The *Arabidopsis thaliana* mobilome and its impact at the species level

Leandro Quadrana[1], Amanda Bortolini Silveira[1], George F. Mayhew[2], Chantal LeBlanc[4], Robert A. Martienssen[3], Jeffrey A. Jeddeloh[2], and Vincent Colot[1*]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Ecole Normale Supérieure, F-75005 Paris, France

[2]Roche NimbleGen, Inc, Madison, WI, 53719, USA

[3]Howard Hughes Medical Institute-Gordon and Betty Moore Foundation, Watson School of Biological Sciences, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

[4]Department of Molecular, Cellular & Developmental Biology. Yale University. 352A Osborn Memorial Laboratories. 165 Prospect St, New Haven, CT, 06511, USA

* Corresponding author: colot@biologie.ens.fr

Transposable elements (TEs) are powerful motors of genome evolution yet a comprehensive assessment of recent transposition activity at the species level is lacking for most organisms. Here, using genome sequencing data for 211 *Arabidopsis thaliana* accessions taken from across the globe, we identify thousands of recent transposition events involving half of the 326 TE families annotated in this plant species. We further show that the composition and activity of the "mobilome" vary extensively between accessions in relation to climate and genetic factors. Moreover, TEs insert equally throughout the genome and are rapidly purged by natural selection from gene-rich regions because they frequently affect genes, in multiple ways. Remarkably, loci controlling adaptive responses to the environment are the most frequent transposition targets observed. These findings demonstrate the pervasive, species-wide impact that a rich mobilome can have and the importance of transposition as a recurrent generator of large-effect alleles.

INTRODUCTION

Transposable elements (TEs) are sequences that move and replicate around the genome. Depending on whether their mobilization relies on a RNA or DNA intermediate, they are classified as retrotransposons (class I) or DNA transposons (class II), respectively (Slotkin and Martienssen 2007). TEs are further subdivided into distinct families, the prevalence of which differs between organisms because of a complex array of factors, including variable transposition activity and diverse selection pressures (Barrón et al. 2014). Given their mobile nature, TEs pose multiple threats to the physical and functional integrity of genomes. In particular, TEs can disrupt genes through insertion and also through excision in the case of DNA transposons. Thus, TE mobilization is a source of both germline and somatic mutations (Richardson et al. 2015; Barrón et al. 2014; Lisch 2013). Although TEs are endogenous mutagens with potentially catastrophic effects, their mobilization might sometimes be beneficial. In fact, soon after their discovery, Barbara McClintock named TEs "controlling elements" to emphasize their role in the control of gene action (McClintock 1956). In mammals, transposition of some *LINE 1* retrotransposons occurs extensively during embryogenesis as well as in the adult brain, again suggesting functional relevance of somatic TE mobilization (Richardson, Morell, and Faulkner 2014). Nonetheless, TEs are under tight control to limit their mutational impact both within and across generations. In plants and mammals, a major control is through epigenetic silencing mechanisms, including DNA methylation and these mechanisms can in turn have "epimutagenic" effects on adjacent genes (Richards, 2006; Slotkin and Martienssen 2007; Weigel and Colot 2012; Heard and Martienssen 2014; Quadrana and Colot 2016).

Despite the many documented short-term as well as evolutionary consequences of TE mobilization (Rebollo, Romanish, and Mager 2012; Trono 2016; Kazazian 2004; Elbarbary, Lucas, and Maquat 2016; Bennetzen and Wang 2014; Lisch 2013), TEs are among the least investigated components of genomes, mainly because they are present in multiple, often degenerated copies, which complicate analysis. Thus, a species-wide view of the mobilome - i.e. of the set of TE families with transposition activity - is lacking for most organisms.

Studies in humans suggest that although at least half of the 3Gb genome is made up of TE sequences, mainly belonging to *LINE 1* and *SINE* retrotransposon families, a few these only and none of the other TE families contain mobile copies (Richardson et al. 2015). In contrast, the number of TE families that have retained transposition activity is much larger in the mouse and these families include several so-called endogenous retroviruses (ERVs) in addition to *LINEs* and *SINEs* (Richardson et al. 2015). In Drosophila, which has a much smaller genome (～120 Mb) characterized by a large repertoire of class I and class II TE families, the situation

is again different with most of these TE families likely mobile (Mackay et al. 2012; Cridland et al. 2013; Kofler, Nolte, and Schlötterer 2015; Rahman et al. 2015). However, in this species and even more so in mammals, the population genetics of the mobilome remains poorly characterized.

The flowering plant *A. thaliana* is particularly attractive for conducting a systematic survey of the mobilome and of its molecular as well as phenotypic impact at the species level. First, like Drosophila, *A. thaliana* has a compact genome and a large repertoire of class I and class II TE families (The Arabidopsis Genome Initiative 2000; Ahmed et al. 2011; Joly-Lopez and Bureau 2014). Thus, most TE families are of relatively small size, which facilitates their study. Second, *A. thaliana* occupies a wide range of habitats across the globe and representative accessions have been extensively characterized both genetically and phenotypically (Weigel and Nordborg 2015). Third, whole genome sequencing has been performed for >1000 *A. thaliana* accessions and DNA methylome as well as transcriptome data are also available for hundreds of these (Schmitz et al. 2013; Long et al. 2013; Dubin et al. 2015; Cao et al. 2011). Finally, genome-wide association studies (GWASs) are straightforward in this species (Weigel and Nordborg 2015).

Here we present a comprehensive assessment of the *A. thaliana* mobilome, which provides important novel insights into the population genetics of TE mobilization and radically changes the prevailing view of limited transposition potential in this species (Hu et al. 2011; Maumus and Quesneville 2014). Specifically, we show that the *A. thaliana* mobilome is composed of a very large number of class I and class II TE families overall, but differs markedly among accessions. We further show that TE mobilization is a complex trait and we have identified environmental as well as genetic factors that influence transposition in nature. These factors include the annual temperature range, the TE themselves and multiple gene loci, notably *MET2a*, which encodes a poorly characterized DNA methyltransferase. In addition, we present compelling evidence that TEs insert throughout the genome with no overt bias and that the mobilome has a pervasive impact on the expression and DNA methylation status of adjacent genes. These and other observations indicate that purifying selection is most probably the main factor responsible for the differential accumulation of TE sequences along the *A. thaliana* genome and notably their clustering in pericentromeric regions. Finally, we reveal the importance of the mobilome as a generator of large-effect alleles at loci underlying adaptive traits. Collectively, our approaches and findings provide a unique framework for detailed studies of the dynamics and impact of transposition in nature.

RESULTS

Composition of the *A. thaliana* mobilome

The reference genome sequence of *A.thaliana* is 125 Mb long (TAIR 10) and contains ~32000, mostly degenerate TE copies that belong to 326 distinct families (The Arabidopsis Genome Initiative 2000; Ahmed et al. 2011). So far, transposition activity has been documented experimentally for nine TE families, mainly based on studies carried out in the reference accession Col-0 (Ito and Kakutani 2014; Tsay et al. 1993). To assess species-wide the composition of the *A.thaliana* mobilome, we used publically available Illumina short genome sequence reads (Schmitz et al. 2013; Schneeberger et al. 2011). First, we looked for TE copy number variation (CNV) between the reference accession Columbia (Col-0) and 211 accessions taken from across the globe. To limit the problem posed by the presence of TEs in multiple copies across the genome, with varying degrees of similarity to each other, we performed an aggregated CNV analysis based on the 11,851 annotated Col-0 TE sequences longer than 300bp (see 'Materials and methods'). CNVs were detected for 263 TE families (Figure 1A and B; Figure 1-Source data 1; see 'Materials and methods'), in keeping with the results of a previous study indicating that the vast majority of the TE sequences annotated in the Col-0 reference genome are absent from that of at least one of 80 accessions analyzed (Cao et al. 2011).

Since CNVs could reflect either recent TE mobilization or the gain or loss of TE copies through other types of chromosomal rearrangements, we then looked among the unmapped Illumina short reads for so-called "split-reads" that contain TE extremities. Crucially, because most TE families generate short target site duplications (TSDs) of fixed size upon insertion, TSDs can serve as signatures of *bona fide* transposition events. We therefore developed a pipeline for the systematic identification of split-reads covering TE junctions that are absent from the reference genome and that produce, when mapped to the insertion site, a sequence overlap of the size of TSDs (3 to 15 bp, depending on the TE family, Figure 1C and D; see 'Materials and methods'). Our pipeline differs in that respect from that used in another study to detect the presence/absence of reference and non-reference TE insertions in the same set of accessions (Stuart et al. 2016). Results produced by our pipeline for the 292 annotated TE families that create TSDs upon transposition were verified visually to eliminate false positives (Figure 1-figure Supplement 2; see 'Materials and methods'). Following this approach, non-reference TE insertions with TSDs were identified for 131 TE families in total (Figure 1-Source data 2), which all also show CNV (Figure 1-Source data 1).

Most (86%) non-reference TE insertions with TSDs are private or shared only by a few accessions and thus

they typically correspond to recently derived alleles, as expected (Figure 1E). Moreover, recent transposition activity is only detected for between four and 66 TE families in any given accession, thus indicating large variations in the composition of the mobilome among accessions (Figure 1F). However, we have probably identified most of the annotated TE families that compose the mobilome at the species level, since the number of TE families defined as mobile by the split-reads approach reaches a plateau after examining just 74 accessions (Figure 1G). The 53 class I *COPIA* families and the 40 class II Mutator-like (*MuDR*) families are the most mobile, as they account for 1408 and 729 of the 2835 non-reference TE insertions with TSDs identified in total, respectively (Figure 1H and Figure 1-Source data 2 and 3). However, the number of non-reference insertions per accession is always small (<16) for any given family (Figure 1-Source data 2), thus suggesting a lack of recent transposition bursts.

The ability to detect non-reference TE insertions with TSDs using split-reads is strongly dependent on read depth as well as sequence composition at the insertion site (Figure 1-figure supplement 1A and B; Hénaff et al. 2015). To assess the extent of this limitation, we used the assembled Ler-1 genome sequence recently obtained using PacBio long reads (see 'Materials and methods'). Although not annotated, this sequence assembly can nonetheless serve to identify by whole genome comparison the Col-0 TEs flanked by TSDs that are absent from the corresponding position in Ler-1 (see 'Materials and Methods'). A total of 142 TEs belonging to 80 distinct families were identified in this way (Figure 1-Source data 4), two numbers that are consistent with estimates obtained using other approaches (Ziolkowski et al. 2009; Hénaff et al. 2015). In contrast, we could detect only 78 Col-0-specific TEs with TSDs belonging to 49 TE families when using the split-reads pipeline to map Col-0 short reads onto the assembled Ler-1 genome. These results indicate therefore that the split-reads approach has a low sensitivity (Figure 1-figure supplement 1C; 45% false negatives and 10% false positives; False Discovery Rate: 15.3%).

To obtain an independent estimation of the composition of the mobilome, we also performed TE sequence capture (TE-capture; Baillie et al. 2011). Briefly, probes were designed to cover the 5' and 3' extremities of 310 TE elements belonging to 181 distinct families, including 117 of the 131 TE families identified as mobile using the split-reads approach (see 'Materials and methods'). Using genomic DNA extracted from 12 randomly chosen accessions (Figure 1-figure supplement 1D and E), we could validate by TE-capture most (87%) of the non-reference TE insertions with TSDs that were detected by the split-reads approach (Figure 1-figure supplement 1F; see 'Materials and methods'). As expected, TE-capture also uncovered many additional non-

reference TE insertions with TSDs for these TE families (Figure 1-figure supplement 1F). However, no such insertions were detected for the other TE families that could be captured but which were not identified as mobile by the split-reads approach in any of the 12 accessions. These results confirm that despite the low sensitivity of the split-reads approach, we have probably identified most of the TE families with TSDs that compose the *A. thaliana* mobilome at the species level. Finally, non-reference insertions were also identified for 30 TE families (including 15 *HELITRON* families) that could not be analyzed using our split-reads pipeline because they do not produce TSDs or have insertion sites located in low complexity regions (Figure 1-figure supplement 1B). Since most of the non-reference insertions for these 30 TE families are present in only one or two of the 12 accessions examined (Figure 1-figure supplement 1G), they likely reflect recent transposition events. Thus, there are altogether at least 165 TE families with recent transposition activity at the species level. Moreover, based on the TE-capture data, we can estimate that depending on their divergence time, any two accessions have accumulated between ~200 and ~300 newly transposed TE copies (Figure 1-figure supplement 1H).

