR analysis. *Nucleic Acids Research* **19**: 1349.

719 Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read
720 extraction. *Bioinformatics* **30**: 2503–2505.

721 Faust GG, Hall IM. 2012. YAHA: fast and flexible long-read alignment with optimal breakpoint
722 detection. *Bioinformatics* **28**: 2417–2424.

723 Gehring M, Pignatta D, Erdmann RM, Bell GW, Scheer E. 2014. *Data from: Natural epigenetic*
724 *polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting*. Dryad Digital

725 Repository.

726 Gendrel AV, Lippman Z, Yordan C, Colot V, Martienssen RA. 2002. Dependence of
727 heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1. *Science*
728 **297**: 1871–1873.

729 Hénaff E, Vives C, Desvoyes B, Chaurasia A, Payet J, Gutierrez C, Casacuberta JM. 2014.
730 Extensive amplification of the E2F transcription factor binding sites by transposons during
731 evolution of Brassica species. *The Plant Journal* **77**: 852–862.

732 Hénaff E, Zapata L, Casacuberta JM, Ossowski S. 2015. Jitterbug: somatic and germline
733 transposon insertion detection at single-nucleotide resolution. *BMC Genomics* **16**: 1–16.

734 Hirochika H, Okamoto H, Kakutani T. 2000. Silencing of retrotransposons in Arabidopsis and
735 reactivation by the ddm1 mutation. *The Plant Cell* **12**: 357–369.

736 Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. 2009.
737 Genome-Wide Demethylation of Arabidopsis Endosperm. *Science* **324**: 1451–1454.

738 Ibarra CA, Feng X, Schoft VK, Hsieh T-F, Uzawa R, Rodrigues JA, Zemach A, Chumak N,
739 Machlicova A, Nishimura T, et al. 2012. Active DNA demethylation in plant companion cells
740 reinforces transposon methylation in gametes. *Science* **337**: 1360–1364.

741 Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. 2011. An siRNA pathway
742 prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472**:
743 115–119.

744 Jeddeloh JA, Stokes TL, Richards EJ. 1999. Maintenance of genomic methylation requires a
745 SWI2/SNF2-like protein. *Nat Genet* **22**: 94–97.

746 Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. 2014. Environmentally
747 responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and
748 epimutations. *Genome Research* **24**: 1821–1829.

749 Josefsson C, Dilkes B, Comai L. 2006. Parent-dependent loss of gene silencing during
750 interspecies hybridization. *Current Biology* **16**: 1322–1328.

751 Jullien PE, Susaki D, Yelagandula R, Higashiyama T, Berger F. 2012. DNA Methylation
752 Dynamics during Sexual Reproduction in Arabidopsis thaliana. *Curr Biol* **22**: 1825–1830.

753 Katoh A, Uenohara K, Akita M, Hashimoto T. 2006. Early steps in the biosynthesis of NAD in
754 Arabidopsis start with aspartate and occur in the plastid. *Plant Physiology* **141**: 851–857.

755 Kawashima T, Berger F. 2014. Epigenetic reprogramming in plant sexual reproduction. *Nature*
756 *Reviews Genetics* **15**: 613–624.

757 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**: 357–
758 359.

759 Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for
760 structural variant discovery. *Genome Biology* **15**: R84.

761 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
762 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
763 format and SAMtools. *Bioinformatics* **25**: 2078–2079.

764 Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V,
765 May B, Kasschau KD, et al. 2004. Role of transposable elements in heterochromatin and
766 epigenetic control. *Nature* **430**: 471–476.

767 Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008.
768 Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**:
769 523–536.

770 Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM. 2015.
771 Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic
772 Stress. *PLoS Genetics* **11**: e1004915.

773 Martienssen RA. 2010. Heterochromatin, small RNA and post-fertilization dysgenesis in
774 allopolyploid and interploid hybrids of Arabidopsis. *New Phytologist* **186**: 46–53.

775 Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of
776 increasing complexity. *Nature Reviews Genetics* **15**: 394–408.

777 Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D,
778 Paszkowski J, Mathieu O. 2009. Selective epigenetic control of retrotransposition in
779 Arabidopsis. *Nature* **461**: 427–430.

780 Miura A, Yonebayashi S, Watanabe K, Toyama T. 2001. Mobilization of transposons by a
781 mutation abolishing full DNA methylation in Arabidopsis. *Nature* **411**: 212–214.

782 Mosher RA, Melnyk CW, Kelly KA, Dunn RM, Studholme DJ, Baulcombe DC. 2009. Uniparental
783 expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. *Nature*
784 **460**: 283–286.

785 Pignatta D, Erdmann RM, Scheer E, Picard CL, Bell GW, Gehring M. 2014. Natural epigenetic
786 polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *eLife* **3**:
787 e03198.

788 Plasterk R. 1991. The Origin of Footprints of the Tc1 Transposon of Caenorhabditis elegans.
789 *EMBO* **10**: 1919–1925.

790 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
791 features. *Bioinformatics* **26**: 841–842.

792 Richards EJ. 2006. Inherited epigenetic variation - revisiting soft inheritance. *Nature Reviews*
793 *Genetics* **7**: 395–U2.

794 Robb S, Lu L, Valencia E, Burnette JM. 2013. The use of RelocaTE and unassembled short
795 reads to produce high-resolution snapshots of transposable element generated diversity in
796 rice. *G3: Genes*| *Genomes*| *Genetics*.

797 Saze H, Scheid OM, Paszkowski J. 2003. Maintenance of CpG methylation is essential for
798 epigenetic inheritance during plant gametogenesis. *Nat Genet* **34**: 65–69.

799 Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen
800 H, Schork NJ, et al. 2013. Patterns of population epigenomic diversity. *Nature* **495**: 193–
801 198.

802 Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J,
803 Warthmann N, et al. 2011. Reference-guided assembly of four diverse Arabidopsis thaliana
804 genomes. *Proc Natl Acad Sci USA* **108**: 10249–10254.

805 Slotkin RK, Vaughn M, Borges F, TanurdZiC M, Becker JD, FeijO JA, Martienssen RA. 2009.
806 Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*
807 **136**: 461–472.

808 Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad*

809 *Sci USA* **100**: 9440–9445.

810 Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive
811 analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell*
812 **152**: 352–364.

813 Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman
814 RE, Neph S, Reynolds AP, et al. 2014. Mapping and dynamics of regulatory DNA and
815 transcription factor networks in A. thaliana. *Cell* **8**: 2015–2030.

816 Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K,
817 Veltman JA, Hehir-Kwa JY. 2014. Mobster: accurate detection of mobile element insertions
818 in next generation sequencing data. 1–11.

819 Vitte C, Fustier M-A, Alix K, Tenaillon MI. 2014. The bright side of transposons in crop evolution.
820 *Briefings in Functional Genomics* **13**: 276–295.

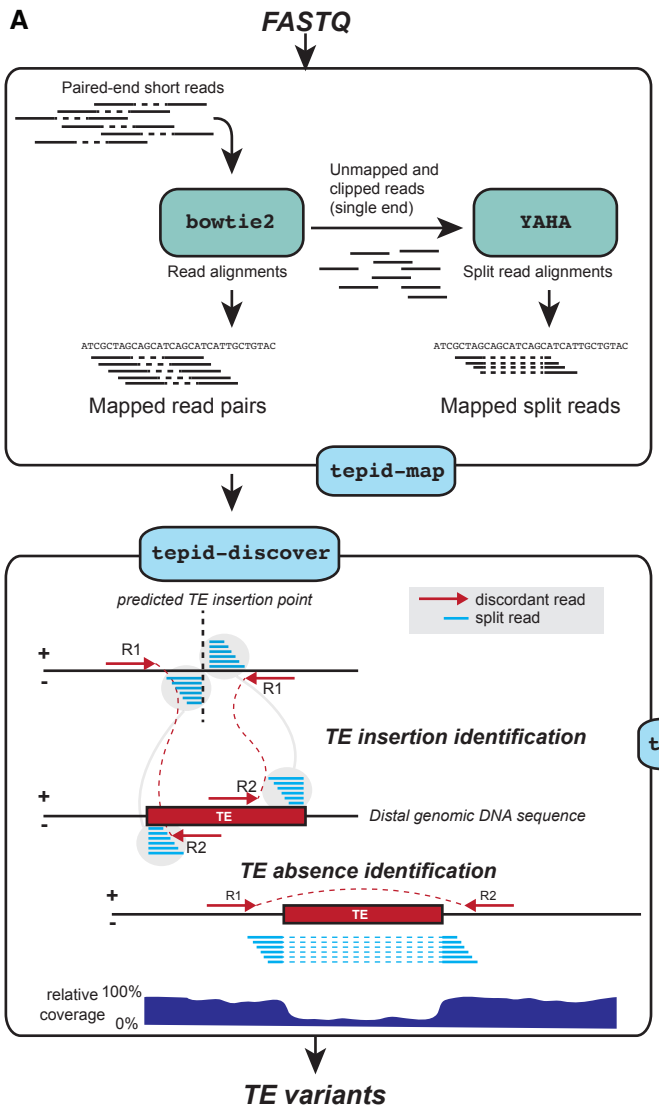
821 Wang G, Ellendorff U, Kemp B, Mansfield JW, Forsyth A, Mitchell K, Bastas K, Liu C-M, Woods-
822 Tör A, Zipfel C, et al. 2008. A genome-wide functional investigation into the roles of
823 receptor-like proteins in Arabidopsis. *Plant Physiology* **147**: 503–517.

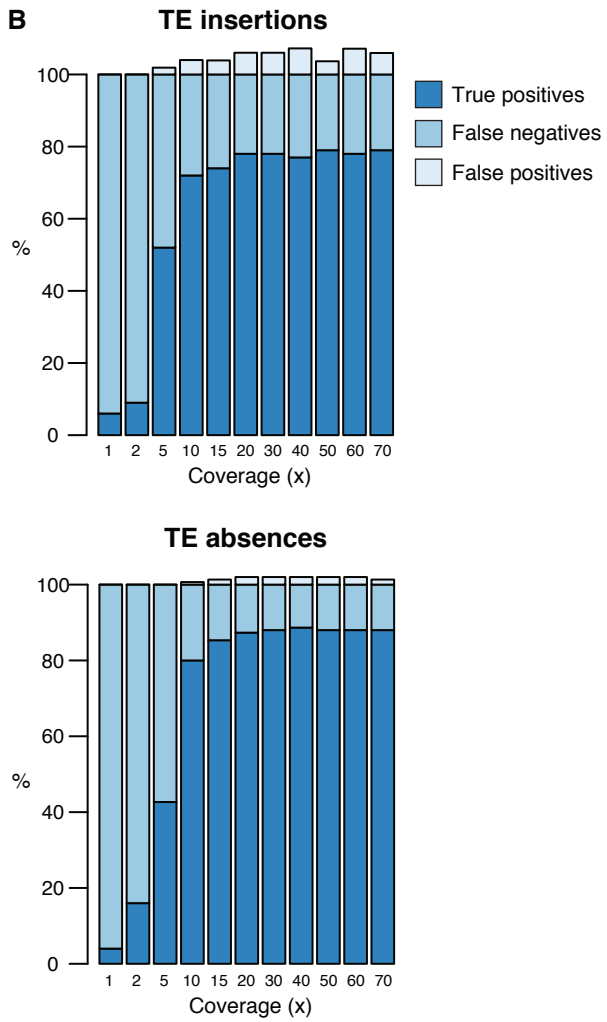
824 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
825 Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic
826 transposable elements. *Nature Reviews Genetics* **8**: 973–982.