TE mobilization as a complex trait

The observation that the composition of the mobilome differs extensively between accessions (Figure 1F) suggests that it is influenced by both environmental and genetic factors. To try to identify such factors, we first established that copy number (CN) correlates positively with the number of TE insertions with TSDs that are detected by TE-capture (Figure 2-figure supplement 1; see 'Materials and methods'). Thus, CNV provides a reliable and quantitative estimator of differential TE mobilization between accessions, which we used to analyze the 113 TE families that were defined as mobile based both on the split-reads approach and TE-capture (Figure 2-Source data 1).

Controlling for population stratification and considering thirteen geo-climatic variables (Hancock et al. 2011), we uncovered robust correlations with CN for 15 class I and class II TE families. Among these, *ATCOPIA2* and *ATCOPIA78* share the highest number of geo-climatic variables correlated with CN (Figure 2-figure supplement 2). Moreover, the strongest correlation is between temperature annual range and CN for *ATCOPIA78* (Figure 2A and B). Given that at least one member of this TE family is transcriptionally induced by heat shock in the Col-0 accession (Ito et al. 2011; Cavrak et al. 2014), *ATCOPIA78* provides a compelling case of a causal link between climate and TE mobilization.

We next explored the possibility of using GWASs to identify genetic variants influencing TE mobilization (see 'Materials and methods'). For 33 TE families, a disproportionately large number of SNPs are associated with CNV, preventing further analysis. For the remaining 80 TE families, SNPs in linkage disequilibrium with each other and associated with CNVs delineate 230 loci. Moreover, 34% of these loci are also identified by GWAS using whole genome sequencing data obtained for another 180 accessions taken from Sweden (Long et al. 2013) (Figure 2-figure supplement 3A). This substantial overlap suggests a similar genetic architecture for the *A. thaliana* mobilome at both the local and global scales, which prompted us to perform a joined GWAS using all 391 accessions in order to increase both sensitivity and specificity. Depending on the TE family, GWAS identified between 0 and 33 loci and collectively, associations explain between 2% and 67% of the variance in CN (Figure 2C).

Among the 334 loci detected in total, 130 encompass sites with reference or non-reference TE insertions (Figure 2C and Supplementary file 1). Furthermore, each of these local (*cis*) genetic variants explains on average 5.2% of the total variance compared with 2.2% for distal (*trans*) genetic variants. The higher explanatory power of *cis* variants is of course to be expected, as the TEs themselves are the primary determinants of the transposition process. Indeed, almost all *cis* SNPs that map to TE sequences in the reference genome are likely causal as they affect sequences involved in transposition, for example the long terminal repeats (LTRs) and the various open reading frames of LTR retrotransposons (Figure 2-figure supplement 3B and C).

While *cis* loci collectively explain most CN variance for class I TE families, this is not the case for class II TE families (Figure 2-figure supplement 3D). Given that class II TEs move by a cut and paste mechanism, some *trans* loci could in fact correspond to sites of excision. However, we could not find evidence of excision footprints, such as small insertions or deletions. Alternatively, the larger fraction of CN variance explained by *trans* loci for class II TE families may in part result from many of these families being non-autonomous, i.e. requiring factor(s) encoded by other TEs for their mobilization. Consistent with this possibility, the proportion of CN variance associated with *trans* loci as well as the number of TE annotations overlapping *trans* loci are higher for non-autonomous than autonomous class II TE families (Figure 2-figure supplement 3E). Although we did not investigate *trans* mobilization in depth, we readily identified one presumed case, involving the non-autonomous and autonomous MuDR families *ATDNA1T9A* and *VANDAL16*, respectively (Figure 2-figure supplement 3F). Of note, CN do not co-vary between these two families, which could indicate that their

transposition is differentially controlled. Finally, "false" *trans* loci could also be caused by non-reference TE insertions that are sufficiently frequent to be in linkage disequilibrium with SNPs used for the GWASs but that we have failed to detect. However, such *trans* loci are not expected to be more prevalent for class II than class I TE families and should be very rare in any case given that the probability of missing moderately frequent (>5%) non-reference TE insertions by both the split-reads approach and TE-capture is low (Figure 2-figure supplement 3G).

Since transposition is controlled by multiple protein activities in addition to those encoded by TEs, we also examined genes located within or adjacent to *trans* loci. Overall, these genes do not appear to be enriched for any particular function and most of them are specific to a single TE family (Figure 2D and Figure 2-figure supplement 3H). These observations indicate either a complex genetic architecture of mobilome variation and/or spurious *trans* associations such as those considered above. Nonetheless, 22 *trans* loci stand out as they show association with CNV for two or more TE families (Figure 2D) and a causal link is evident in two cases. Indeed, the locus associated with CNV for respectively the retrotransposon and DNA transposon families *ATGP2* and *ATENSPM2* encodes the transcription factor ARF23, which recognizes motifs that are enriched in the sequence of these TEs. The second locus is associated with CNV for the largest number of TE families (four class I and three class II families, Figure 2E) and encodes the MET2a protein, a poorly characterized homolog of the main DNA maintenance methyltransferase MET1. Moreover, one of the *MET2a* SNPs is presumably causal as it leads to a non-synonymous amino-acid substitution (G519E) in a conserved domain of the protein (Figure 2F) that in the mammalian homolog Dnmt1 is required for the targeting to replication foci (Klein et al. 2011). A role for *MET2a* is also supported by the observation that *met2a* mutant plants (Stroud et al. 2013) lose some DNA methylation exclusively over mobile TE families. Furthermore, loss of methylation is more pronounced when only considering the seven TE families that show a *MET2a* association (Figure 2G). Intriguingly, CHG sites (where H=A, T or C), which are poor substrates for MET1 or Dnmt1 compared to CG sites (Law and Jacobsen 2010), are the most affected in the *met2a* mutant. Whether or not this observation reflects an atypical recognition specificity for MET2a remains to be determined. Finally, we note that GWASs failed to detect any association with genes known to be involved in the epigenetic silencing of TEs (Ito and Kakutani 2014) such as *MET1* and *DDM1*, presumably because of their essential function.

Genome localization of newly inserted TEs

In *A. thaliana* as in many other eukaryotes, TE sequences tend to cluster in pericentromeric regions (The Arabidopsis Genome Initiative 2000). Mechanistically, such clustering may result from insertion bias, selective constraints or differential elimination of TE copies through ectopic homologous recombination (Barrón et al. 2014). To distinguish between these possibilities, we looked at the genomic location of the 2835 distinct non-reference TE insertions with TSDs detected with the split-reads approach and found that they are distributed almost evenly along chromosomes (Figure 3A). Since a similar distribution is observed for the non-reference TE insertions with TSDs detected specifically using TE capture (Figure 3-figure supplement 1A), we can rule out an ascertainment bias of the split-reads approach towards non-reference TE insertions located along the chromosome arms. However, there is a clear trend towards a more pericentromeric localization when only considering non-reference TE insertions with TSDs that are shared by two or more accessions and that are thus presumably more ancestral (Figure 3B and Figure 3-figure supplement 1B). Moreover, the density of non-reference TE insertions with TSDs positively correlates with the recombination rate but negatively with gene density (Figure 3-figure supplement 1C, D and E). Finally, except for the *COPIA* families, non-reference TE insertions with TSDs are globally under-represented within genes, where they are expected to be most detrimental (Figure 3C and Figure 3-figure supplement 1F and Figure 3-figure supplement 2A). Collectively, these observations provide strong evidence that TEs are preferentially purged over time from the chromosome arms because of their deleterious effects on adjacent genes rather than as a consequence of ectopic homologous recombination.

Although most TE families show not overt insertion bias at the genome scale, there are clear local insertion preferences. In agreement with previous observations (Fu et al. 2013; Miyao et al. 2003), private non-reference *COPIA* and *MuDR* insertions with TSDs are enriched at coding sequences and transcriptional start sites (TSS), respectively (Figure 3-figure supplement 2A). In addition, insertion sites for most TE superfamilies are enriched in specific DNA sequence motifs or exhibit biased sequence composition (Figure 3-figure supplement 2B and C; Supplementary file 2). For example, *LINEs* tend to insert within poly(A) tracks, as expected for this superfamily of non-LTR retrotransposons, which integrate into the genome via poly(A)-dependent, target site–primed reverse transcription (TPRT; Beck et al. 2011).

Impact of newly inserted TEs on the expression of adjacent genes

Transcriptome analyses in the reference accession Col-0 have revealed that *A.thaliana* genes nearest to TE

sequences are expressed at lower levels compared with the genome-wide distribution of gene expression, suggesting that TE insertions tend to reduce the expression of neighboring genes (Hollister and Gaut 2009). To investigate more directly the impact of TEs on the genes within or near which they insert, we examined RNA-seq data available for 144 accessions (Schmitz et al. 2013). Specifically, we considered all non-reference TE insertions with TSDs and calculated for each gene located within 1kb of them (1616 genes in total), the ratio between the expression level in the accession(s) harboring the insertion and the median expression level in the accessions devoid of the insertion. Expression ratios expected under the null hypothesis (no effect of the TE insertions) were calculated by taking 10^6 randomly chosen sets of 1616 genes and assigning for each set the TE insertion 'presence/absence' label randomly among the 144 accessions (see 'Materials and Methods'). Comparison of the distribution of the observed and expected expression ratios indicates that for a large fraction of genes, expression is indeed significantly altered when TEs insert within or near them (Figure 3D, $P<1.9\times10^{-5}$). These alterations are most pronounced for the *COPIA* insertions, which are overrepresented in genes and less pronounced for the *MuDR* insertions, despite the latter being overrepresented around the TSS of genes (Figure 3E and Figure 3-figure supplement 2A). Although other TE superfamilies show similar trends, we could not draw firm conclusions in these cases because of insufficient statistical power (Figure 3-figure supplement 3). This notwithstanding, it is clear that TE insertions induce both increases and decreases in gene expression with equal frequency (Figure 3D and E). Thus our findings contradict the prevailing view of a dampening effect of TE insertions on the expression of adjacent genes (Hollister and Gaut 2009) and suggest instead a stronger selection against TE insertions when they occur close to highly expressed genes.

To complement the re-analysis of transcriptome data, we also measured by RT-qPCR the expression level of 19 genes with recent *COPIA* or *MuDR* insertions, using nine different accessions grown under control conditions or subjected to heat shock. *COPIA* insertions were found to have more dramatic and systematic effects on gene expression in stressed plants (Figure 3-figure supplement 4) which in the case of *ATCOPIA78* can be related to its transcriptional sensitivity to heat shock (Ito et al. 2011; Cavrak et al. 2014). On the other hand, we could not detect any effect of the MuDR insertions under the two conditions tested (Figure 3-figure supplement 4). These findings are in agreement with those of the transcriptome analysis and indicate in addition that the effect of TE insertions on the expression of adjacent genes can vary substantially in relation to the environment.