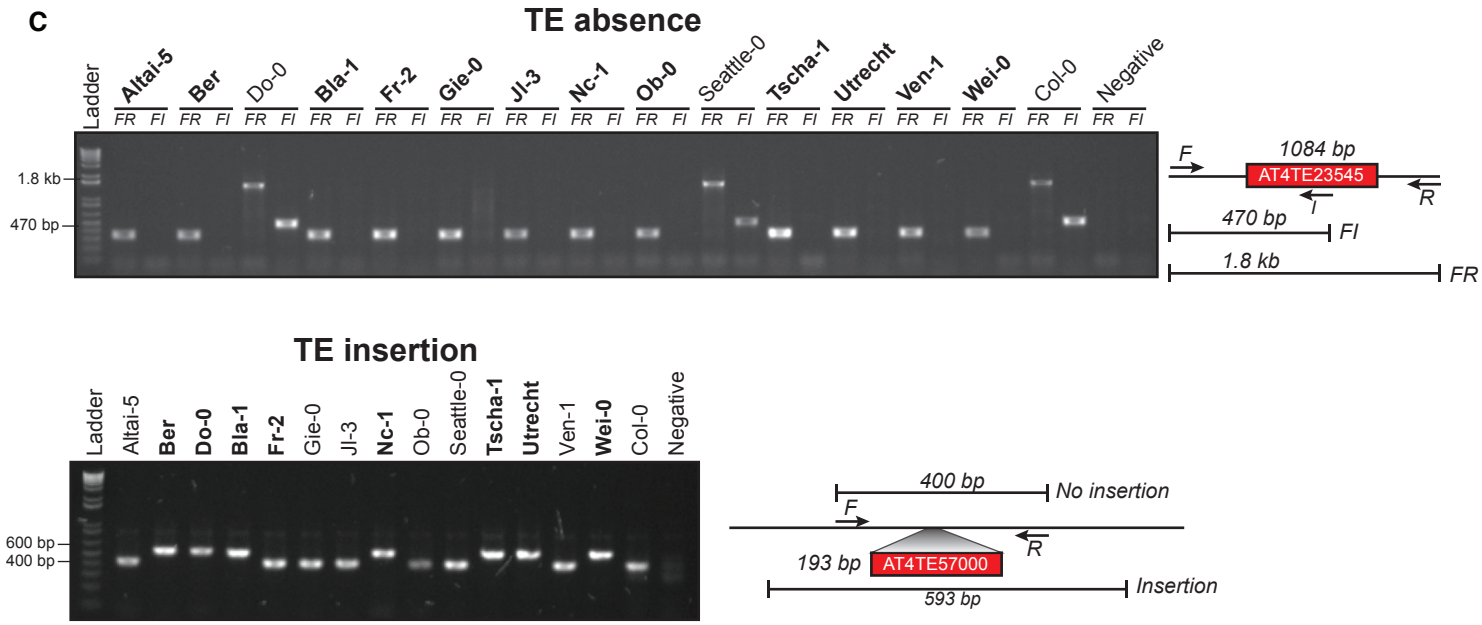
827 Zemach A, Kim MY, Hsieh P-H, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL,
828 Zilberman D. 2013. The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA
829 Methyltransferases to Access H1-Containing Heterochromatin. *Cell* **153**: 193–205.

830 Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen H, Henderson IR, Shinn P,
831 Pellegrini M, Jacobsen SE, et al. 2006. Genome-wide High-Resolution Mapping and
832 Functional Analysis of DNA Methylation in Arabidopsis. *Cell* **126**: 1189–1201.

833 Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of
834 Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation
835 and transcription. *Nat Genet* **39**: 61–69.







D

	True positives	False positives	False negatives
Absences	77.10%	9.16%	13.74%
Insertions	70.71%	9.29%	20.00%

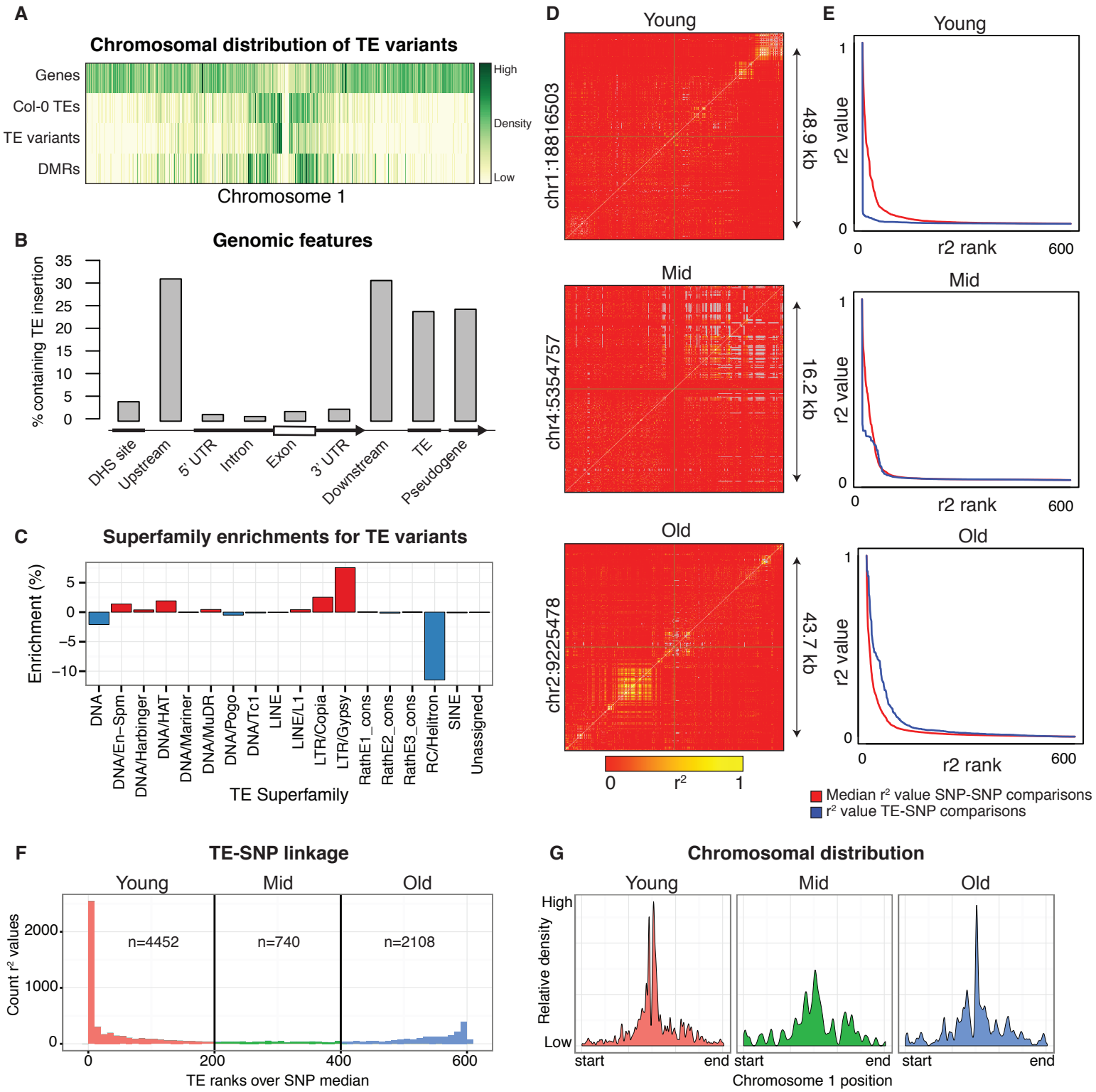
837 **Figure 1. Design and testing of the TE variant discovery pipeline**

838 (A) Principle of TE variant discovery using split and discordant read mapping positions. Paired
839 end reads are first mapped to the reference genome using Bowtie2 (Langmead and Salzberg
840 2012). Soft-clipped or unmapped reads are then extracted from the alignment and re-mapped
841 using Yaha, a split read mapper (Faust and Hall 2012). All read alignments are then used by
842 TEPID to discover TE variants relative to the reference genome, in the 'tepid-discover' step.
843 When analyzing groups of related samples, these variants can be further refined using the
844 'tepid-refine' step, which examines in more detail the genomic regions where there was a TE
845 variant identified in another sample, and calls the same variant for the sample in question using
846 lower read count thresholds as compared to the 'tepid-discover' step, in order to reduce false
847 negative variant calls within a group of related samples.

848 (B) Testing of the TEPID pipeline using simulated TE variants in the Arabidopsis Col-0 genome
849 (TAIR10), for a range of sequencing coverage levels.

850 (C) PCR validation examples for TE presence/absence variant predictions. Accessions that
851 were predicted to contain a TE insertion or TE absence are marked in bold. For the example TE
852 absence variant, two primer sets were used; forward (F) and reverse (R) or internal (I).
853 Accessions with a TE absence will not produce the FI band and produce a shorter FR product,
854 with the change in size matching the size of the deleted TE. For the example TE insertion, one
855 primer set was used, spanning the TE insertion site. A band shift of approximately 200 bp can
856 be seen, corresponding to the size of the inserted TE.

857 (D) Summary table of PCR validation results for TE variants.


A
Chromosomal distribution of TE variants
Genes
Col-0 TEs
TE variants
DMRs
High
Density
Low
Chromosome 1
B
Genomic features
% containing TE insertion
35
30
25
20
15
10
5
0
DHS site
Upstream
5' UTR
Intron
Exon
3' UTR
Downstream
TE
Pseudogene
C
Superfamily enrichments for TE variants
Enrichment (%)
5
0
−5
−10
DNA
DNA/En−Spm
DNA/Harbinger
DNA/HAT
DNA/Mariner
DNA/MuDR
DNA/Pogo
DNA/Tc1
LINE
LINE/L1
LTR/Copia
LTR/Gypsy
RathE1_cons
RathE2_cons
RathE3_cons
RC/Helitron
SINE
Unassigned
TE Superfamily
D
Young
chr1:18816503
48.9 kb
Mid
chr4:5354757
16.2 kb
Old
chr2:9225478
43.7 kb
0
r²
1
E
Young
Mid
Old
r2 value
r2 rank
0
600
1
Median r² value SNP-SNP comparisons
r² value TE-SNP comparisons
F
TE-SNP linkage
Young
Mid
Old
Count r² values
2000
1000
0
n=4452
n=740
n=2108
0
200
400
600
TE ranks over SNP median
G
Chromosomal distribution
Young
Mid
Old
High
Relative density
Low
start
end
Chromosome 1 position

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

860 (A) Distribution of identified TE variants on chromosome 1, with distributions of all Col-0 genes,
861 Col-0 TEs, and population DMRs.

862 (B) Frequency of TE variants at different genomic features.

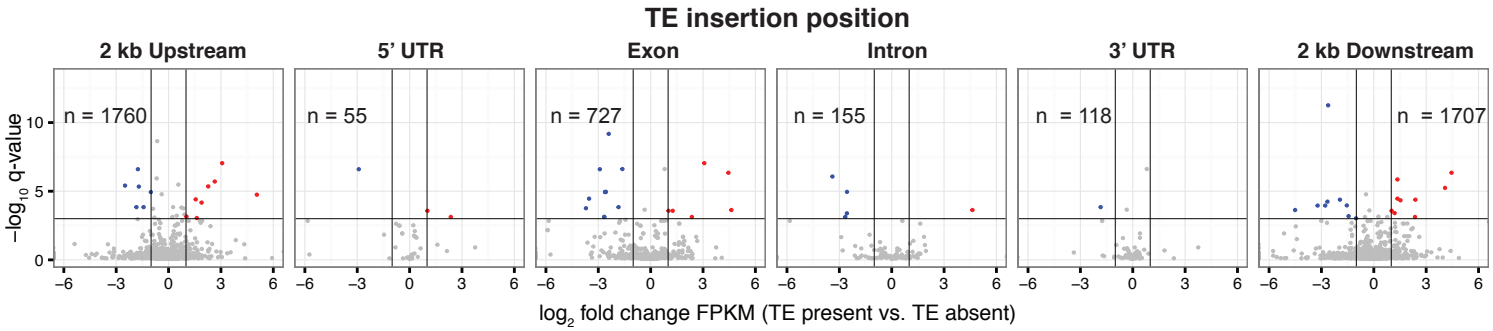
863 (C) Enrichment and depletion of TE variants categorized by TE superfamily compared to the
864 expected frequency due to genomic occurrence.

865 (D) r^2 correlation matrices for individual representative young, mid, and old TE variants.

866 (E) Rank order plots for individual representative young, mid, and old TE variants (matching
867 those shown in D). Red line indicates the median r^2 value for each rank across SNP-based
868 values. Blue line indicates r^2 values for TE-SNP comparisons.