Impact of newly inserted TEs on the DNA methylation status of adjacent sequences

291 TE sequences are typically targeted by the RNA-directed DNA methylation (RdDM) machinery in *A. thaliana*
292 (Lippman et al. 2004; Lister et al. 2008; Cokus et al. 2008) and we have previously provided genome-wide
293 evidence that DNA methylation can spread from RdDM targets to flanking sequences, with consequences on
294 gene expression (Ahmed et al. 2011). To investigate the effect of new TE insertions on the DNA methylation
295 status of adjacent sequences, we used MethylC-Seq data available for 140 accessions (Schmitz et al. 2013).
296 Analysis of this data set first indicated that mobile TE families have on average higher CG, CHG and CHH
297 methylation than non-mobile TE families (Figure 4A). Furthermore, DNA methylation is also higher for most
298 mobile TE families in the accessions that have experienced recent transposition activity (Figure 4-figure
299 supplement 1A). These observations prompted us to examine in addition methylome data obtained for several
300 mutation accumulation (MA) lines (Becker et al. 2011; Schmitz et al. 2011). We found that mobile TE families
301 suffer less sporadic DNA methylation loss than non-mobile families (Figure 4B). These findings are entirely
302 consistent with DNA methylation playing an important role in the control of TE mobility and suggest in turn
303 that most of the recent TE insertions we have identified are present in the methylated state. Moreover, given
304 that DNA methylation is likely established over newly inserted TE copies by RdDM in a progressive manner
305 across multiple generations (Teixeira et al. 2009; Marí-Ordóñez et al. 2013), unmethylated non-reference TE
306 insertions should be mainly private and reflect very recent transposition events.
307 Based on these considerations, we next analyzed the DNA methylation status of 1543 TE insertion sites for
308 which reliable data could be extracted across all 140 accessions (Figure 4C). Approximately 10% of insertion
309 sites are methylated in most accessions, including systematically in the one(s) containing the TE. As expected,
310 these sites are preferentially located within TE-rich, pericentromeric regions. In contrast, another 40% of sites
311 are devoid of methylation in the accession(s) containing the TE insertion as well as in most of the other
312 accessions. This absence of adjacent DNA methylation could indicate either that the TE insertions themselves
313 are unmethylated or else that DNA methylation does not spread from them. Finally, 50% of sites are methylated
314 exclusively or almost exclusively in the accession(s) with the TE insertion (Figure 4-figure supplement 1B),
315 thus suggesting that at these sites TEs are methylated and that DNA methylation spread into adjacent sequences.
316 Why some sites may be refractory to DNA methylation spreading when others are not is unclear, as we did not
317 identify any feature that could distinguish them, such as the identity of the TE or the sequence composition at
318 the insertion site.
319 Further analysis of DNA methylation associated with TE insertions indicates that it involves the three sequence

contexts and that it generally extends for up to 300 bp on both sides of the insertion (Figure 4D and E; Figure 4-figure supplement 1B), a distance that closely matches that previously reported for the spreading of DNA methylation from RdDM targets (Ahmed et al. 2011). For 243 insertion sites however, DNA methylation extends over much longer distances (up to 3.5 kb) on one or the other side of the insertion (Figure 4D and E). While most of these sites lie within or close to genes, the TE insertions are not preferentially orientated with respect to gene transcription (Figure 4F), which rules out sense-antisense transcription as the likely trigger for this long-distance DNA methylation. Proximity to pericentromeric heterochromatin can also be ruled out, because most of the genes with long-distance DNA methylation are located on the chromosome arms (Figure 4-figure supplement 1C). To explore potential mechanisms further, we made used of the wealth of epigenomic data available in Col-0 to examine the 142 Col-0 TE insertions with TSDs that are absent from the assembled Ler-1 genome sequence (Figure 1-Source data 4). Methylome data (Stroud et al. 2013) indicate that 121 of these 142 Col-0 TE copies are methylated and that DNA methylation tends to extend into flanking sequences, predominantly over short distances, but occasionally over much longer distances (<300pb: 63 TE insertions; >1kb: 36 TE insertions; Figure 1-Source data 4). These results confirm those obtained for the non-reference TE insertions. In addition, analysis of Col-0 small RNA-seq data (Fahlgren et al. 2009) indicates that in contrast to short-distance DNA methylation, long-distance DNA methylation aligns with 24-nt siRNAs (Figure 4-figure supplement 1D). Thus, genes affected by the latter type of DNA methylation have presumably become secondary targets of RdDM, as was shown for a transgene (Kanno et al. 2008; Daxinger et al. 2009). Moreover, genes affected by long-distance DNA methylation in accessions other than Col-0 tend in this accession, where they are by definition in the ancestral state, to be most highly expressed in both pollen and seeds and more highly expressed in these two organs than genes affected by short-distance methylation (Figure 4G). Given that RdDM activity is also maximal in these organs (Teixeira and Colot 2010), our observations suggest that secondary RdDM results from the concomitance of strong transcription and strong RdDM at target loci.

Finally, our analysis of TE-associated DNA methylation indicates that it accounts for at least 7% of the so-called gene C-DMRs (i.e. regions of differential methylation at CG, CHG and CHH sites) identified in nature, which are typically low frequency gain of DNA methylation variants (Schmitz et al. 2013). These and similar findings reported recently (Stuart et al. 2016) confirm and extend previous results that first indicated that many natural gene C-DMRs are not *bona fide* epialleles but rather new alleles caused by TE insertions (Schmitz et al.

2013). However, close examination of one TE-insertion allele shared among 13 accessions indicates that it is subject to epigenetic variation in nature, as it is present in the unmethylated state in one accession (Figure 4-figure supplement 1B).

TE insertions as motors of adaptive changes

Although TEs tend to insert with no overt bias at the genome scale (Figure 3A), we detected nineteen 10kb windows with a high load of non-reference TE insertions (Figure 5A). Such enrichment could result from insertion preference or else an absence of strong negative selection. In fact, three of these 10kb windows span genes encoding nucleotide-binding domain and leucine-rich repeat containing (NLR) proteins, which function as immune receptors in plants and are known to be under strong diversifying selection (Chae et al. 2014). Moreover, a fourth 10kb window spans the gene *FLC*, which encodes a key repressor of flowering and is one of the main genetic factors causing natural variation in the onset of flowering (Ietswaart, Wu, and Dean 2012). Remarkably, the *FLC* locus has the highest number of non-reference TE insertions (seven in total) across the genome. These insertions belong to several COPIA families and affect four distinct *FLC* haplotypes in total (Figure 5A and Figure 5-figure supplement 1A). Moreover, five insertions are located within the first intron (Figure 5B), which plays an important role in the epigenetic regulation of *FLC* in response to cold (Ietswaart, Wu, and Dean 2012). Although four of these insertions as well as another intronic insertion were previously identified among early flowering accessions (Liu et al. 2004; Lempe et al. 2005), causality could not be established unequivocally because of numerous other sequence polymorphisms in complete linkage disequilibrium. To obtain direct proof of a causal role for the seven TE insertions we identified, we used publically available transcriptomic (Schmitz et al. 2013) as well as phenotypic data (Li et al. 2010; Lempe et al. 2005) and compared *FLC* expression as well as flowering time among accessions that have the same *FLC* haplotype but differ by the presence or absence of a TE insertion (see 'Material and Methods'). Results of these comparisons indicate that the TE-containing accessions have systematically much reduced *FLC* expression and flower much earlier than their TE-free counterparts (Figure 5C and D; Figure 5-figure supplement 1B and C). Thus, we can conclude that TEs are recurrent generators of major effect *FLC* alleles, which in turn suggests that they contribute significantly to the high level of allelic heterogeneity observed at this locus (Li et al. 2014)

DISCUSSION

We have shown that the *A. thaliana* mobilome is particularly rich at the species level, being composed of at least half of the 326 TE families that are annotated in the reference genome. This finding is at odds with the prevailing view that most TE families are mere molecular fossils in *A. thaliana*, since they contain a much lower proportion of "young", i.e. non-degenerated TE copies than in the close relative *A. lyrata* (Hu et al. 2011; Maumus and Quesneville 2014). Furthermore, we provide definitive evidence that TEs insert throughout the genome, with no overt bias towards pericentromeric regions, which contrasts with the observed clustering of annotated TE sequences around centromeres. However, these discrepancies are easily resolved, since we have also shown that despite the richness of the *A. thaliana* mobilome, most TEs tend to be rapidly purged by natural selection in this species when they insert in the chromosome arms, which are gene-rich. Indeed, our systematic survey indicates that TEs have pervasive effects on the expression and DNA methylation status of genes near or within which they insert. Incidentally, the deleterious effects associated with most transposition events in *A. thaliana* may also explain in part the fact that we did not detect any recent transposition bursts, as these should be strongly counter-selected. Furthermore, because *A. thaliana* is predominantly self-fertilizing, the purging of deleterious TE insertions should be accelerated in this species compared to *A. lyrata*, which is an obligated out-crosser. Given this difference in mating systems, the TE population dynamics of *A. thaliana* and *A. lyrata* are expected to differ significantly (Lockton and Gaut 2010). Thus, homologous recombination could play a more prominent role in the elimination of TE insertions in *A. lyrata*, as is thought to be the case in *D. melanogaster* (Barrón et al. 2014). However, comprehensive studies similar to those presented here remain to be performed for *A. lyrata* in order to identify conclusively the forces that shape the TE landscape in this species.

We have also shown that the composition and activity of the mobilome vary greatly between accessions. GWASs revealed that this variation is caused in part by sequence polymorphisms within the TEs themselves (*cis* variation), which is in agreement with empirical data and theoretical models indicating that TE families contain only one or a few active, autonomous (i.e. master) TE copies at any one time (Beck et al. 2011). The fact that we could readily detect such *cis* variants may again be linked to the mating system of *A. thaliana*, which on the one hand should increase the probability that disabling mutations accumulate within the few active TE copies that are present within a given lineage, before these copies could transpose further; and on the other hand should decrease the probability of acquiring new active copies through crosses.

Another important result of the GWASs is that natural variation at the *MET2a* locus, which encodes a poorly characterized DNA methyltransferase, has a significant impact on mobilome composition and activity across

accessions, being associated with differential transposition activity for seven distinct class I and class II TE families. While the role of MET2a in transposition control remains to be determined experimentally, it is noteworthy that none of the epigenetic repressors of TE activity identified through genetic screens, such as MET1 or DDM1, are associated with natural variation of the mobilome, presumably because of their essential function. Altogether, these findings illustrate the power of GWASs in identifying the genetic factors affecting transposition in nature.

Being a complex trait, TE mobilization is also modulated by environmental factors, and we have identified temperature annual range as a clear contributor to the variation in *ATCOPIA78* mobilization across accessions. Remarkably, this TE family has generated several rare alleles with large effects at the *FLC* locus, which is a major genetic determinant of the onset of flowering in nature. It is therefore tempting to speculate that *ATCOPIA78* may endow *A. thaliana* with a unique ability to adapt to global warming and the associated increase in droughts by facilitating the creation of early flowering *FLC* alleles. Additionally, these observations may provide insights into how *A. thaliana* has been able to colonize efficiently the entire northern hemisphere from a few glacial refugia located in Southern Europe (François et al. 2008).

In summary, our findings have far reaching implications, as they indicate that part of the missing heritability that plagues many GWASs may be accounted for by recent and thus rare TE insertion alleles with large effects (Vinkhuyzen et al. 2013; Brachi, Morris, and Borevitz 2011). More generally, our study highlights the need for similar species-wide explorations of the mobilome in a variety of organisms in order to assess the true mutational and epimutational impact of transposition as well as its contribution to natural phenotypic variation. In this respect, it can be anticipated that the advent of long read sequencing technologies will greatly facilitate such studies, especially in organisms with large, repeat-rich genomes.

MATERIALS AND METHODS

Data sources. Whole genome sequencing data (Illumina short paired-end reads) for *A. thaliana* accessions were obtained from the NCBI SRA archive for 210 accessions collected worldwide (Schmitz et al. 2013) SRA012474 as well as for the 180 Swedish accessions (Long et al. 2013; SRA052536). Illumina short paired-end reads from Ler-1 (Schneeberger et al. 2011) and the assembled Ler-1genome obtained using PacBio long reads (unpublished) were retrieved from http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/current/Ler-1/Reads/ and https://github.com/

PacificBiosciences/DevNet/wiki/Arabidopsis-P5C3, respectively. High quality SNP imputations for these accessions were obtained from the 1001 Genomes Project (http://1001genomes.org/). Processed MethylC-seq and RNA-seq data from accessions collected worldwide (Schmitz et al. 2013) were obtained from NCBI GEO (GSE43857 and GSE43858, respectively). Processed BS-seq data for the *met2a* mutant (Stroud et al. 2013) were obtained from NCBI GEO (GSE39901). Flowering time data were retrieved from the phenotypic information deposited at https://www.arabidopsis.org/servlets/TairObject?type=germplasm&id=6530472136 as well as (Y. Li et al. 2010; Lempe et al. 2005).

Local climate data and coordinates for the 211 accessions analyzed using CNV were retrieved from (Hancock et al. 2011) (http://bergelson.uchicago.edu/wp-content/uploads/2015/04/allvars948_notnormd_011311.txt.zip). TE-capture sequencing data have been deposited in the ENA short read archive under project number PRJEB11706 and accessions codes ERR1121179 to ERR1121190.