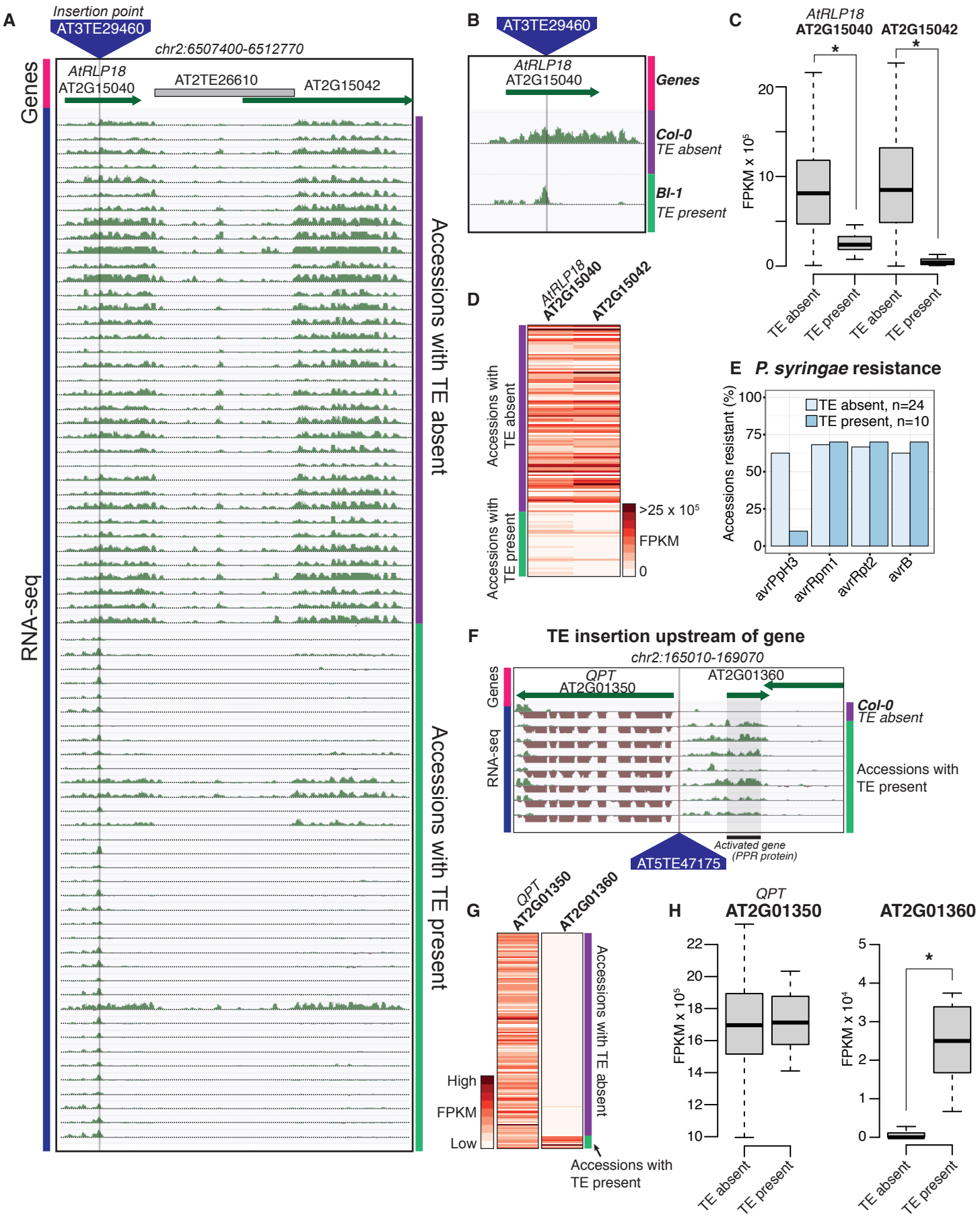
869 (F) Histogram of the number of TE r^2 ranks (0-600) that are above the SNP-based median r^2
870 value for testable TE variants. TE variants with <200 ranks over the SNP median were classified
871 as "young" elements, as they are not yet linked to the surrounding SNPs. TE variants with 200-
872 400 ranks over the SNP median were classified as "mid" aged. TE variants with >400 ranks
873 over the SNP median were classified as "old" elements, as they were linked to the surrounding
874 SNPs.

875 (G) Chromosomal distribution of TE variants by age.


TE insertion position
2 kb Upstream
5' UTR
Exon
Intron
3' UTR
2 kb Downstream
n = 1760
n = 55
n = 727
n = 155
n = 118
n = 1707
−log10 q-value
log2 fold change FPKM (TE present vs. TE absent)

877 **Figure 3. Differential transcript abundance associated with genic TE variant**
878 **presence/absence**

879 Volcano plots showing transcript abundance differences for genes associated with TE insertion
880 variants at different positions, indicated in the plot titles. Genes with significantly different
881 transcript abundance in accessions with a TE insertion compared to accessions without a TE
882 insertion are colored blue (lower transcript abundance in accessions containing TE insertion) or
883 red (higher transcript abundance in accessions containing TE insertion). Vertical lines indicate
884 ±2 fold change in FPKM.


A
Insertion point
AT3TE29460
chr2:6507400-6512770
Genes
AtRLP18
AT2G15040
AT2TE26610
AT2G15042
RNA-seq
Accessions with TE absent
Accessions with TE present
B
AT3TE29460
AtRLP18
AT2G15040
Genes
Col-0
TE absent
Bl-1
TE present
C
AtRLP18
AT2G15040
AT2G15042
FPKM x 10^5
TE absent
TE present
TE absent
TE present
D
AtRLP18
AT2G15040
AT2G15042
Accessions with TE absent
Accessions with TE present
>25 x 10^5
FPKM
0
E
P. syringae resistance
Accessions resistant (%)
TE absent, n=24
TE present, n=10
avrPphB
avrRpm1
avrRpt2
avrB
F
TE insertion upstream of gene
chr2:165010-169070
Genes
RNA-seq
QPT
AT2G01350
AT2G01360
Col-0
TE absent
Accessions with
TE present
Activated gene
(PPR protein)
AT5TE47175
G
QPT
AT2G01350
AT2G01360
High
FPKM
Low
Accessions with TE absent
Accessions with
TE present
H
QPT
AT2G01350
AT2G01360
FPKM x 10^5
FPKM x 10^4
TE absent
TE present
TE absent
TE present

886 **Figure 4. Effects of TE variants on local gene expression**

887 (A) Genome browser representation of RNA-seq data for genes *AtRLP18* (AT2G15040) and a
888 leucine-rich repeat family protein (AT2G15042) for all accessions containing a TE insertion
889 within the exon of the gene *AtRLP18*, and for a random subset of accessions not containing the
890 TE insertion within the exon of *AtRLP18*.

891 (B) Magnified view of the TE insertion variant within the *AtRLP18* exon.

892 (C) Boxplots comparing transcript abundance of *AtRLP18* and AT2G15042 in accessions with
893 and without the TE insertion in the *AtRLP18* exon. Asterisk indicates statistical significance ($p <$
894 5.8×10^{-10}; Mann-Whitney U test). Boxes represent the interquartile range (IQR) from quartile 1
895 to quartile 3. Boxplot upper whiskers represent the maximum value, or the upper value of the
896 quartile 3 plus 1.5 times the IQR (whichever is smaller). Boxplot lower whisker represents the
897 minimum value, or the lower value of the quartile 1 minus 1.5 times the IQR (whichever is
898 larger).

899 (D) Heatmap showing *AtRLP18* and AT2G15042 RNA-seq FPKM values for all accessions.

900 (E) Percentage of accessions with resistance to *Pseudomonas syringae* transformed with
901 different *avr* genes, for accessions containing or not containing a TE insertion in *AtRLP18.*

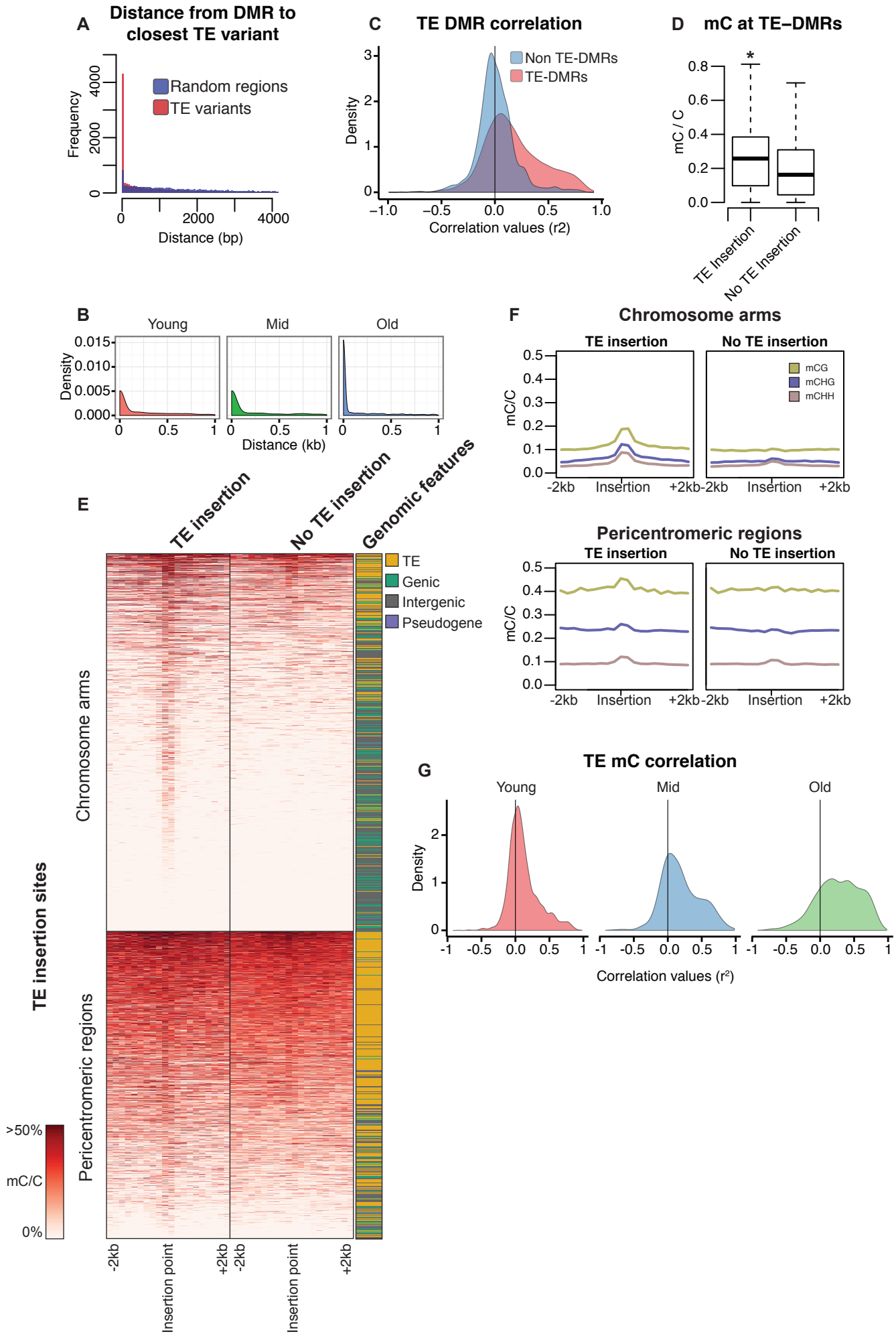
902 (F) Genome browser representation of RNA-seq data for a PPR protein-encoding gene
903 (AT2G01360) and *QPT* (AT2G01350), showing transcript abundance for these genes in
904 accessions containing a TE insertion variant in the upstream region of these genes.

905 (G) Heatmap representation of RNA-seq FPKM values for *QPT* and a gene encoding a PPR
906 protein (AT2G01360), for all accessions. Note that scales are different for the two heatmaps,
907 due to the higher transcript abundance of *QPT* compared to AT2G01360. Scale maximum for
908 AT2G01350 is 3.1×10^{5}, and for AT2G01360 is 5.9×10^{4}.

909 (H) Boxplots showing RNA-seq FPKM differences for *QPT* and AT2G01360 associated with

910 presence/absence of a TE variant in the gene upstream region. Asterisk indicates statistical

911 significance ($p < 1.8 \times 10^{-7}$, Mann-Whitney U test). Boxplots were constructed as for C.


A
Distance from DMR to closest TE variant
Random regions
TE variants
Frequency
Distance (bp)
C
TE DMR correlation
Non TE-DMRs
TE-DMRs
Density
Correlation values (r2)
D
mC at TE–DMRs
mC / C
TE Insertion
No TE Insertion
B
Young
Mid
Old
Density
Distance (kb)
F
Chromosome arms
TE insertion
No TE insertion
mCG
mCHG
mCHH
mC/C
-2kb
Insertion
+2kb
Pericentromeric regions
E
TE insertion
No TE insertion
Genomic features
TE
Genic
Intergenic
Pseudogene
Chromosome arms
Pericentromeric regions
TE insertion sites
>50%
mC/C
0%
Insertion point
G
TE mC correlation
Young
Mid
Old
Density
Correlation values (r²)

913 **Figure 5. TE variants are associated with DNA methylation changes**

914 (A) Distance from each population DMR to closest TE variant (red), compared with a set of
915 randomly selected regions of the genome, of the same size as all TE variants (blue).