CNV analysis. We performed an aggregated CNV analysis for each TE family by considering only annotated TE sequences longer than 300bp as a joined pseudo-annotation. Families composed in the Col-0 reference genome exclusively of TE sequences shorter than 350bp (seven families, *AR12*, *AR3*, *ATCLUST1*, *ATDNATA1*, *ATMSAT1*, *ATSINE1* and *ATSINE3*), that contain too many (>1000) copies or that overlap with genomic regions with abnormal coverage (four families, *ATREP15*, *ATREP10D*, *ATREP3* and *HELITRONY3*; see next section) were not considered. Illumina pair-end short reads were mapped onto the TAIR10 reference genome using Bowtie2 (using the arguments --mp 13 --rdg 8,5 --rfg 8,5 --very-sensitive) and PCR-duplicates were removed using Picard. Read coverage (RC) was computed in non-overlapping windows of 100bp spanning the joined pseudo-annotation for each TE family. RC over each bin was corrected for GC content (Yoon et al. 2009) and normalized to the genome-wide RD. Normalized RC across joined pseudo-annotations were compared between each accession and two independent re-sequenced Col-0 reference genomes (Schmitz et al. 2013; Long et al. 2013). CNVs were detected by performing a distribution-independent permutation test for each of the two references. In order to be as stringent as possible, the maximum *P*-value from these two comparisons were considered. These *P*-values were compared with an empirical null-distribution of *P*-values constructed by randomizing a million times the TE annotation labels over the 100bp windows. We defined the false discovery rate (FDR) of an observed *P*-value as the fraction of significant (that is, below the observed *P*-value in question) hits found in the randomized set. We considered statistically significant TE-CNVs when FDR was below 10^{-5}.

Filtering out genomic regions with aberrant coverage or low sequence complexity. Read mapping to the reference *A. thaliana* genome sequence revealed several regions with very high coverage, which correspond mainly to the 180bp centromeric repeat unit, the 45S and 5S rDNA repeat units, ATHILA TE sequences, telomeres and plastid DNA-like sequences. These as well as the few regions not suitable for sequence alignment because of low complexity were identified by mapping genome sequencing reads obtained for two independent lines of the Col-0 accession (Schmitz et al. 2013; Long et al. 2013) onto the TAIR10 genome sequence. RC was calculated on consecutive non-overlapping windows of 100bp across the entire genome. After correction for GC content (Yoon et al. 2009), consecutive windows (allowing one window gap) with a RC greater or lower than three median absolute deviations from the median RC signal were merged to define larger segments. These sequences spanned 1,125,487 bp (0.94% of the TAIR10 genome sequence) and were excluded from all analyses.

Identification of non-reference TE insertions with TSDs using split-reads. The split-read analysis was performed in four steps: (i) extraction of reads not mapping to the reference genome; (ii) forced mapping to a collection of TE sequence extremities and soft-clipping of mapped reads; (iii) mapping of clipped reads to the reference genome; (iv) identification of clipped reads that reveal target site duplications (TSDs). Briefly, for each accession we retrieved reads that do not map to the TAIR10 reference genome (containing the SAM flag 4), but we did not make use of information about discordant paired-end reads, unlike in the Jitterbug approach (Hénaff et al. 2015). Unmapped reads were then aligned (using Bowtie2 in --local mode to allow for soft clip alignments) to a collection of 5' and 3' TE sequence extremities (300bp) obtained from the Col-0 for each TE family or from Repbase Update (Jurka et al. 2005) in the case of the ARNOLDY2, ATCOPIA62, ATCOPIA95, TA12, TAG1 families, which do not contain copies with intact extremities in the Col-0 genome. Next, we selected all reads with one end (≥20nt) mapping to a TE extremity (by locating reads where the CIGAR string contained only one 'S' character with a value equal or greater to 20). These reads were recursively soft clipped by 1nt and mapped to the TAIR10 reference genome using Bowtie2 (using the arguments --end-to-end --mp 13 --rdg 8,5 --rfg 8,5 --very-sensitive) until the soft-clipped read length reached 20nt. Read clusters composed of five or more reads clipped from the same extremity and overlapping with read clusters composed of reads clipped from the other extremity were taken to indicate the presence of a bona fide TE insertion only if the size

494 of the overlap was equal or less than 2-fold larger than that reported for TSDs for the corresponding TE family.
495 Putative non-reference TE insertions overlapping genomic regions with aberrant coverage (as defined above) or
496 located within inner pericentromeres (Clark et al. 2007) or spanning the corresponding donor TE sequence were
497 filtered out. Presence or absence of these putative non-reference TE insertions was verified in other accessions
498 by relaxing the parameters used to detect them in the first place. Specifically, we asked for the presence of a
499 minimum of two rather than ten soft-clipped reads spanning the corresponding TSD coordinates. This improved
500 the discovery of putative TE-insertions that are shared by more than one accession.
501
502 **Estimation of sensitivity.** The rate of false negatives was estimated indirectly by first searching for TE
503 insertions with TSDs that are present in the Col-0 reference genome but absent from the Ler-1 assembled
504 genome sequence recently obtained using PacBio long-reads (unpublished). Specifically, annotated TE
505 sequences together with 1kb of upstream and downstream sequences were aligned using BLAT to the Ler-1
506 genome sequence. Heavily truncated TE sequences (covering less than 50% of full-length copies) were not
507 considered, as most of these are unlikely to have been mobilized recently. Out of the 7931 TE sequences
508 analyzed, 200 are not present in the Ler-1 genome sequence and among these, 142 have TSDs (Table S2). This
509 value was then compared to that obtained by mapping Col-0 Illumina short split-reads to the Ler-1 genome
510 sequence. A total of 64 TE insertions with TSDs could not be detected using that approach, thus giving a false
511 negative rate of 45% (Figure 1-figure supplement 1C). Furthermore, 15 TE insertions were called that are in
512 fact not present in the Col-0 genome sequence, thus giving a false positive rate of 10%; (i.e. FDR of 15.3%;
513 Figure 1-figure supplement 1C).
514
515 **Manual filtering of residual false positives.** We analyzed two re-sequenced Col-0 genomes (Schmitz et al.
516 2013; Long et al. 2013) using our split-read pipeline. Non-reference TE insertions with TSDs detected in other
517 accessions and also identified in at least one of the two Col-0 genomes were filtered out, as they are most likely
518 errors. Additionally, after manual inspection using the Integrative Genomics Viewer (IGV) (Robinson et al.
519 2011), we eliminated all putative non-reference TE insertions with TSDs for which read-clusters comprise a
520 disproportionately high number of reads (more than three MADs over the median whole genome coverage) as
521 well as all putative non-reference TE insertions with TSDs that are present in 50% of accessions, as these are
522 frequently the result of sequencing errors or mapping artifacts. Indeed, the pattern of 5' and 3' split-reads at the

insertion site differs radically between false and *bona fide* non-reference TE insertions, which enable their discrimination by visual inspection (Figure 1-figure supplement 2). We also observed a high rate of non-reference TE insertions with TSDs within TE sequences, mostly in the case of the ATHILA and GYPSY families. Visual inspection of all these insertions indicates that they most likely result from mapping artifacts and they were therefore excluded manually. Finally, we examined 12 non-reference TE insertions with TSDs by PCR and all were validated (Supplementary file 3).

TE sequence capture.

A complete description of our TE-capture design and protocol will be published elsewhere. Briefly 2.1M biotinylated capture probes (Roche-NimbleGen) were designed to cover 200bp at each end of 310 potentially mobile TEs (Supplementary file 4), belonging to 181 of the 326 TE families annotated in TAIR10. These 181 TE families include 117 of the 131 mobile TE families identified using the split-read approach as well as most other TE families for which non-degenerate and thus potentially mobile copies are present in the Col-0 genome. Mobile or potentially mobile TE families with skewed sequence composition or very high copy number (including all ATHILA families) were excluded from the design in order to avoid overt biases in the capture of TE sequences. Twelve DNA sequence libraries for the accessions Pi-0, Bl-1, Sp-0, Ta-0, Lip-0, Bla-1, Stw-0, Cvi-0, Kondara, Jm-0, Gre-0 and Rubezhnoe-1 were prepared following the Illumina TruSeq paired-end kit, with some modifications. One µg of genomic DNA was sheared to a median fragment size of 400-450bp. Fragments were subjected to end repair, A-tailing and index adapter ligation following the manufacturer's instructions. Libraries were then amplified through 7 cycles of ligation-mediated PCR using the KAPA HiFi Hot Start Ready Mix and primers AATGATACGGCGACCACCGAGA and CAAGCAGAAGACGGCATACGAG at a final concentration of 2µM. PCR products were cleaned-up using the Qiaquick PCR Purification kit following the manufacturer's instructions, except that DNA was eluted in PCR grade water. Amplified DNA libraries were then pooled and one µg of the multiplex pool was used in the first hybridization step to capture probes (72 hours). Captured DNA was recovered using Streptavidin Dynabeads, washed and PCR amplified (5 cycles) as above. Amplified captured DNA was then subjected to a second round of hybridization (12 hours), recovery and amplification (14 cycles). Hybridization, captured DNA recovery and washing were performed as described in the NimbleGen SeqCapture EZ protocol. Pre-capture and post-second capture PCR products were run on an Agilent Bioanalyser DNA 1000 chip.

552 Enrichment for capture TE sequences was confirmed by qPCR and estimated at 50-100 fold depending on the
553 TE. Pair-end sequencing was performed with Illumina HiSeq 2000 and 100bp reads. Between 10.7 and 16.1
554 million pairs were sequenced per library. Pairs were mapped to the TAIR10 reference genome using Bowtie2
555 with the arguments --mp 13 --rdg 8,5 --rfg 8,5 --very-sensitive -X 1000. Enrichment factor was ~250 fold on
556 average (Figure 1-figure supplement 1E). Discordant pairs were remapped using Bowtie2 with the parameters -
557 -mp 13 --rdg 8,5 --rfg 8,5 -D 20 -R 10 -N 0 -L 15 -i S,1,0.50 -k 100. TE insertions were detected in each
558 accession using discordantly mapped reads and the algorithm Hydra (Quinlan et al. 2010) using the arguments:
559 pairDiscordants: -d 50000 –z 1000 –x0 –r 100; dedupDiscordants: -s3; hydra: -ms 50 –li –use all –mld (10*mad)
560 –mno (median+(20*mad). Comparing the results of TE-capture with those obtained using our split-read
561 pipeline gives respectively false negative and positive rates of 56% and 7% (i.e. FDR of 14%; Figure 1-figure
562 supplement 1E). These values are similar to those obtained using the assembled genome sequence of Ler-1 and
563 confirm the low sensitivity of the split-read pipeline.
564
565 **GWAS of CNVs.** Unlike for the initial CNV analysis, we considered this time only the annotated TE sequences
566 that are at least half the size of the corresponding reference TE sequence (obtained from the *A. thaliana*
567 RepeatMasker repeat library, http://repeatmasker.org) as this improved significantly the correlation between
568 CNV and the number of non-reference TE insertions. For each of the 113 families identified as mobile by both
569 the split-reads approach and TE-capture, aggregated CNV was obtained for each accession by summing the
570 CNVs estimated for each annotated copy. GWAS on this CNV values was carried out with imputed SNPs
571 (MAF > 5%) from the 211 accessions collected worldwide (Schneeberger et al. 2011; Schmitz et al. 2013), the
572 180 Swedish accessions (Long et al. 2013) and the joined dataset using a linear mixed model (LMM) (Kang et
573 al. 2010). Kinship matrix was included in the model as a random effect to control population structure. Out of
574 the 113 TE families, 33 displayed a genome inflation factor (GIF) greater than 1.10 and were excluded from
575 subsequent analysis. A conservative threshold value (P-value < 1×10^{-8}) was set to call statistically associated
576 SNPs. Proximal associated SNPs in linkage disequilibrium (r^2 > 0.2) were identified using Plink (Purcell et al.
577 2007) and combined in blocks to build statistical associated intervals, which were expanded by 1 or 5kb on
578 either side. Variance explained by the leading SNP within each locus was calculated using the following
579 equation: $\sigma^2_{SNP} = 2(MAF)(1-MAF)\frac{\beta^2}{\sigma^2_y}$, in which MAF is the minor allele frequency, β is the SNP effect

estimated by the LMM and σ_y^2 is the variance of the phenotype Y. Total variance explained by *cis* and *trans* loci was computed as the sum of the single-locus explained variances, under the assumption of additive contributions. The probability P(f) of missing a non-reference TE insertion with TSD underlying a "false" *trans* locus through both the split reads and TE-capture approaches was calculated using the following equation: $P(f) = FN \times \binom{n}{12-k} \times f^{12-k} \times (1-f)^k$ where f is the frequency of the minor allele for the *trans* locus under consideration (reported in the Supplementary File 1), FN is the false negative rate for the split-reads bioinformatic pipeline (0.56), n is the number of accessions analyzed by TE-capture (12) and k is the number of accessions without the non-reference TE insertion with TSD among the 12 accessions analyzed by TE-capture (k ranges from 0 to 12 and calculations were all performed using k=12). .

Climate analysis. We selected 12 geo-climatic variables representing different ecological layers: Aridity, number of frosty days, number of consecutive frost-free days, day length in the spring, maximum temperature in the warmest month, minimum temperature in the coldest month, temperature annual range, photosynthetically active radiation, precipitation in the wettest month, precipitation in the driest month, relative humidity in the summer, landscape slope and thermal (Hancock et al. 2011). CNs for the 113 families confirmed as mobile by the split-read approach and TE-capture were used to calculate a partial Mantel correlation with the 13 geo-climatic variables. Population structure kinship was included in the test to control population stratification. Partial Mantel tests were conducted using the "ecodist" package in R. A threshold of $P < 0.01$, corrected for multiple testing (8.33×10^{-4}), was set to call statistically correlated variables. In addition to using the partial Mantel test, we also applied linear models to regress CN as a function of the climatic variables. Although this method does not control for population structure, it largely confirmed the associations found by the partial Mantel test.

Characterization of *cis* and *trans* associations. Protein coding genes, miRNA genes, ncRNAs and TE annotations were retrieved from (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff). GWAS intervals were defined as *cis* associations if they overlap with a TE annotation or a non-reference TE insertion with TSDs of the same family. All other GWAS intervals were defined as *trans* associations. The *cis* associations are over-represented 34 times (P-value < 1×10^{-16}) when compared to randomly chosen genomic

609 intervals of the same size, which is consistent with TE activity being primarily determined by the TE sequence
610 itself. All genomic annotations overlapping with *trans* intervals (within 1 or 5kb) were considered as being
611 putatively causal.
612
613 **Analysis of the localization of non-reference TE insertions with TSDs along chromosomes.** To assess if
614 non-reference TE insertions with TSDs are enriched in pericentromeric regions, their number within these
615 regions was compared with that expected from a random distribution. Insertions with different allele counts
616 (private, shared by 2-10 accessions, shared by >10 accessions) were considered separately. TE distribution in
617 the reference genome (TAIR10) was obtained by counting the number of TE sequences located within
618 pericentromeres (minus genomic regions showing aberrant coverage and inner pericentromeres). The expected
619 distribution for the 2835 non-reference TE insertions with TSDs was calculated by randomizing 10^6 times their
620 position across the chromosomes (genomic regions showing coverage deviation, the inner pericentromeres, or
621 coordinates spanning the corresponding donor TE sequence were excluded). This set of random positions was
622 used as a control for all subsequent analyses. Insertion distribution over genes and neighboring sequences was
623 performed using a meta-gene. Briefly, protein coding gene features were extracted from the TAIR10 annotation
624 and coordinates of non-reference TE insertions with TSDs were crossed with the set of genic features according
625 to the following stepwise hierarchy: 5' UTR > 3' UTR> exon > intron > intergenic regions. For insertions that
626 do not overlap protein-coding genes, the distance to the closest gene was calculated and reported as negative or
627 positive distance according to the gene orientation. Expected insertion distribution under the null hypothesis
628 was retrieved by applying this procedure for each of the 10^6 randomized sets of insertions. To assess if non-
629 reference TE insertions with TSDs are enriched within small clusters, we divided the genome into 10kb non-
630 overlapping windows and we counted the number of insertions events within them. The observed and expected
631 (random) densities of non-reference TE insertions with TSDs per window were compared and significant
632 enrichment was declared when the number of insertions found within a window was in the upper 0.005% tail of
633 the random distribution.
634
635 **Reconstruction of the historical recombination landscape**. Historical recombination was estimated using
636 LDhat (McVean et al. 2004) as described before (Choi et al. 2013). Briefly, biallelic SNPs (MAF $\geq$ 0.1) from
637 the 210 accessions collected worldwide were selected and split into blocks of 5,000 SNPs with overlap of 500

638 SNPs. SNPs located in inner centromeres were excluded from the analysis. Blocks of SNPs were formatted
639 using the 'convert' program. A likelihood lookup table was generated for 210 individuals with program
640 'complete' using the following parameters: –n 210 –rhomax 100 –n_pts 100 –theta 0.001. Population-scaled
641 recombination rate (ρ/kb, ρ = 4Ner, where Ne is the effective population size and r is the per-generation
642 recombination rate) was calculated using the 'interval' program with the following parameters: -its 60000000 –
643 bpen 5 –sam 40000. Recombination rates for contiguous blocks were joined at overlap position 250.
644 Population-scaled recombination rate map is provided in Supplementary file 5.
645
646 **Identification of DNA sequence motifs overrepresented at non-refrence TE insertion sites with TSDs.**
647 Sequence spanning non-reference insertion sites were analyzed using Bioprospector Release 2 (Liu, Brutlag,
648 and Liu 2001) using the following parameters: -r 1 -n 200 -a 1 -W 'TSD size' . Background sequence
649 distribution for the reference genome was obtained using the "genomebg" program. Sequence logo was
650 produced using the seqLogo package version 1.36.0.
651
652 **Assessing the impact of non-reference TE insertions with TSDs on gene expression**. All genes with a non-
653 reference TE insertion with TSDs within 1kb were retrieved and their expression analyzed using transcriptome
654 data available for 144 accessions (Schmitz et al. 2013). For each gene, we calculated the ratio between the
655 median gene expression level for the accessions harboring the TE insertion and the median gene expression
656 level for accessions lacking that insertion. Distribution plots of observed gene expression ratios were compared
657 to the expected distribution under the null hypothesis (random effect). This expected distribution was obtained
658 by calculating the gene expression ratio for 10^6 randomly chosen sets of genes for which the TE insertion
659 presence/absence "label" was randomly assigned between the accessions. Statistical significance of differences
660 between the observed and expected distributions was determined using the Kolmogorov-Smirnov test.
661
662 **Expression levels of genes affected by non-reference insertions with TSDs.** RNA was extracted using the
663 RNeasy plant mini kit (Qiagen from plants grown under normal conditions (10 days old seedlings grown in
664 liquid medium) or subjected to heat shock treatment (Ito et al. 2011). RT-qPCR was performed as described
665 previously (Silveira et al. 2013). Primers details are given in Supplementary file 3. RT-qPCR results (one
666 biological replicate only) are indicated relative to those obtained for a gene (*AT5G13440*) that shows invariant

667 expression under multiple conditions.

668

669 **Gene expression profile of genes affected by long-distance DNA methylation.** Expression data in the Col-0
670 accession were obtained from http://www.weigelworld.org/resources/
671 microarray/AtGenExpress/AtGE_dev_gcRMA.txt.zip/at_download/file for the 224 and 162 genes affected
672 respectively by short- and long-distance DNA methylation in accessions with non-reference TE insertions with
673 TSDs (by definition, these TE insertions are absent in Col-0). Triplicate data for each developmental time point
674 was averaged and then normalized across the developmental time-point series. Average expression level was
675 then calculated for each time point for all genes affected by short-distance DNA methylation and compared to
676 the average calculated for all genes affected by long-distance DNA methylation. Statistical significance of
677 differences between these two averages was calculated using the non-parametric Mann-Whitney U test.

678

679 **Determination of *FLC* haplotypes**. Haplotype analysis was performed as described previously (Li et al. 2014).
680 Briefly, SNPs within 100 kb of *FLC* were retrieved for the 211 worldwide accessions and used as input into
681 fastPHASE version 1.4.0 (Scheet and Stephens 2006). Default parameters were kept, except for the -Pzp option.
682 For each SNP, haplotype membership with the highest likelihood was assigned.

683

684 **Code availability.** Source code for the split-read pipeline can be accessed at
685 https://github.com/LeanQ/SPLITREADER
686

687

688 **REFERENCES AND NOTES**

689

690 Ahmed, Ikhlak, Alexis Sarazin, Chris Bowler, Vincent Colot, and Hadi Quesneville. 2011. "Genome-Wide
691 Evidence for Local DNA Methylation Spreading from Small RNA-Targeted Sequences in Arabidopsis."
692 *Nucleic Acids Research* 39 (16): 6919–31. doi:10.1093/nar/gkr324.

693 Baillie, J Kenneth, Mark W Barnett, Kyle R Upton, Daniel J Gerhardt, Todd A. Richmond, Fioravante De Sapio,
694 Paul M. Brennan, et al. 2011. "Somatic Retrotransposition Alters the Genetic Landscape of the Human
695 Brain." *Nature* 479 (7374): 534–37. doi:10.1038/nature10531.

696 Barrón, Maite G., Anna-Sophie Fiston-Lavier, Dmitri a. Petrov, and Josefa González. 2014. "Population
697 Genomics of Transposable Elements in Drosophila." *Annual Review of Genetics* 48 (1): 561–81.
698 doi:10.1146/annurev-genet-120213-092359.

699 Beck, Christine R, José Luis Garcia-Perez, Richard M Badge, and John V Moran. 2011. "LINE-1 Elements in
700 Structural Variation and Disease." *Annu Rev Genomics Hum Genet* 12: 187–215. doi:10.1146/annurev-

genom-082509-141802.

Becker, Claude, Jörg Hagmann, Jonas Müller, Daniel Koenig, Oliver Stegle, Karsten Borgwardt, and Detlef Weigel. 2011. "Spontaneous Epigenetic Variation in the *Arabidopsis thaliana* Methylome." *Nature* 480 (480): 245–52. doi:10.1038/nature10555.

Bennetzen, Jeffrey L, and Hao Wang. 2014. "The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes." *Annual Review of Plant Biology* 65. Annual Reviews: 505–30. doi:10.1146/annurev-arplant-050213-035811.

Brachi, Benjamin, Geoffrey P Morris, and Justin O Borevitz. 2011. "Genome-Wide Association Studies in Plants: The Missing Heritability Is in the Field." Genome Biology 12 (10): 232. doi:10.1186/gb-2011-12-10-232.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, et al. 2011. "Whole-Genome Sequencing of Multiple *Arabidopsis thaliana* Populations." *Nature Genetics* 43 (10): 956–63. doi:10.1038/ng.911.

Cavrak, Vladimir V, Nicole Lettner, Suraj Jamge, Agata Kosarewicz, Laura Maria Bayer, and Ortrun Mittelsten Scheid. 2014. "How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation." *PLoS Genet* 10 (1): e1004115. doi:10.1371/journal.pgen.1004115.

Chae, Eunyoung, Kirsten Bomblies, Sang-Tae Kim, Darya Karelina, Maricris Zaidem, Stephan Ossowski, Carmen Martín-Pizarro, et al. 2014. "Species-Wide Genetic Incompatibility Analysis Identifies Immune Genes as Hot Spots of Deleterious Epistasis." *Cell* 159: 1341–51. doi:10.1016/j.cell.2014.10.049.

Choi, Kyuha, Xiaohui Zhao, Krystyna a Kelly, Oliver Venn, James D Higgins, Nataliya E Yelina, Thomas J Hardcastle, et al. 2013. "Arabidopsis Meiotic Crossover Hot Spots Overlap with H2A.Z Nucleosomes at Gene Promoters." *Nature Genetics* 45 (11): 1327–36. doi:10.1038/ng.2766.

Clark, Richard M., Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, et al. 2007. "Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*." *Science* 317: 338–42. doi:10.1126/science.1138632.

Cokus, Shawn J, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. 2008. "Shotgun Bisulphite Sequencing of the Arabidopsis Genome Reveals DNA Methylation Patterning." *Nature* 452 (7184): 215–19. doi:10.1038/nature06745.

Cridland, Julie M., Stuart J. Macdonald, Anthony D. Long, and Kevin R. Thornton. 2013. "Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources." *Molecular Biology and Evolution* 30 (10): 2311–27. doi:10.1093/molbev/mst129.