916 (B) Density of distance measurements between population DMRs and the closest TE variant,
917 grouped by age classification of TE variant.

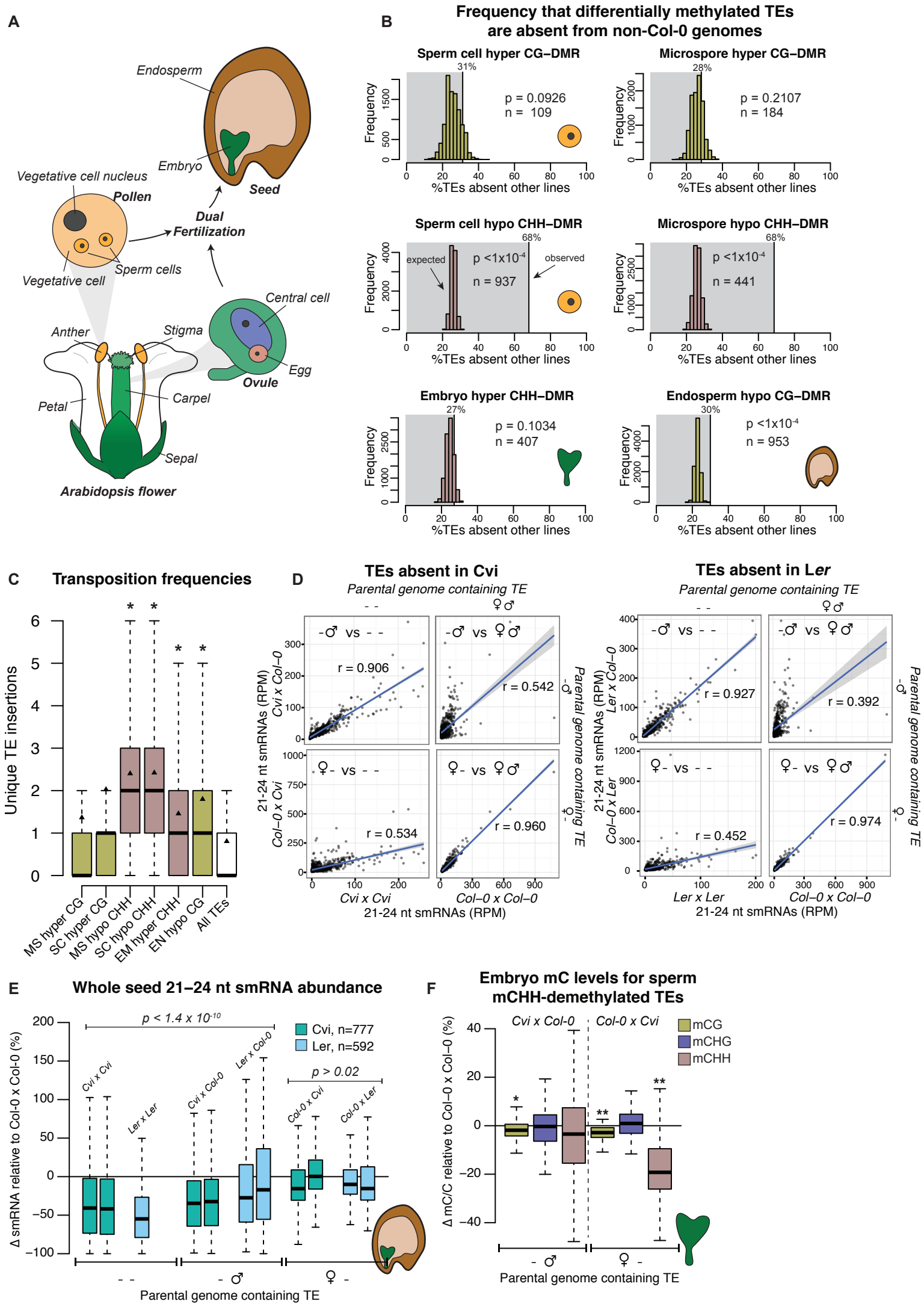
918 (C) Density of Pearson correlation values between DNA methylation levels at population DMRs
919 and the presence/absence of the nearest TE variants. Pearson correlation values for population
920 DMRs within 1 kb of a TE variant (TE-DMRs) are shown in red, while values for population
921 DMRs further than 1 kb from a TE variant (non TE-DMRs) are shown in blue.

922 (D) Boxplot showing DNA methylation levels at TE-DMRs for accessions containing the TE
923 insertion and accessions not containing the TE insertion. Asterisk indicates statistical
924 significance ($p < 2.2 \times 10^{-16}$, Welch's t-test). Boxplots were constructed as for Fig. 4C.

925 (E) DNA methylation levels in 200 bp windows ± 2 kb from TE insertion sites, for accessions
926 with and without the TE insertion. Heatmap is separated into in regions of the genome less than
927 3 Mb from a centromere (pericentromeric regions) or greater than 3 Mb from a centromere
928 (chromosome arms), and sorted by total DNA methylation at each locus.

929 (F) Average DNA methylation levels in mCG, mCHG and mCHH contexts in regions
930 surrounding TE insertion variant sites for pericentromeric insertions and insertions in the
931 euchromatic chromosome arms.

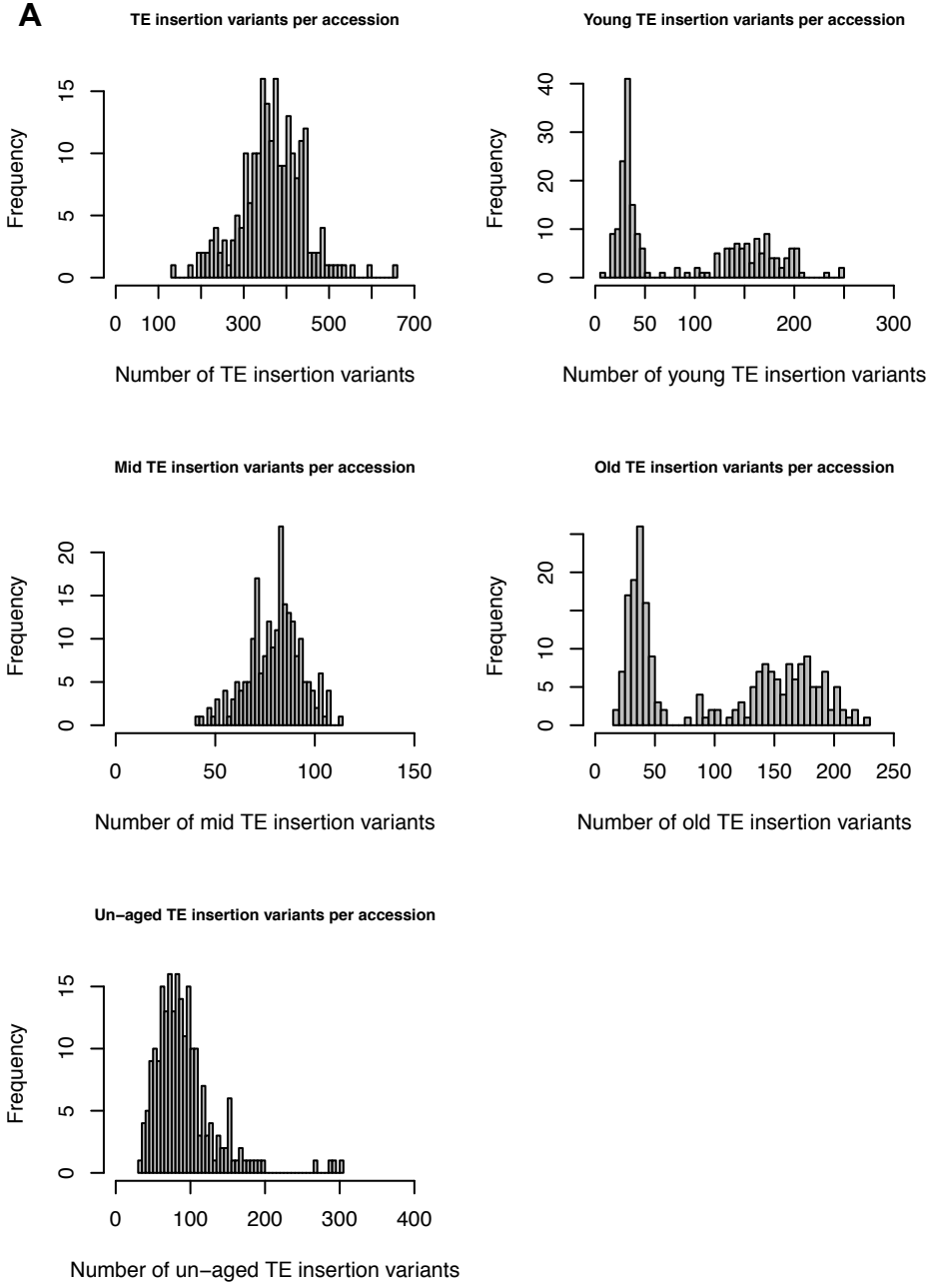
932 (G) Density of Pearson correlation values between DNA methylation levels in 200 bp regions
933 flanking TE variant sites and the presence/absence of the TE variant, separated by TE age
934 groups. Only TEs that were able to be assigned an age group are shown (total 7,300; minor
935 allele frequency >3%).


A
Endosperm
Embryo
Seed
Vegetative cell nucleus
Pollen
Dual Fertilization
Sperm cells
Vegetative cell
Central cell
Anther
Stigma
Egg
Ovule
Petal
Carpel
Sepal
Arabidopsis flower
B
Frequency that differentially methylated TEs are absent from non-Col-0 genomes
Sperm cell hyper CG–DMR
31%
p = 0.0926
n = 109
Microspore hyper CG–DMR
28%
p = 0.2107
n = 184
Sperm cell hypo CHH–DMR
68%
expected
observed
p <1x10-4
n = 937
Microspore hypo CHH–DMR
68%
p <1x10-4
n = 441
Embryo hyper CHH–DMR
27%
p = 0.1034
n = 407
Endosperm hypo CG–DMR
30%
p <1x10-4
n = 953
Frequency
%TEs absent other lines
C
Transposition frequencies
Unique TE insertions
MS hyper CG
SC hyper CG
MS hypo CHH
SC hypo CHH
EM hyper CHH
EN hypo CG
All TEs
D
TEs absent in Cvi
Parental genome containing TE
r = 0.906
r = 0.542
r = 0.534
r = 0.960
21-24 nt smRNAs (RPM)
Cvi x Cvi
Col–0 x Col–0
TEs absent in Ler
r = 0.927
r = 0.392
r = 0.452
r = 0.974
Ler x Ler
E
Whole seed 21–24 nt smRNA abundance
p < 1.4 x 10-10
Cvi, n=777
Ler, n=592
p > 0.02
Cvi x Cvi
Ler x Ler
Cvi x Col-0
Ler x Col-0
Col-0 x Cvi
Col-0 x Ler
Δ smRNA relative to Col-0 x Col-0 (%)
Parental genome containing TE
F
Embryo mC levels for sperm mCHH-demethylated TEs
Cvi x Col-0
Col-0 x Cvi
mCG
mCHG
mCHH
Δ mC/C relative to Col–0 x Col–0 (%)
Parental genome containing TE