Daxinger, Lucia, Tatsuo Kanno, Etienne Bucher, Johannes van der Winden, Ulf Naumann, Antonius J M Matzke, and Marjori Matzke. 2009. "A Stepwise Pathway for Biogenesis of 24-Nt Secondary siRNAs and Spreading of DNA Methylation." *The EMBO Journal* 28 (1): 48–57. doi:10.1038/emboj.2008.260.

Dubin, Manu J, Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J Osborne, Francesco Paolo Casale, Philipp Drewe, et al. 2015. "DNA Methylation in Arabidopsis Has a Genetic Basis and Shows Evidence of Local Adaptation." *eLife* 4: e05255. doi:10.7554/eLife.05255.

Elbarbary, Reyad A., Bronwyn A. Lucas, and Lynne E. Maquat. 2016. "Retrotransposons as Regulators of Gene Expression." *Science* 351 (6274): aac7247. doi:10.1126/science.aac7247.

Fahlgren, Noah, Christopher M Sullivan, Kristin D Kasschau, Elisabeth J Chapman, Tyler W H Backman, Scott A Givan, and James C Carrington. 2009. "Computational and Analytical Framework for Small RNA Profiling by High-Throughput Sequencing." *RNA*, no. 15: 992–1002. doi:10.1261/rna.1473809.

François, Olivier, Michael G. B. Blum, Mattias Jakobsson, and Noah A. Rosenberg. 2008. "Demographic History of European Populations of *Arabidopsis thaliana*." *PLoS Genetics* 4 (5): e1000075. doi:10.1371/journal.pgen.1000075.

Fu, Yu, Akira Kawabe, Mathilde Etcheverry, Tasuku Ito, Atsushi Toyoda, Asao Fujiyama, Vincent Colot, Yoshiaki Tarutani, and Tetsuji Kakutani. 2013. "Mobilization of a Plant Transposon by Expression of the

749 Transposon-Encoded Anti-Silencing Factor." *The EMBO Journal* 32 (17): 2407–17.
750 doi:10.1038/emboj.2013.169.

751 Hancock, Angela M, Benjamin Brachi, Nathalie Faure, Matthew W Horton, Lucien B Jarymowycz, F Gianluca
752 Sperone, Chris Toomajian, Fabrice Roux, and Joy Bergelson. 2011. "Adaptation to Climate Across the
753 *Arabidopsis thaliana* Genome." *Science* 334 (6052): 83–86. doi:10.1126/science.1209244.

754 Heard, Edith, and Robert A. Martienssen. 2014. "Transgenerational Epigenetic Inheritance: Myths and
755 Mechanisms." *Cell* 157 (1): 95–109. doi:10.1016/j.cell.2014.02.045.

756 Hénaff, Elizabeth, Luís Zapata, Josep M. Casacuberta, and Stephan Ossowski. 2015. "Jitterbug: Somatic and
757 Germline Transposon Insertion Detection at Single-Nucleotide Resolution." *BMC Genomics* 16 (1): 768.
758 doi:10.1186/s12864-015-1975-5.

759 Hill, W G, and Alan Robertson. 1966. "The Effect of Linkage on Limits to Artificial Selection." *Genetics*
760 *Research* 8 (03): 269–94. doi:10.1017/S0016672300010156.

761 Hollister, Jesse D, and Brandon S Gaut. 2009. "Epigenetic Silencing of Transposable Elements: A Trade-off
762 between Reduced Transposition and Deleterious Effects on Neighboring Gene Expression." *Genome*
763 *Research* 19 (8): 1419–28. doi:10.1101/gr.091678.109.

764 Hu, Tina T, Pedro Pattyn, Erica G Bakker, Jun Cao, Jan-Fang Cheng, Richard M Clark, Noah Fahlgren, et al.
765 2011. "The *Arabidopsis lyrata* Genome Sequence and the Basis of Rapid Genome Size Change." *Nature*
766 *Genetics* 43 (5): 476–81. doi:10.1038/ng.807.

767 Ietswaart, Robert, Zhe Wu, and Caroline Dean. 2012. "Flowering Time Control: Another Window to the
768 Connection between Antisense RNA and Chromatin." *Trends in Genetics* 28 (9): 445–53.
769 doi:10.1016/j.tig.2012.06.002.

770 Ito, Hidetaka, Etienne Bucher, Marie Mirouze, Isabelle Vaillant, and Jerzy Paszkowski. 2011. "An siRNA
771 Pathway Prevents Transgenerational Retrotransposition in Plants Subjected to Stress." *Nature* 472 (7341):
772 115–19. doi:10.1038/nature09861.

773 Ito, Hidetaka, and Tetsuji Kakutani. 2014. "Control of Transposable Elements in *Arabidopsis thaliana*."
774 *Chromosome Research* 22 (2): 217–23. doi:10.1007/s10577-014-9417-9.

775 Joly-Lopez, Zoé, and Thomas E. Bureau. 2014. "Diversity and Evolution of Transposable Elements in
776 Arabidopsis." *Chromosome Research* 22 (2): 203–16. doi:10.1007/s10577-014-9418-8.

777 Jurka, Jerzy, Vladimir V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. 2005. "Repbase
778 Update, a Database of Eukaryotic Repetitive Elements." *Cytogenetic and Genome Research* 110 (1-4):
779 462–67.

780 Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-yee Kong, Nelson B. Freimer, Chiara
781 Sabatti, and Eleazar Eskin. 2010. "Variance Component Model to Account for Sample Structure in
782 Genome-Wide Association Studies." *Nature Genetics* 42 (4): 348–54. doi:10.1038/ng.548.

783 Kanno, Tatsuo, Etienne Bucher, Lucia Daxinger, Bruno Huettel, Gudrun Böhmdorfer, Wolfgang Gregor, David
784 P Kreil, Marjori Matzke, and Antonius J M Matzke. 2008. "A Structural-Maintenance-of-Chromosomes
785 Hinge Domain–containing Protein Is Required for RNA-Directed DNA Methylation." *Nature Genetics* 40
786 (5): 670–75. doi:10.1038/ng.119.

787 Kazazian, Haig H. 2004. "Mobile Elements: Drivers of Genome Evolution." *Science* 303 (5664): 1626–32.
788 doi:10.1126/science.1089670.

789 Klein, Christopher J, Maria-Victoria Botuyan, Yanhong Wu, Christopher J Ward, Garth a Nicholson, Simon
790 Hammans, Kaori Hojo, et al. 2011. "Mutations in DNMT1 Cause Hereditary Sensory Neuropathy with
791 Dementia and Hearing Loss." *Nature Genetics* 43 (6): 595–600. doi:10.1038/ng.830.

792 Kofler, Robert, Viola Nolte, and Christian Schlötterer. 2015. "Tempo and Mode of Transposable Element
793 Activity in Drosophila." *PLoS Genetics* 11 (7): e1005406. doi:10.1371/journal.pgen.1005406.

794 Law, Julie A., and Steven E. Jacobsen. 2010. "Establishing , Maintaining and Modifying DNA Methylation
795 Patterns in Plants and Animals." *Nature Reviews Genetics* 11 (3): 204–20. doi:10.1038/nrg2719.

796 Lempe, Janne, Sureshkumar Balasubramanian, Sridevi Sureshkumar, Anandita Singh, Markus Schmid, and
797 Detlef Weigel. 2005. “Diversity of Flowering Responses in Wild *Arabidopsis thaliana* Strains.” *PLoS*
798 *Genetics* 1 (1): e6. doi:10.1371/journal.pgen.0010006.

799 Li, Peijin, Daniele Filiault, MS Mathew S Box, Envel Kerdaffrec, Cock van Oosterhout, Amity M Wilczek,
800 Johanna Schmitt, et al. 2014. “Multiple FLC Haplotypes Defined by Independent Cis-Regulatory
801 Variation Underpin Life History Diversity in *Arabidopsis thaliana*.” *Genes & Development* 28 (15):
802 1635–40. doi:10.1101/gad.245993.114.

803 Li, Y., Y. Huang, J. Bergelson, M. Nordborg, and J. O. Borevitz. 2010. “Association Mapping of Local Climate-
804 Sensitive Quantitative Trait Loci in *Arabidopsis thaliana*.” *Proceedings of the National Academy of*
805 *Sciences* 107 (49): 21199–204. doi:10.1073/pnas.1007431107.

806 Lippman, Zachary, Anne-Valerie Gendrel, Michael Black, Matthew W Vaughn, Neilay Dedhia, W Richard
807 McCombie, Kimberly Lavine, et al. 2004. “Role of Transposable Elements in Heterochromatin and
808 Epigenetic Control.” *Nature* 430 (6998): 471–76. doi:10.1038/nature02651.

809 Lisch, Damon. 2013. “How Important Are Transposons for Plant Evolution?” *Nature Reviews Genetics* 14 (1):
810 49–61. doi:10.1038/nrg3374.

811 Lister, Ryan, Ronan C O Malley, Julian Tonti-filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and
812 Joseph R Ecker. 2008. “Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis.”
813 *Cell* 133 (3): 523–36. doi:10.1016/j.cell.2008.03.029.

814 Liu, Jun, Yuehui He, Richard Amasino, and Xuemei Chen. 2004. “siRNAs Targeting an Intronic Transposon in
815 the Regulation of Natural Flowering Behavior in Arabidopsis.” *Genes & Development*, no. 732: 2873–78.
816 doi:10.1101/gad.1217304.

817 Liu, X, D L Brutlag, and J S Liu. 2001. “Bioprospector: Discovering Conserved DNA Motifs in Upstream
818 Regulatory Regions of Co-Expressed Genes.” *Pacific Symposium on Biocomputing* 6: 127–38.

819 Lockton, Steven, and Brandon S Gaut. 2010. “The Evolution of Transposable Elements in Natural Populations
820 of Self-Fertilizing *Arabidopsis thaliana* and Its Outcrossing Relative *Arabidopsis lyrata*.” *BMC*
821 *Evolutionary Biology* 10: 10. doi:10.1186/1471-2148-10-10.

822 Long, Quan, Fernando a Rabanal, Dazhe Meng, Christian D Huber, Ashley Farlow, Alexander Platzer, Qingrun
823 Zhang, et al. 2013. “Massive Genomic Variation and Strong Selection in *Arabidopsis thaliana* Lines from
824 Sweden.” *Nature Genetics* 45 (8): 884–90. doi:10.1038/ng.2678.

825 Mackay, Trudy F. C., Stephen Richards, Eric a. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu,
826 Sònia Casillas, et al. 2012. “The *Drosophila melanogaster* Genetic Reference Panel.” *Nature* 482 (7384):
827 173–78. doi:10.1038/nature10811.

828 Marí-Ordóñez, Arturo, Antonin Marchais, Mathilde Etcheverry, Antoine Martin, Vincent Colot, and Olivier
829 Voinnet. 2013. “Reconstructing de Novo Silencing of an Active Plant Retrotransposon.” *Nature Genetics*
830 45 (9): 1029–39. doi:10.1038/ng.2703.

831 Maumus, Florian, and Hadi Quesneville. 2014. “Ancestral Repeats Have Shaped Epigenome and Genome
832 Composition for Millions of Years in *Arabidopsis thaliana*.” *Nature Communications* 5 (May): 4104.
833 doi:10.1038/ncomms5104.

834 McClintock, Barbara. 1956. “Intranuclear Systems Controlling Gene Action and Mutation.” In *Brookhaven*
835 *Symposia in Biology*, 58–74.

836 McVean, Gilean A T, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. 2004.
837 “The Fine-Scale Structure of Recombination Rate Variation in the Human Genome.” *Science* 304 (5670):
838 581–84. doi:DOI: 10.1126/science.1092500.

839 Miyao, Akio, Katsuyuki Tanaka, Kazumasa Murata, Hiromichi Sawaki, Shin Takeda, Kiyomi Abe, Yoriko
840 Shinozuka, Katsura Onosato, and Hirohiko Hirochika. 2003. “Target Site Specificity of the Tos17
841 Retrotransposon Shows a Preference for Insertion Withn Genes and against Insertion in Retrotransposon-
842 Rich Regions of the Genome.” *The Plant Cell* 15 (August): 1771–80. doi:10.1105/tpc.012559.ements.

843 Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian

844 Maller, Pamela Sklar, Paul I W De Bakker, and Mark J Daly. 2007. “PLINK: A Tool Set for Whole-
845 Genome Association and Population-Based Linkage Analyses.” *The American Journal of Human Genetics*
846 81 (3): 559–75. doi:10.1086/519795.

847 Quadrana, Leandro, and Vincent Colot. 2016. “Plant Transgenerational Epigenetics.” *Annual Review of*
848 *Genetics* 50 (1): null. doi:10.1146/annurev-genet-120215-035254.

849 Quinlan, Aaron R, Royden a Clark, Svetlana Sokolova, Mitchell L Leibowitz, Yujun Zhang, Matthew E Hurles,
850 Joshua C Mell, and Ira M Hall. 2010. “Genome-Wide Mapping and Assembly of Structural Variant
851 Breakpoints in the Mouse Genome.” *Genome Research* 20 (5): 623–35. doi:10.1101/gr.102970.109.

852 Rahman, Reazur, Gung-Wei Chirn, Abhay Kanodia, Yuliya A. Sytnikova, Björn Brembs, Casey M. Bergman,
853 and Nelson C. Lau. 2015. “Unique Transposon Landscapes Are Pervasive across *Drosophila*
854 *melanogaster* Genomes.” *Nucleic Acids Research* 43 (22): 10655–72. doi:10.1093/nar/gkv1193.

855 Rebollo, Rita, Mark T Romanish, and Dixie L Mager. 2012. “Transposable Elements: An Abundant and Natural
856 Source of Regulatory Sequences for Host Genes.” *Annual Review of Genetics* 46 (January): 21–42.
857 doi:10.1146/annurev-genet-110711-155621.

858 Richardson, Sandra R, Santiago Morell, and Geoffrey J Faulkner. 2014. “L1 Retrotransposons and Somatic
859 Mosaicism in the Brain.” *Annual Review of Genetics* 48 (1): 1–27. doi:10.1146/annurev-genet-120213-
860 092412.

861 Richardson, Sandra R., Aurélien J. Doucet, Huira C. Kopera, John B. Moldovan, José Luis Garcia-Perez, and
862 John V. Moran. 2015. “The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes.”
863 *Microbiology Spectrum* 3 (2): 1–26. doi:10.1128/microbiolspec.MDNA3-0061-2014.

864 Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and
865 Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.
866 doi:10.1038/nbt0111-24.

867 Scheet, Paul, and Matthew Stephens. 2006. “A Fast and Flexible Statistical Model for Large-Scale Population
868 Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase.” *The American*
869 *Journal of Human Genetics* 78 (4): 629–44. doi:10.1086/502802.

870 Schmitz, Robert J., Matthew D. Schultz, Mathew G. Lewsey, Ronan C. O’Malley, Mark A. Urich, Ondrej
871 Libiger, Nicholas J. Schork, and Joseph R. Ecker. 2011. “Transgenerational Epigenetic Instability Is a
872 Source of Novel Methylation Variants.” *Science* 334 (6054). American Association for the Advancement
873 of Science: 369–73. doi:10.1126/science.1212959.

874 Schmitz, Robert J., Matthew D. Schultz, Mark A. Urich, Joseph R. Nery, Mattia Pelizzola, Ondrej Libiger,
875 Andrew Alix, et al. 2013. “Patterns of Population Epigenomic Diversity.” *Nature* 495 (7440): 193–98.
876 doi:10.1038/nature11968.

877 Schneeberger, Korbinian, Stephan Ossowski, Felix Ott, Juliane D Klein, Xi Wang, Christa Lanz, Lisa M Smith,
878 et al. 2011. “Reference-Guided Assembly of Four Diverse *Arabidopsis thaliana* Genomes.” *Proceedings*
879 *of the National Academy of Sciences* 108 (25): 10249–54. doi:10.1073/pnas.1107739108.

880 Silveira, Amanda Bortolini, Charlotte Trontin, Sandra Cortijo, Joan Barau, Luiz Eduardo Vieira Del Bem,
881 Olivier Loudet, Vincent Colot, and Michel Vincentz. 2013. “Extensive Natural Epigenetic Variation at a
882 de Novo Originated Gene.” *PLoS Genetics* 9 (4): e1003437. doi:10.1371/journal.pgen.1003437.

883 Slotkin, R Keith, and Robert Martienssen. 2007. “Transposable Elements and the Epigenetic Regulation of the
884 Genome.” *Nature Reviews. Genetics* 8 (4): 272–85. doi:10.1038/nrg2072.

885 Stroud, Hume, Maxim V.C. Greenberg, Suhua Feng, Yana V. Bernatavichute, and Steven E. Jacobsen. 2013.
886 “Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis
887 Methylome.” *Cell* 152 (1-2): 352–64. doi:10.1016/j.cell.2012.10.054.

888 Stuart, Tim, Steven R Eichten, Jonathan Cahn, Justin Borevitz, and Ryan Lister. 2016. “Population Scale
889 Mapping of Transposable Element Diversity Reveal Link to Gene Regulation and Epigenetc Variation.”
890 *bioRxiv* 0897: 0–3. doi:10.1101/039511.

891 Teixeira, Felipe K, and Vincent Colot. 2010. “Repeat Elements and the Arabidopsis DNA Methylation

892 Landscape." *Heredity* 105 (1). Nature Publishing Group: 14–23. doi:10.1038/hdy.2010.52.

893 Teixeira, Felipe K, Fabiana Heredia, Alexis Sarazin, François Roudier, Martine Boccara, Constance Ciaudo,
894 Corinne Cruaud, et al. 2009. "A Role for RNAi in the Selective Correction of DNA Methylation Defects."
895 *Science* 323 (5921): 1600–1604. doi:10.1126/science.1165313.

896 The Arabidopsis Genome Initiative. 2000. "Analysis of the Genome Sequence of the Flowering Plant
897 *Arabidopsis thaliana*." *Nature* 408 (6814): 796–815. doi:10.1038/35048692.

898 Trono, Didier. 2016. "Transposable Elements, Polydactyl Proteins, and the Genesis of Human-Specific
899 Transcription Networks." *Cold Spring Harbor Symposia on Quantitative Biology* LXXX (ii).
900 doi:10.1101/sqb.2015.80.027573.

901 Tsay, Yi Fang, Mary J. Frank, Tania Page, Caroline Dean, and Nigel M. Crawford. 1993. "Identification of a
902 Mobile Endogenous Transposon in *Arabidopsis thaliana*." *Science* 260 (5106): 342–44.
903 doi:10.1126/science.8385803.

904 Vinkhuyzen, Anna a.E., Naomi R. Wray, Jian Yang, Michael E. Goddard, and Peter M. Visscher. 2013.
905 "Estimation and Partition of Heritability in Human Populations Using Whole-Genome Analysis Methods."
906 *Annual Review of Genetics* 47 (1): 75–95. doi:10.1146/annurev-genet-111212-133258.

907 Weigel, Detlef, and Vincent Colot. 2012. "Epialleles in Plant Evolution." *Genome Biology* 13 (10): 249.
908 doi:10.1186/gb-2012-13-10-249.

909 Weigel, Detlef, and Magnus Nordborg. 2015. "Population Genomics for Understanding Adaptation in Wild
910 Plant Species." *Annual Review of Genetics* 49 (1): 315–38. doi:10.1146/annurev-genet-120213-092110.

911 Yoon, Seungtai, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. 2009. "Sensitive and
912 Accurate Detection of Copy Number Variants Using Read Depth of Coverage." *Genome Research* 19 (9):
913 1586–92. doi:10.1101/gr.092981.109.

914 Ziolkowski, Piotr A., Grzegorz Koczyk, Lukasz Galganski, and Jan Sadowski. 2009. "Genome Sequence
915 Comparison of Col and Ler Lines Reveals the Dynamic Nature of Arabidopsis Chromosomes." *Nucleic*
916 *Acids Research* 37 (10): 3189–3201. doi:10.1093/nar/gkp183.

917
918
919 **Acknowledgements.** The authors declare no competing financial interest, except GFM and JAJ who
920 declare a competing interest as employees of Roche NimbleGen Inc. We thank members of the Colot
921 lab and especially Mathilde Etcheverry for discussions. We thank Edith Heard and Pierre Capy for
922 critical reading of an earlier version of the manuscript. This work was supported by the European
923 Union Seventh Framework Programme Network of Excellence EpiGeneSys (Award 257082, to VC),
924 the Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE, ANR-11-IDEX-0001-02 PSL*
925 Research University and ANR-12-ADAP-0020-01 (to VC) and the Chaire Blaise Pascal (to RAM).
926 LQ was the recipient of postdoctoral fellowships from the ANR-10-LABX-54 MEMO LIFE and
927 ANR-11-IDEX-0001-02 PSL* Research University. ABS was the recipient of postdoctoral
928 fellowships from the ANR-12-ADAP-0020-01 and from the Brazilian National Council for Scientific
929 and Technological Development (CNPq). LQ and VC conceived and designed the study. LQ

performed all of the bioinformatic analyses. ABS, GFM and JAJ designed the TE-sequence capture, ABS performed the experiment and CL and RAM performed the sequencing of captured fragments. ABS performed all of the other experimental analyses. LQ and VC wrote the paper, with additional input from all authors. Correspondence and requests for materials should be addressed to colot@biologie.ens.fr.

FIGURES LEGENDS

Figure 1. Overview of the *A. thaliana* mobilome. (**A**) Genome browser tracks showing normalized sequencing coverage over the two full-length *ATCOPIA31* elements annotated in the reference genome (Col-0). CNV is detected as increased or decreased coverage in other accessions. Number of copies is indicated on the right. (**B**) Heat map representing CNVs (log2 ratio) for 317 TE families and 211 *A. thaliana* accessions. TE families with statistically significant CNV in at least one accession are indicated. Figure 1-Source data 1 contains absolute copy number estimation of TE sequences. (**C**) Schematic representation of the bioinformatics pipeline to identify non-reference TE insertions with TSD using split-reads. 1- Reads are mapped on a collection of TE extremities from annotated TE sequences and reference sequences (Repbase update). 2- Reads aligning partially over TE extremities are extracted and clipped. 3- The unmapped portion of these split-reads are re-mapped on the Arabidopsis reference genome. 4- Non-reference TE insertions with TSDs are identified by searching for overlapping clusters of 5' and 3' split-reads. (**D**) Genome browser tracks showing split-reads for two non-reference *ATCOPIA31* insertions and TSD reconstruction. Figure 1-Source data 2 contains the coordinates of all non-reference TE insertions with TSDs. (**E**) Distribution frequency of allele counts for non-reference TE insertions with TSDs. (**F**) Number of mobile TE families per accession identified using split-read and TE-sequence capture. (**G**) Cumulative plot of the number of mobile TE families detected with increasing numbers of accessions. (**H**) The total number of non-reference TE insertions with TSDs is indicated in relation to the number of accessions with such insertions, for each of the 131 mobile TE families. Asterisks indicate the nine TE families with experimental evidence of transposition (Ito and Kakutani 2014; Tsay et al. 1993). Figure 1-Source data 3 contains the total number of distinct non-reference TE insertions with TSD for each TE family and super-family. Figure 1-figure supplement 1 shows TE-capture results. Figure 1- figure supplement 2 contains IGV screenshots showing the pattern of split-reads characteristic of true- and false-positive non-

reference TE insertions with TSDs.

Figure 2. Environmental and genetic factors associated with differential mobilome activity. (**A**) Copy number (CN, red circles) of *ATCOPIA78* in accessions distributed across the globe. Annual temperature range is also shown. (**B**) Partial Mantel correlation between *ATCOPIA78* CN and annual temperature range. (**C**) Fraction of CNV variance explained by SNPs (*cis,* and *trans)* and partial Mantel correlation with geo-climatic variables. (**D**) Distribution of *cis* and *trans* loci in the joined analysis (391 accessions) and number of TE families associated with a given *trans* locus. A complete list of the GWAS results is provided in Supplementary file 1. (**E**) Manhattan plots displaying GWAS results for the seven TE families with a *MET2a* association. The leading SNP within each interval is indicated as a red diamond. Colors indicate the extent of linkage disequilibrium (r^2) with the leading SNP. (**F**) Schematic view of the MET2a protein (TD: targeting domain; BAH: bromo adjacent homology domain) and sequence alignment of the TD. The amino acid substitution (G519E) that is present in some accessions is indicated (red arrow). (**G**) Average DNA methylation level over non-mobile, mobile and *MET2a*-associated TE families in WT and *met2a* Col-0 seedlings (Stroud et al. 2013). Statistically significant differences are indicated (MWU test). Figure 2-figure supplement 1 shows the positive correlation between CN and number of non-reference TE insertions with TSDs. Figure 2-figure supplement 2 shows climate association to CNVs. Figure 2-figure supplement 3 shows GWAS results for CNVs.