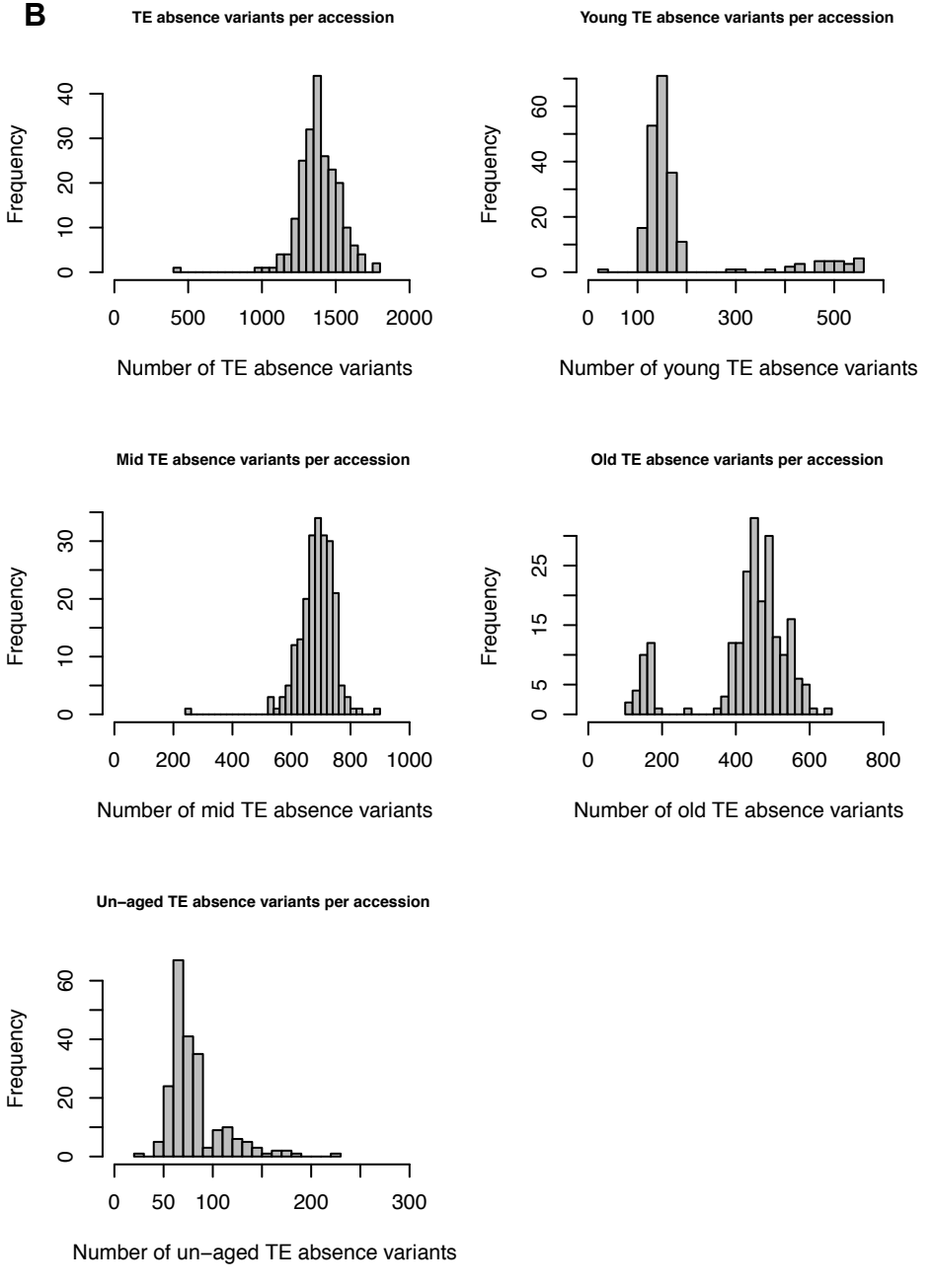
937 **Figure 6. Genetic differences between parents is associated with demethylation of**
938 **transposable elements in hybrid progeny**
939 (A) Diagram of the developing pollen, ovule, and seed. During dual fertilization, one sperm
940 fertilizes the egg forming the embryo, while the other fertilizes the central cell to form the
941 endosperm.
942 (B) Percentage of TEs differentially methylated in germline cell types that were absent in at least
943 one non-Col-0 accession (shaded region) and expected percentage (histogram). 68% of TEs
944 that lose mCHH in the sperm cell or microspore are absent from non-reference accessions,
945 significantly more often than expected ($p < 1 \times 10^{-4}$, determined by resampling 10,000 times).
946 Embryo DMRs are in comparison to aerial tissues, endosperm DMRs are in comparison to the
947 embryo.
948 (C) Frequency of transposition for TEs differentially methylated in the microspore (MS), sperm
949 cell (SC), embryo (EM) and endosperm (EN). Black triangles represent the mean number of
950 unique TE insertion sites caused by elements in each list, asterisk indicates significance ($p < 1 \times$
951 10^{-5}, Welch's t-test). Boxplots were constructed as for Fig. 4C.
952 (D) Correlation between whole seed 21-24 nt smRNA levels (normalized reads per million
953 values in 300 bp windows; RPM) for TEs present in maternal or paternal genomes only,
954 compared with smRNA levels in crosses where the TE is present or absent in both parental
955 genomes. Direction of each cross is represented on axes as *female x male*. TE variants are
956 always present in the Col-0 genome and absent in the non-Col-0 genome. R values are
957 Pearson's correlation coefficient.
958 (E) Whole seed 21-24 nt smRNA abundance (RPM, as for D) in TEs absent in Cvi for crosses
959 between Cvi and Col-0, or absent in L*er* for crosses between L*er* and Col-0. Direction of each
960 cross is represented as *female x male* on boxplot labels. Replicates are plotted side-by-side, p-

961 values are the result of a Mann-Whitney U test using averaged replicate data. Boxplots were
962 constructed as for Fig. 4C.
963 (F) Embryonic DNA methylation levels 6 days after pollination for Cvi-absent TEs that lose
964 mCHH in sperm, following reciprocal crosses between Col-0 and Cvi. Boxplots were
965 constructed as for Fig. 4C.

A


TE insertion variants per accession
Frequency
0
5
10
15
0
100
300
500
700
Number of TE insertion variants
Young TE insertion variants per accession
Frequency
0
10
20
30
40
0
50
100
200
300
Number of young TE insertion variants
Mid TE insertion variants per accession
Frequency
0
5
10
15
20
0
50
100
150
Number of mid TE insertion variants
Old TE insertion variants per accession
Frequency
0
5
10
20
0
50
100
150
200
250
Number of old TE insertion variants
Un−aged TE insertion variants per accession
Frequency
0
5
10
15
0
100
200
300
400
Number of un−aged TE insertion variants

B

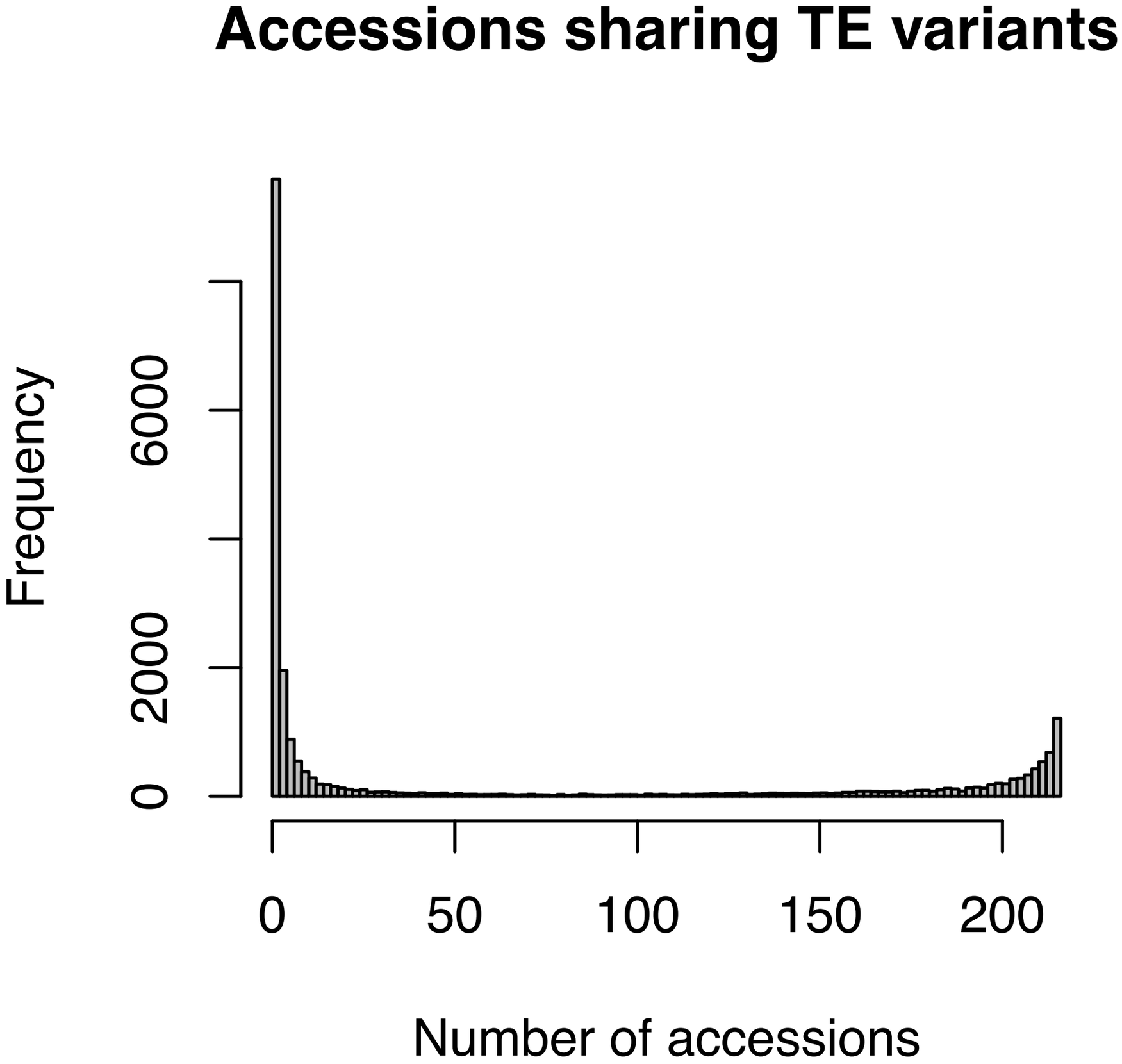

TE absence variants per accession
Frequency
0
10
20
30
40
0
500
1000
1500
2000
Number of TE absence variants
Young TE absence variants per accession
Frequency
0
20
40
60
0
100
300
500
Number of young TE absence variants
Mid TE absence variants per accession
Frequency
0
10
20
30
0
200
400
600
800
1000
Number of mid TE absence variants
Old TE absence variants per accession
Frequency
0
5
15
25
0
200
400
600
800
Number of old TE absence variants
Un−aged TE absence variants per accession
Frequency
0
20
40
60
0
50
100
200
300
Number of un−aged TE absence variants

967 **Figure S1. Number of TE presence/absence variants identified for each wild accession.**

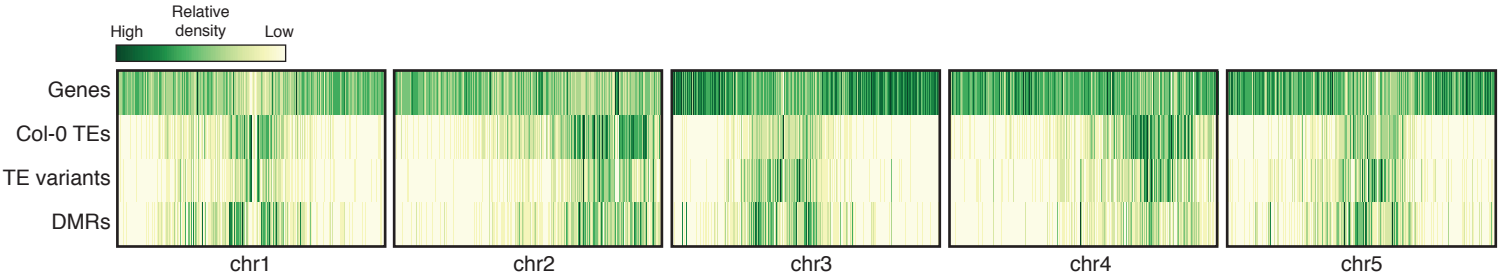
968 (A) Histograms showing number of TE insertion variants identified for each accession,

969 separated by all TE insertions, and the different TE insertion age classes.

970 (B) TE absence variants as for A.

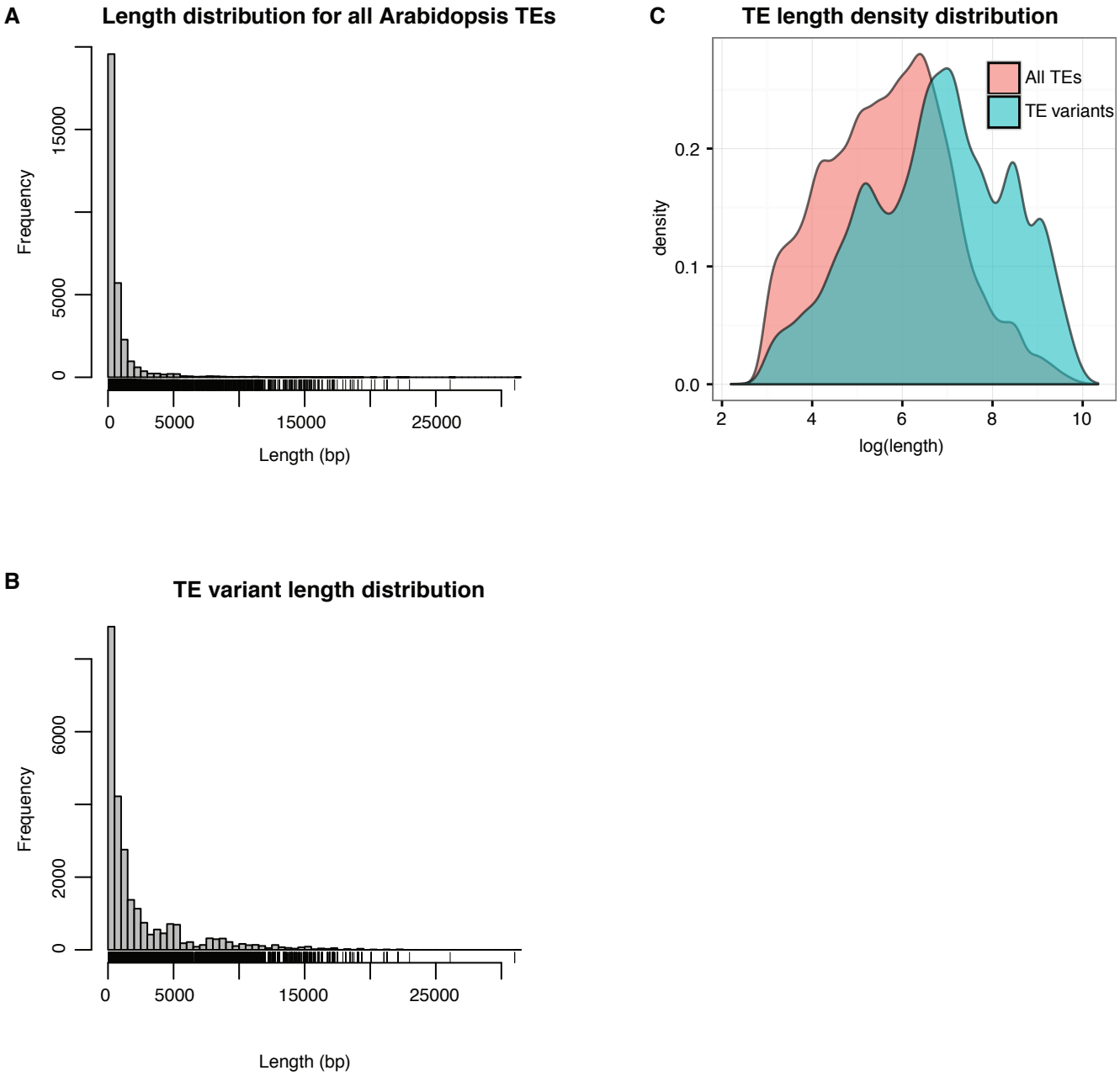

Accessions sharing TE variants
Frequency
0
2000
6000
0
50
100
150
200
Number of accessions

972 **Figure S2. Histogram showing number of accessions sharing each TE variant identified.**


Relative density
High
Low
Genes
Col-0 TEs
TE variants
DMRs
chr1
chr2
chr3
chr4
chr5

974 **Figure S3. Chromosomal distributions of genes, Col-0 TEs, TE variants and population**

975 **DMRs**

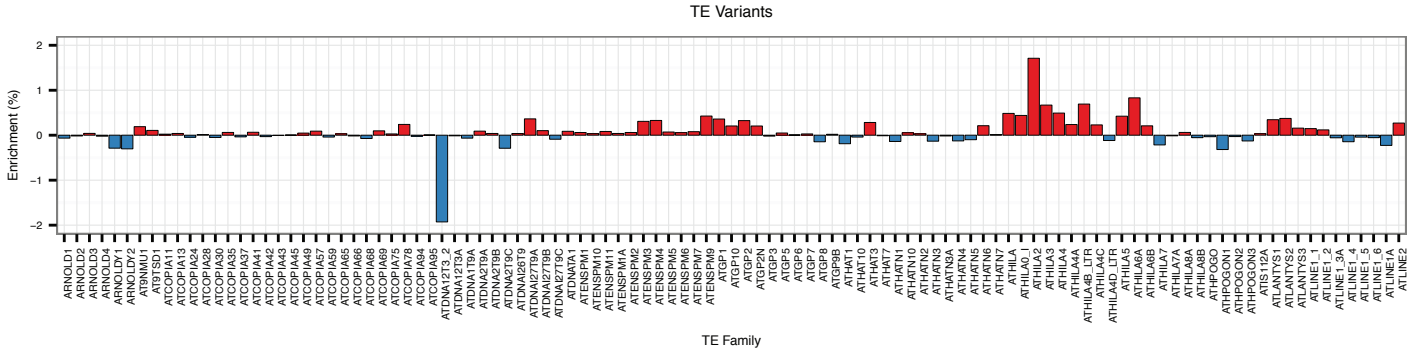

A
Length distribution for all Arabidopsis TEs
Frequency
0
5000
15000
0
5000
15000
25000
Length (bp)
C
TE length density distribution
All TEs
TE variants
density
0.0
0.1
0.2
2
4
6
8
10
log(length)
B
TE variant length distribution
Frequency
0
2000
6000
0
5000
15000
25000
Length (bp)

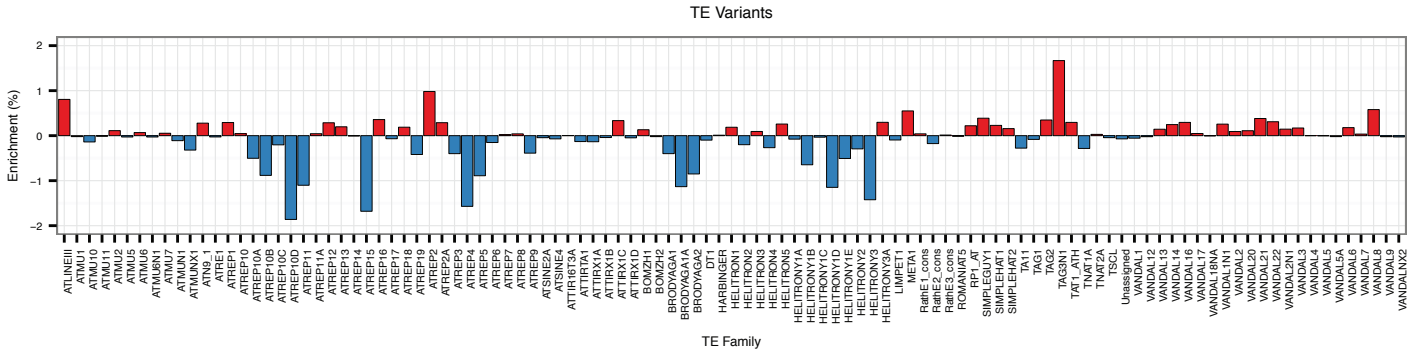
977 **Figure S4. Length distributions for all Col-0 TEs and TE variants**

978 (A) Histogram showing lengths of all annotated TEs in the Col-0 reference genome.

979 (B) Histogram showing lengths of all TE variants.

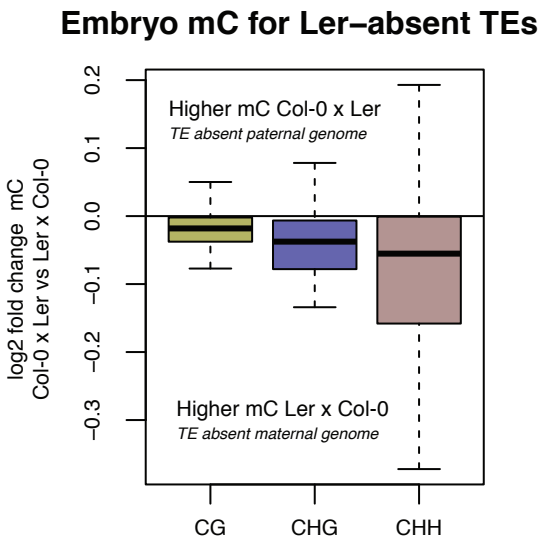
980 (C) Density distribution of $\log_{10}$ TE length for all Col-0 TEs (red) and TE variants (blue).


TE Variants
Enrichment (%)
2
1
0
−1
−2
ARNOLD1
ARNOLD2
ARNOLD3
ARNOLD4
ARNOLDY1
ARNOLDY2
AT9NMU1
AT9TSD1
ATCOPIA11
ATCOPIA13
ATCOPIA24
ATCOPIA28
ATCOPIA30
ATCOPIA35
ATCOPIA37
ATCOPIA41
ATCOPIA42
ATCOPIA43
ATCOPIA45
ATCOPIA49
ATCOPIA57
ATCOPIA59
ATCOPIA65
ATCOPIA66
ATCOPIA68
ATCOPIA69
ATCOPIA75
ATCOPIA78
ATCOPIA94
ATCOPIA95
ATDNA12T3_2
ATDNA12T3A
ATDNA1T9A
ATDNA2T9A
ATDNA2T9B
ATDNA2T9C
ATDNAI26T9
ATDNAI27T9A
ATDNAI27T9B
ATDNA27T9C
ATDNATA1
ATENSPM1
ATENSPM10
ATENSPM11
ATENSPM2
ATENSPM3
ATENSPM4
ATENSPM5
ATENSPM6
ATENSPM7
ATENSPM9
ATGP1
ATGP10
ATGP2
ATGP2N
ATGP3
ATGP5
ATGP6
ATGP7
ATGP8
ATGP9B
ATHAT1
ATHAT10
ATHAT13
ATHAT7
ATHATN1
ATHATN10
ATHATN2
ATHATN3
ATHATN3A
ATHATN4
ATHATN5
ATHATN6
ATHATN7
ATHILA
ATHILAO_I
ATHILA2
ATHILA3
ATHILA4
ATHILA4A
ATHILA4B_LTR
ATHILA4C
ATHILA4D_LTR
ATHILA5
ATHILA6A
ATHILA6B
ATHILA7
ATHILA7A
ATHILA8A
ATHILA8B
ATHPOGO
ATHPOGON1
ATHPOGON2
ATHPOGON3
ATIS112A
ATLANTYS1
ATLANTYS2
ATLANTYS3
ATLINE1_1
ATLINE1_2
ATLINE1_3A
ATLINE1_4
ATLINE1_5
ATLINE1_6
ATLINE1A
ATLINE2
TE Family


TE Variants
Enrichment (%)
2
1
0
−1
−2
ATLINEIII
ATMU1
ATMU10
ATMU11
ATMU5
ATMU6
ATMU6N1
ATMU7
ATMUN1
ATMUNX1
ATN9_1
ATRE1
ATREP1
ATREP7
ATREP10A
ATREP10B
ATREP10C
ATREP10D
ATREP11
ATREP11A
ATREP12
ATREP13
ATREP14
ATREP15
ATREP16
ATREP17
ATREP18
ATREP19
ATREP2
ATREP2A
ATREP3
ATREP4
ATREP5
ATREP6
ATREP7
ATREP8
ATREP9
ATSINE2A
ATSINE4
ATTIR16T3A
ATTIRTA1
ATTIRX1A
ATTIRX1B
ATTIRX1C
BOMZH1
BOMZH2
BRODYAGA1
BRODYAGA1A
BRODYAGA2
DT1
HARBINGER
HELITRON1
HELITRON10
HELITRON2
HELITRON3
HELITRON4
HELITRON5
HELITRONY1A
HELITRONY1B
HELITRONY1C
HELITRONY1D
HELITRONY1E
HELITRONY2
HELITRONY3
HELITRONY3A
LIMPET1
META1
RathE1_cons
RathE2_cons
RathE3_cons
ROMANIAT5
RP1_AT
SIMPLEHAT1
SIMPLEHAT6
SIMPLEHAT7
SIMPLEHAT2
TA11
TAG1
TAG2
TAG3N1
TAT1_ATH
TNAT1A
TNAT2A
TSGL
Unassigned
VANDAL1
VANDAL12
VANDAL13
VANDAL14
VANDAL16
VANDAL17
VANDAL18NA
VANDAL1N1
VANDAL2
VANDAL20
VANDAL21
VANDAL22
VANDAL2N1
VANDAL3
VANDAL4
VANDAL5
VANDAL5A
VANDAL6
VANDAL7
VANDAL8
VANDAL9
VANDALNX2
TE Family

982 **Figure S5. TE variant enrichment and depletion for all TE families**

983 The percentage of TE variants by TE families were compared to the percentage expected due
984 to the genomic frequency of elements of each TE family. Families with more TE
985 presence/absence variants than expected are plotted in red, indicating percentage enrichment
986 of those elements, while those with a lower number of TE variants than expected are plotted in
987 blue, indicating percentage depletion for that TE family.


Embryo mC for Ler−absent TEs
log2 fold change mC
Col-0 x Ler vs Ler x Col-0
0.2
0.1
0.0
−0.1
−0.2
−0.3
Higher mC Col-0 x Ler
TE absent paternal genome
Higher mC Ler x Col-0
TE absent maternal genome
CG
CHG
CHH

989 **Figure S6. TE DNA methylation in embryos resulting from a Col-0 x L*er* cross**

990 DNA methylation levels in sperm-demethylated TEs absent from L*er* but present in Col-0.

991 Levels are $\log_2$ fold change mC/C, for each DNA methylation context, between L*er* x Col-0 and

992 Col-0 x L*er* DNA methylation levels. Positive values indicate higher methylation in the L*er* x Col-

993 0 (*female x male*) cross, whereas negative values represent higher methylation in the Col-0 x

994 L*er* cross. Boxplots were constructed as for Fig. 4C.

995 # TABLES

996 **Table 1. Mapping of paired-end reads providing evidence for TE presence/absence**
997 **variants in L*er* to Col-0 and L*er* reference genomes.**

	Concordant	Discordant	Split	Unmapped	Total
Col-0 mapped	0	993	9,513	0	10,206
Ler mapped	10,073	92	34	7	10,206

998

999 **Table 2. Significantly differentially expressed genes dependent on TE presence/absence.**

TE insertion point	AGI	Gene name	Function	Pearson correlation	log2 fold change FPKM	p-value	q-value
Exon	AT1G28620	-	Lipase	0.27	-3.71	2.74E-06	1.73E-04
	AT1G31580	*ECS1*	Cell wall protein related to Xanthomonas campestris pv. campestris resistance	0.28	-2.91	9.43E-10	2.49E-07
	AT1G48820	-	Terpenoid cyclase	0.27	-2.66	1.86E-05	7.46E-04
	AT2G15040	*AtRLP18*	Disease resistance protein	0.49	-1.62	5.80E-10	2.42E-07
	AT2G16520	-	-	-0.53	3.07	1.76E-10	9.05E-08
	AT2G16810	-	F-box protein	-0.48	2.36	1.86E-05	7.46E-04
	AT3G19520	-	-	0.68	-2.61	9.98E-08	1.14E-05
	AT3G26230	*CYP71B24*	Cytochrome P450	0.53	-2.40	6.44E-13	6.63E-10
	AT3G46482	-	-	-0.48	4.46	2.05E-09	4.52E-07
	AT3G62460	-	Putative endonuclease	0.20	-2.56	9.03E-08	1.12E-05
	AT4G05500	-	F-box protein	-0.44	1.26	5.13E-06	2.66E-04
	AT4G15056	-	-	-0.60	4.63	4.25E-06	2.34E-04
	AT5G16990	-	Oxidative stress tolerance	-0.49	1.02	5.26E-06	2.66E-04
	AT5G38565	-	F-box protein	0.35	-1.84	2.13E-06	1.44E-04
	AT5G67310	*CYP81G1*	Cytochrome P450	0.22	-3.53	3.61E-07	3.47E-05
Intron	AT1G48820	-	Terpenoid cyclase	0.27	-2.66	1.86E-05	7.46E-04
	AT1G52710	-	Cytochrome-c oxidase activity	0.30	-2.57	8.30E-06	4.07E-04
	AT2G03913	-	Defensin-like family protein	-0.72	8.35	1.64E-17	5.06E-14
	AT3G59190	-	F-box protein	0.29	-3.39	4.13E-09	8.50E-07
	AT3G62460	-	Putative endonuclease	0.20	-2.56	9.03E-08	1.12E-05
	AT4G15056	-	-	-0.60	4.63	4.25E-06	2.34E-04
3' UTR	AT5G38565	-	F-box protein	0.35	-1.84	2.13E-06	1.44E-04
5' UTR	AT1G31580	*ECS1*	Cell wall protein related to Xanthomonas campestris pv. campestris resistance	0.28	-2.91	9.43E-10	2.49E-07
	AT2G16810	-	F-box protein	-0.48	2.36	1.86E-05	7.46E-04
	AT5G16990	-	Oxidative stress tolerance	-0.49	1.02	5.26E-06	2.66E-04
Upstream	AT1G48700	-	Oxidoreductase activity	-0.42	1.56	4.31E-07	3.91E-05
	AT1G61415	-	-	-0.38	1.90	8.52E-07	6.74E-05
	AT2G01360	-	PPR protein	-0.84	5.06	1.80E-07	1.80E-05
	AT2G16520	-	-	-0.53	3.07	1.76E-10	9.05E-08
	AT2G16530	*PPRD2*	Oxidoreductase activity	0.42	-1.01	1.05E-07	1.16E-05
	AT2G17080	-	-	-0.49	2.65	1.15E-08	1.97E-06
	AT2G25510	-	-	0.43	-2.49	2.52E-08	3.89E-06

	AT4G11290	*AtPrx39*	Peroxidase activity, cold tolerance	0.35	-1.43	2.15E-06	1.44E-04
	AT4G15620	*CASPL1E2*	-	0.38	-1.70	3.29E-08	4.61E-06
	AT5G28010	-	Defence response	0.38	-1.76	9.69E-10	2.49E-07
	AT5G32620	-	-	-0.36	2.27	3.03E-08	4.45E-06
	AT5G38565	-	F-box protein	0.35	-1.84	2.13E-06	1.44E-04
	AT5G49500	-	Signal recognition particle	-0.41	1.62	2.29E-05	8.93E-04
	AT5G66660	-	-	-0.39	1.03	1.55E-05	7.03E-04
Downstream	AT1G02770	-	-	-0.46	1.36	3.71E-07	3.47E-05
	AT1G33930	-	GTP binding	0.55	-2.61	3.57E-15	5.51E-12
	AT1G77960	-	-	0.15	-4.51	4.22E-06	2.34E-04
	AT2G12460	-	-	0.29	-2.65	7.07E-07	5.74E-05
	AT2G16810	-	F-box protein	-0.48	2.36	1.86E-05	7.46E-04
	AT2G19400	-	Kinase	0.42	-1.54	1.40E-06	1.08E-04
	AT2G27650	-	Ubiquitin thiolesterase activity	-0.41	1.20	7.88E-06	3.93E-04
	AT3G46482	-	-	-0.48	4.46	2.05E-09	4.52E-07
	AT4G03050	*AOP3*	Glucosinolate biosynthesis	-0.38	2.39	4.82E-07	4.13E-05
	AT4G07820	-	-	0.29	-2.78	1.48E-06	1.11E-04
	AT5G16990	-	Oxidative stress tolerance	-0.49	1.02	5.26E-06	2.66E-04
	AT5G22608	-	-	0.33	-1.94	4.76E-07	4.13E-05
	AT5G23770	*AtDUF8*	-	-0.56	1.53	5.50E-07	4.59E-05
	AT5G24570	-	-	0.27	-1.01	2.43E-05	9.37E-04
	AT5G27340	-	-	-0.41	4.08	4.26E-08	5.72E-06
	AT5G35796	-	-	-0.50	1.36	7.65E-09	1.39E-06
	AT5G45840	-	Protein kinase activity	0.25	-3.22	1.51E-06	1.11E-04
	AT5G66455	-	PPR pseudogene	0.32	-1.44	1.39E-05	6.52E-04

1000

1001 **Table S1. TE superfamily enrichments for TE variants**

TE superfamily	Genomic count	Variant count	Genomic frequency	Variant frequency	Enrichment (%)
DNA	1829	798	5.86E+00	3.76E-02	-2.10
DNA/En-Spm	941	933	3.02E+00	4.40E-02	1.38
DNA/Harbinger	379	339	1.22E+00	1.60E-02	0.38
DNA/HAT	1035	1109	3.32E+00	5.23E-02	1.91
DNA/Mariner	151	98	4.84E-01	4.62E-03	-0.02
DNA/MuDR	5410	3777	1.73E+01	1.78E-01	0.45
DNA/Pogo	344	127	1.10E+00	5.98E-03	-0.50
DNA/Tc1	95	37	3.05E-01	1.74E-03	-0.13
LINE	81	45	2.60E-01	2.12E-03	-0.05
LINE/L1	1366	1017	4.38E+00	4.79E-02	0.41
LTR/Copia	1781	1747	5.71E+00	8.23E-02	2.52
LTR/Gypsy	4181	4442	1.34E+01	2.09E-01	7.53
RathE1_cons	213	150	6.83E-01	7.07E-03	0.02
RathE2_cons	74	14	2.37E-01	6.60E-04	-0.17
RathE3_cons	104	72	3.33E-01	3.39E-03	0.01
RC/Helitron	12945	6372	4.15E+01	3.00E-01	-11.48
SINE	131	64	4.20E-01	3.02E-03	-0.12
Unassigned	129	80	4.14E-01	3.77E-03	-0.04

1002

1003 **Table S2. TE family enrichments for TE variants.**

TE family	Genomic count	Variant count	Genomic frequency	Variant frequency	Enrichment (%)
ARNOLD1	39	13	1.30E-01	6.48E-04	-0.06
ARNOLD2	68	42	2.26E-01	2.09E-03	-0.02
ARNOLD3	45	38	1.49E-01	1.90E-03	0.04
ARNOLD4	23	11	7.64E-02	5.49E-04	-0.02
ARNOLDY1	237	100	7.87E-01	4.99E-03	-0.29
ARNOLDY2	288	131	9.56E-01	6.53E-03	-0.30
AT9NMU1	45	68	1.49E-01	3.39E-03	0.19
AT9TSD1	55	58	1.83E-01	2.89E-03	0.11
ATCOPIA11	26	22	8.63E-02	1.10E-03	0.02
ATCOPIA13	35	31	1.16E-01	1.55E-03	0.04
ATCOPIA24	27	9	8.97E-02	4.49E-04	-0.04
ATCOPIA28	81	57	2.69E-01	2.84E-03	0.02
ATCOPIA30	33	12	1.10E-01	5.99E-04	-0.05
ATCOPIA35	22	27	7.31E-02	1.35E-03	0.06
ATCOPIA37	32	14	1.06E-01	6.98E-04	-0.04
ATCOPIA41	39	39	1.30E-01	1.95E-03	0.07
ATCOPIA42	35	17	1.16E-01	8.48E-04	-0.03
ATCOPIA43	29	18	9.63E-02	8.98E-04	-0.01
ATCOPIA45	20	14	6.64E-02	6.98E-04	0.00
ATCOPIA49	34	32	1.13E-01	1.60E-03	0.05
ATCOPIA57	48	50	1.59E-01	2.49E-03	0.09
ATCOPIA59	20	5	6.64E-02	2.49E-04	-0.04
ATCOPIA65	34	29	1.13E-01	1.45E-03	0.03
ATCOPIA66	20	10	6.64E-02	4.99E-04	-0.02
ATCOPIA68	43	14	1.43E-01	6.98E-04	-0.07
ATCOPIA69	25	36	8.30E-02	1.80E-03	0.10
ATCOPIA75	20	19	6.64E-02	9.48E-04	0.03
ATCOPIA78	24	64	7.97E-02	3.19E-03	0.24
ATCOPIA94	39	21	1.30E-01	1.05E-03	-0.02
ATCOPIA95	72	49	2.39E-01	2.44E-03	0.01
ATDNA12T3_2	660	53	2.19E+00	2.64E-03	-1.93
ATDNA12T3A	22	12	7.31E-02	5.99E-04	-0.01
ATDNA1T9A	36	11	1.20E-01	5.49E-04	-0.06
ATDNA2T9A	48	50	1.59E-01	2.49E-03	0.09
ATDNA2T9B	26	25	8.63E-02	1.25E-03	0.04
ATDNA2T9C	242	103	8.04E-01	5.14E-03	-0.29
ATDNAI26T9	25	24	8.30E-02	1.20E-03	0.04
ATDNAI27T9A	116	150	3.85E-01	7.48E-03	0.36
ATDNAI27T9B	43	49	1.43E-01	2.44E-03	0.10
ATDNAI27T9C	108	54	3.59E-01	2.69E-03	-0.09
ATDNATA1	28	36	9.30E-02	1.80E-03	0.09
ATENSPM1	30	32	9.96E-02	1.60E-03	0.06

ATENSPM10	81	61	2.69E-01	3.04E-03	0.04
ATENSPM11	47	48	1.56E-01	2.39E-03	0.08
ATENSPM1A	77	59	2.56E-01	2.94E-03	0.04
ATENSPM2	114	88	3.79E-01	4.39E-03	0.06
ATENSPM3	98	127	3.25E-01	6.33E-03	0.31
ATENSPM4	50	99	1.66E-01	4.94E-03	0.33
ATENSPM5	111	88	3.69E-01	4.39E-03	0.07
ATENSPM6	151	112	5.01E-01	5.59E-03	0.06
ATENSPM7	74	65	2.46E-01	3.24E-03	0.08
ATENSPM9	97	150	3.22E-01	7.48E-03	0.43
ATGP1	144	168	4.78E-01	8.38E-03	0.36
ATGP10	165	151	5.48E-01	7.53E-03	0.21
ATGP2	48	97	1.59E-01	4.84E-03	0.32
ATGP2N	54	77	1.79E-01	3.84E-03	0.20
ATGP3	60	36	1.99E-01	1.80E-03	-0.02
ATGP5	86	67	2.86E-01	3.34E-03	0.05
ATGP6	25	18	8.30E-02	8.98E-04	0.01
ATGP7	47	37	1.56E-01	1.85E-03	0.03
ATGP8	92	32	3.06E-01	1.60E-03	-0.15
ATGP9B	27	22	8.97E-02	1.10E-03	0.02
ATHAT1	105	32	3.49E-01	1.60E-03	-0.19
ATHAT10	35	15	1.16E-01	7.48E-04	-0.04
ATHAT3	23	72	7.64E-02	3.59E-03	0.28
ATHAT7	20	11	6.64E-02	5.49E-04	-0.01
ATHATN1	69	18	2.29E-01	8.98E-04	-0.14
ATHATN10	48	43	1.59E-01	2.14E-03	0.06
ATHATN2	46	37	1.53E-01	1.85E-03	0.03
ATHATN3	70	20	2.32E-01	9.98E-04	-0.13
ATHATN3A	20	10	6.64E-02	4.99E-04	-0.02
ATHATN4	56	12	1.86E-01	5.99E-04	-0.13
ATHATN5	42	8	1.39E-01	3.99E-04	-0.10
ATHATN6	43	71	1.43E-01	3.54E-03	0.21
ATHATN7	71	50	2.36E-01	2.49E-03	0.01
ATHILA	198	229	6.58E-01	1.14E-02	0.48
ATHILA0_I	138	180	4.58E-01	8.98E-03	0.44
ATHILA2	413	618	1.37E+00	3.08E-02	1.71
ATHILA3	243	296	8.07E-01	1.48E-02	0.67
ATHILA4	250	265	8.30E-01	1.32E-02	0.49
ATHILA4A	310	254	1.03E+00	1.27E-02	0.24
ATHILA4B_LTR	143	234	4.75E-01	1.17E-02	0.69
ATHILA4C	206	183	6.84E-01	9.13E-03	0.23
ATHILA4D_LTR	106	47	3.52E-01	2.34E-03	-0.12
ATHILA5	131	172	4.35E-01	8.58E-03	0.42
ATHILA6A	247	331	8.20E-01	1.65E-02	0.83
ATHILA6B	134	131	4.45E-01	6.53E-03	0.21

ATHILA7	134	46	4.45E-01	2.29E-03	-0.22
ATHILA7A	31	18	1.03E-01	8.98E-04	-0.01
ATHILA8A	97	77	3.22E-01	3.84E-03	0.06
ATHILA8B	78	41	2.59E-01	2.04E-03	-0.05
ATHPOGO	25	10	8.30E-02	4.99E-04	-0.03
ATHPOGON1	174	52	5.78E-01	2.59E-03	-0.32
ATHPOGON2	27	12	8.97E-02	5.99E-04	-0.03
ATHPOGON3	118	53	3.92E-01	2.64E-03	-0.13
ATIS112A	173	122	5.75E-01	6.09E-03	0.03
ATLANTYS1	153	171	5.08E-01	8.53E-03	0.34
ATLANTYS2	179	194	5.94E-01	9.68E-03	0.37
ATLANTYS3	142	126	4.72E-01	6.28E-03	0.16
ATLINE1_1	61	70	2.03E-01	3.49E-03	0.15
ATLINE1_2	41	51	1.36E-01	2.54E-03	0.12
ATLINE1_3A	160	95	5.31E-01	4.74E-03	-0.06
ATLINE1_4	102	39	3.39E-01	1.95E-03	-0.14
ATLINE1_5	91	52	3.02E-01	2.59E-03	-0.04
ATLINE1_6	129	75	4.28E-01	3.74E-03	-0.05
ATLINE1A	289	147	9.60E-01	7.33E-03	-0.23
ATLINE2	138	146	4.58E-01	7.28E-03	0.27
ATLINEIII	196	292	6.51E-01	1.46E-02	0.81
ATMU1	59	35	1.96E-01	1.75E-03	-0.02
ATMU10	87	30	2.89E-01	1.50E-03	-0.14
ATMU11	20	11	6.64E-02	5.49E-04	-0.01
ATMU2	64	65	2.13E-01	3.24E-03	0.11
ATMU5	24	10	7.97E-02	4.99E-04	-0.03
ATMU6	26	31	8.63E-02	1.55E-03	0.07
ATMU6N1	39	20	1.30E-01	9.98E-04	-0.03
ATMU7	32	32	1.06E-01	1.60E-03	0.05
ATMUN1	107	49	3.55E-01	2.44E-03	-0.11
ATMUNX1	155	39	5.15E-01	1.95E-03	-0.32
ATN9_1	65	99	2.16E-01	4.94E-03	0.28
ATRE1	22	9	7.31E-02	4.49E-04	-0.03
ATREP1	498	390	1.65E+00	1.95E-02	0.29
ATREP10	63	51	2.09E-01	2.54E-03	0.05
ATREP10A	279	85	9.27E-01	4.24E-03	-0.50
ATREP10B	491	150	1.63E+00	7.48E-03	-0.88
ATREP10C	123	41	4.08E-01	2.04E-03	-0.20
ATREP10D	1295	489	4.30E+00	2.44E-02	-1.86
ATREP11	841	339	2.79E+00	1.69E-02	-1.10
ATREP11A	95	72	3.15E-01	3.59E-03	0.04
ATREP12	22	72	7.31E-02	3.59E-03	0.29
ATREP13	106	110	3.52E-01	5.49E-03	0.20
ATREP14	47	30	1.56E-01	1.50E-03	-0.01
ATREP15	1003	331	3.33E+00	1.65E-02	-1.68

ATREP16	47	103	1.56E-01	5.14E-03	0.36
ATREP17	34	9	1.13E-01	4.49E-04	-0.07
ATREP18	391	298	1.30E+00	1.49E-02	0.19
ATREP19	189	42	6.28E-01	2.09E-03	-0.42
ATREP2	164	306	5.45E-01	1.53E-02	0.98
ATREP2A	116	135	3.85E-01	6.73E-03	0.29
ATREP3	1439	878	4.78E+00	4.38E-02	-0.40
ATREP4	832	239	2.76E+00	1.19E-02	-1.57
ATREP5	624	237	2.07E+00	1.18E-02	-0.89
ATREP6	181	90	6.01E-01	4.49E-03	-0.15
ATREP7	164	114	5.45E-01	5.69E-03	0.02
ATREP8	75	58	2.49E-01	2.89E-03	0.04
ATREP9	201	56	6.67E-01	2.79E-03	-0.39
ATSINE2A	33	13	1.10E-01	6.48E-04	-0.04
ATSINE4	98	51	3.25E-01	2.54E-03	-0.07
ATTIR16T3A	106	71	3.52E-01	3.54E-03	0.00
ATTIRTA1	95	37	3.15E-01	1.85E-03	-0.13
ATTIRX1A	63	15	2.09E-01	7.48E-04	-0.13
ATTIRX1B	44	21	1.46E-01	1.05E-03	-0.04
ATTIRX1C	39	93	1.30E-01	4.64E-03	0.33
ATTIRX1D	28	9	9.30E-02	4.49E-04	-0.05
BOMZH1	28	45	9.30E-02	2.24E-03	0.13
BOMZH2	47	27	1.56E-01	1.35E-03	-0.02
BRODYAGA1	251	87	8.34E-01	4.34E-03	-0.40
BRODYAGA1A	586	163	1.95E+00	8.13E-03	-1.13
BRODYAGA2	525	179	1.74E+00	8.93E-03	-0.85
DT1	123	62	4.08E-01	3.09E-03	-0.10
HARBINGER	90	62	2.99E-01	3.09E-03	0.01
HELITRON1	130	124	4.32E-01	6.18E-03	0.19
HELITRON2	194	89	6.44E-01	4.44E-03	-0.20
HELITRON3	53	54	1.76E-01	2.69E-03	0.09
HELITRON4	199	79	6.61E-01	3.94E-03	-0.27
HELITRON5	56	89	1.86E-01	4.44E-03	0.26
HELITRONY1A	271	165	9.00E-01	8.23E-03	-0.08
HELITRONY1B	414	146	1.37E+00	7.28E-03	-0.65
HELITRONY1C	120	73	3.98E-01	3.64E-03	-0.03
HELITRONY1D	756	273	2.51E+00	1.36E-02	-1.15
HELITRONY1E	447	196	1.48E+00	9.78E-03	-0.51
HELITRONY2	198	73	6.58E-01	3.64E-03	-0.29
HELITRONY3	1399	646	4.65E+00	3.22E-02	-1.42
HELITRONY3A	49	92	1.63E-01	4.59E-03	0.30
LIMPET1	116	58	3.85E-01	2.89E-03	-0.10
META1	138	202	4.58E-01	1.01E-02	0.55
RathE1_cons	213	150	7.07E-01	7.48E-03	0.04
RathE2_cons	74	14	2.46E-01	6.98E-04	-0.18

RathE3_cons	104	72	3.45E-01	3.59E-03	0.01
ROMANIAT5	49	30	1.63E-01	1.50E-03	-0.01
RP1_AT	87	102	2.89E-01	5.09E-03	0.22
SIMPLEGUY1	116	155	3.85E-01	7.73E-03	0.39
SIMPLEHAT1	56	83	1.86E-01	4.14E-03	0.23
SIMPLEHAT2	73	80	2.42E-01	3.99E-03	0.16
TA11	157	49	5.21E-01	2.44E-03	-0.28
TAG1	28	2	9.30E-02	9.98E-05	-0.08
TAG2	86	127	2.86E-01	6.33E-03	0.35
TAG3N1	97	399	3.22E-01	1.99E-02	1.67
TAT1_ATH	87	117	2.89E-01	5.84E-03	0.29
TNAT1A	162	51	5.38E-01	2.54E-03	-0.28
TNAT2A	38	31	1.26E-01	1.55E-03	0.03
TSCL	81	45	2.69E-01	2.24E-03	-0.04
Unassigned	113	61	3.75E-01	3.04E-03	-0.07
VANDAL1	80	42	2.66E-01	2.09E-03	-0.06
VANDAL12	37	20	1.23E-01	9.98E-04	-0.02
VANDAL13	38	54	1.26E-01	2.69E-03	0.14
VANDAL14	25	66	8.30E-02	3.29E-03	0.25
VANDAL16	42	87	1.39E-01	4.34E-03	0.29
VANDAL17	178	128	5.91E-01	6.38E-03	0.05
VANDAL18NA	24	15	7.97E-02	7.48E-04	0.00
VANDAL1N1	52	86	1.73E-01	4.29E-03	0.26
VANDAL2	43	47	1.43E-01	2.34E-03	0.09
VANDAL20	51	56	1.69E-01	2.79E-03	0.11
VANDAL21	64	119	2.13E-01	5.94E-03	0.38
VANDAL22	48	94	1.59E-01	4.69E-03	0.31
VANDAL2N1	51	63	1.69E-01	3.14E-03	0.14
VANDAL3	177	152	5.88E-01	7.58E-03	0.17
VANDAL4	102	68	3.39E-01	3.39E-03	0.00
VANDAL5	85	56	2.82E-01	2.79E-03	0.00
VANDAL5A	56	33	1.86E-01	1.65E-03	-0.02
VANDAL6	89	95	2.96E-01	4.74E-03	0.18
VANDAL7	36	31	1.20E-01	1.55E-03	0.04
VANDAL8	123	198	4.08E-01	9.88E-03	0.58
VANDAL9	43	24	1.43E-01	1.20E-03	-0.02
VANDALNX2	37	19	1.23E-01	9.48E-04	-0.03