Figure 3. Genomic localization of non-reference TE insertions. (**A**) Density of non-reference TE insertions with TSDs (blue) and of annotated TE sequences (red) along the reference sequence of chromosome 1. Inner pericentromeric regions are masked. (**B**) Fraction of private and shared non-reference TE insertions with TSDs and of annotated TE sequences in outer pericentromeric regions. Statistically significant differences are indicated (chi square test). (**C**) Observed/expected ratio (O/E) of private non-reference TE insertions with TSDs in and around genes. Errors bars are defined as 95% confidence intervals. (**D**) Cumulative distribution of gene expression ratios between alleles harboring and lacking non-reference TE insertions. Statistically significant differences were calculated using the KS test. (**E**) As D, but only for COPIA (green) or MuDR (red) non-reference TE insertions with TSDs. Figure 3-figure supplement 1 shows detailed analysis of the distribution of non-reference TE insertions with TSDs along the genome. Figure 3-figure supplement 2 shows local TE insertion preferences. Figure 3-figure supplement 3 shows global expression levels of gene affected by non-

reference TE insertions. Figure 3-figure supplement 4 shows expression levels of genes affected in some accessions by a non-reference insertion with TSD in plants grown under control conditions or subjected to heat stress.

Figure 4. DNA methylation of non-reference TE insertion sites. (**A**) Boxplot representation of average DNA methylation level for mobile and non-mobile TE families across all accessions. (**B**) O/E ratio of spontaneous DMRs identified in mutation accumulation lines (Becker et al. 2011; Schmitz et al. 2011) for non-mobile and mobile TE families. Statistically significant differences were calculated using a chi square test. (**C**) Average DNA methylation level in 50bp windows upstream and downstream of 1543 insertions sites for accessions lacking or containing a given non-reference TE insertion with TSD. **(D**) Genome browser tracks showing examples of insertion sites respectively associated with short- and long-distance DNA methylation. (**E**) Meta-analysis of DNA methylation around non-reference TE insertions sites. (**F**) Distribution of non-reference TE insertions associated with short- or long-distance DNA methylation according to their position relative to genes (stacked bar plot) and proportion of insertions in the two possible orientations relative to the closest gene (pie charts). (**G**) Average expression level in different organs and at different developmental time points (in Col-0) of genes with non-reference TE insertions with TSDs and affected by short- (blue) or long-distance (red) DNA methylation. Error bars are s.e.m. Statistical significance of differences was calculated using a MWU test. Figure 4-figure supplement 1 shows DNA methylation of TE families and impact on sequences flanking non-reference TE insertions with TSDs.

Figure 5. Local enrichment of non-reference TE insertions with TSDs. (A) Density of non-reference TE insertions with TSDs in 10kb windows. The 19 regions statistically enriched in such insertions are indicated by red bars. (**B**) Position and identity of the seven non-reference TE insertions with TSDs spanning the *FLC* locus. (**D**) and (**E**) Level of *FLC* expression (**D**) and flowering time (**E**) for accessions of same *FLC* haplotype but differing by the presence or absence of the relevant TE insertion. Errors bars are s.e.m. Figure 5-figure supplement 1 shows the reconstruction of the *FLC* haplotypes and additional analyses of the effect on flowering timeof non-reference TE insertions with TSDs at the locus.

Figure 1-figure supplement 1. Validation of the *A. thaliana* mobilome by TE-capture. (**A**) The number of non-reference TE insertions with TSDs identified by the split-read pipeline is plotted against the corresponding genome sequencing coverage for each accession. Accessions analyzed by TE-capture are highlighted in red. (**B**) Genome browser tracks showing examples of non-reference TE insertions identified by TE-capture only. (**C**) Overlap between TE insertions with TSDs identified specifically in Col-0 using the Ler-1 genome assembly as a reference and either whole genome sequence alignment or the split-reads pipeline. The percentage of false positives (FP), true positives (TP) and false negatives (FN) as well as the false discovery rate (FDR) are indicated. (**D**) Description of the TE-capture design and workflow. (**E**) TE-capture enrichment of target sequences. (**F**) Overlap between non-reference insertions with TSDs identified by split-read analysis and TE-capture. The percentage of FP, TP and FN as well as the FDR are indicated. (**G**) Distribution frequency of allele counts for non-reference TE insertions identified using the split-read approach and TE-capture among the 12 accessions analysed. (**H**) Number of SNPs plotted against the number of non-reference TE insertions identified by TE-capture between any two accessions.

Figure 1-figure supplement 2. Visual inspection of true- and false-positive non-reference TE insertions with TSDs. IGV screenshots showing split-reads for non-reference TE insertions with TSDs that are validated or not by TE-capture (true- and false-positives, respectively). Split-reads are shown for 12 different accessions. Accessions containing the non-reference TE insertion with TSD are indicated in red.

Figure 2-figure supplement 1. Pearson correlation between TE CN and number of TE sequences identified by TE capture.

Figure 2-figure supplement 2. Climate association to TE CNV. Heat map representing partial Mantel correlation coefficient between TE CN and geo-climatic variables. TE families with statistically significant correlations ($P < 8.33\times10^{-4}$) are indicated.

Figure 2-figure supplement 3. **GWAS of CNVs**. (**A**) Overlap between GWAS results obtained using CNVs and SNPs from world-wide accessions and from Swedish accessions. (**B**) Manhattan plot of the GWAS results for *ATCOPIA69* CNVs. The leading SNP (red diamond) is located within the TE itself. Colors indicate the

extent of linkage disequilibrium (r^2) to the leading SNP. Distribution of CN values associated with the leading SNP and the common allele are shown on the right. (**C**) Schematic overview of *ATCOPIA69*. LTR, long terminal repeat; gag, nucleocapsid protein; pro, protease. The sequence alignment on the right indicates the position of the amino acid change (red arrow) caused by the leading SNP in the transposase protein. Conserved amino acids are highlighted in blue. (**D**) Average fraction of CN variance explained by *cis* and *trans* loci for Class I and Class II TE families. (**E**) Average fraction of CN variance explained by *cis* and *trans* loci for autonomous and non-autonomous class II TE families. Ratio of the observed over expected (O/E) number of TE annotations overlapping *trans* loci for autonomous and non-autonomous class II TE families. Statistically significant differences were calculated by resampling 10,000 times the coordinates of the *trans* loci. (**F**) Manhattan plot of the GWAS results for *ATDNA1T9A*. Distribution of CN values associated with the leading SNP and the common allele are shown on the right. Note that both *ATDNA1T9A* and *VANDAL16* show similar insertion preference towards the TSS of genes. (**G**) Probability of missing a non-reference TE insertion with TSD as a function of the allele frequency of the *trans* locus identified by GWAS. (**H**) Gene ontology of genes overlapping or close to *trans* loci.

Figure 3-figure supplement 1. Distribution of non-reference TE insertions with TSDs along the genome. (**A**) Density of non-reference TE insertions with TSDs detected by split-reads (black) or TE-sequence capture (red) across the five chromosomes. (**B**) Fraction of non-reference TE insertions with TSDs detected in pericentromeric regions using TE-sequence capture as a function of allele frequency (f). (**C**) Historical recombination landscape estimated using genome sequencing data for 211 accessions. (**D**) Density of non-reference TE insertions with TSDs as a function of the density of coding sequences. (**E**) Density of non-reference TE insertions with TSDs as a function of the recombination rate. The correlation between these two variables still holds after correcting for the partial correlation between gene density and recombination rate (r=0.32, $P < 4^{-23}$). (**F**) Distribution frequency of allele counts for non-reference TE insertions with TSDs located either within or close to genes or away from genes.

Figure 3-figure supplement 2. Local TE insertion preferences. (**A**) Metagene analysis of the distribution of private non-reference TE insertions with TSDs for four Class II and two Class I TE superfamilies. UTR, untranslated transcribed region. (**B**) Sequence motifs for non-reference insertion sites. (**C**) GC-content for non-

reference insertion sites (including 50bp upstream and downstream). Blue bars represent the GC content for 100bp-long sequences randomly chosen from the reference genome sequence. Bars represent average GC-content ± SD. Statistically significant differences were calculated using a permutation test

Figure 3-figure supplement 3. Global expression levels of genes affected by non-reference TE insertions of different TE superfamilies. Cumulative distribution of gene expression ratios between alleles harboring and lacking non-reference TE insertions for the different TE superfamilies. The number of genes analyzed is indicates in each case. Statistically significant differences between the observed and expected distributions of expression ratios were calculated using a KS test.

Figure 3-figure supplement 4. Expression levels of selected genes affected by non-reference TE insertions with TSDs. Accession(s) with a non-reference TE insertion with TSD are indicated in red for each gene. Plants were grown under standard conditions (Ctrl.) or subjected to a heat shock (HS). RT-qPCR results (three technical replicates) are indicated relative to those obtained for a gene that shows invariant expression under multiple conditions (see 'Material and Methods').

Figure 4-figure supplement 1. DNA methylation of TE families and impact on sequences flanking non-reference TE insertions with TSDs. (**A**) Boxplot representation of average DNA methylation levels for non-mobile TE families (across all accessions) and for mobile TE families (separately for accessions with or without evidence of mobility). (**B**) Genome browser tracks for one insertion site. DNA methylation and split-reads are indicated whenever present. Two accessions (Zdr-1 and Knox-18) have DNA methylation (red arrows) but lack split reads supportive of the presence of the non-reference TE insertion (black crosses). Another accession (Anholt-1) has no DNA methylation (red cross) yet contains the non-reference TE insertion (black arrow). Another 12 accessions have the non-reference TE insertion and all have DNA methylation at the insertion site. (**C**) Expected density distribution of the fraction of non-reference TE insertions with TSDs that are located in the chromosome arms and are associated with long-distance DNA methylation. The observed fraction is indicated by the vertical line. (**D**) Genome browser tracks showing the density of 24-nt siRNAs over two Col-0 TE insertions with TSDs that are associated with long- and short-distance DNA methylation, respectively (top and bottom panels). Brackets indicate the absence of the insertions in the other accessions.

Figure 5-figure supplement 1. Reconstruction of the *FLC* haplotypes and additional analyses of the effect on flowering time of non-reference TE insertions with TSDs at the locus. (**A**) SNPs identify 20 distinct haplotypes at the *FLC* locus (+/- 50Kb). The seven non-reference TE insertions with TSDs (indicated A to G, as in Figure 5B) located within *FLC* affect four distinct haplotypes, as shown on the right. (**B**) and (**C**) Flowering time (Y. Li et al. 2010; Lempe et al. 2005) associated with *FLC* alleles belonging to the same haplotype but differing by the presence or absence of a non-reference TE insertion with TSD.

Figure 1-Source data 1. **Copy number estimation of TE sequences.** (**A**) Copy number estimation based on read coverage for the 317 TE families analyzed across 211 *A. thaliana* accessions collected worldwide. Column descriptions are provided in (**B**).

Figure 1-Source data 2. **Coordinates of non-reference TE insertions with TSDs**. (**A**) Coordinates and presence or absence call (1 and 0, respectively) across the 211 *A. thaliana* accessions. Description of columns is provided in (**B**).

Figure 1-Source data 3. **Number of distinct non-reference TE insertions with TSDs identified by the split-reads approach for each TE family and super-family.**

Figure 1-Source data 4. **TE insertions with TSDs present in Col-0 but absent in Ler-1**. (**A**) Genomic coordinates of the insertion in Col-0 and of the corresponding empty site in Ler-1. Description of columns is provided in (**B**).

Figure 2-Source data 1. **Copy number estimation used for the geo-climatic associations and GWASs**. (**A**) Copy number estimation based on read coverage for the 131 mobile TE families analyzed across 211 *A. thaliana* accessions collected worldwide. (**B**) Copy number estimation based on read coverage for the same 131 mobile TE families across 180 *A. thaliana* accessions from Sweden.

Supplementary file 1. Summary of GWAS results for CNV. Distribution of CN values, Manhattan plot and

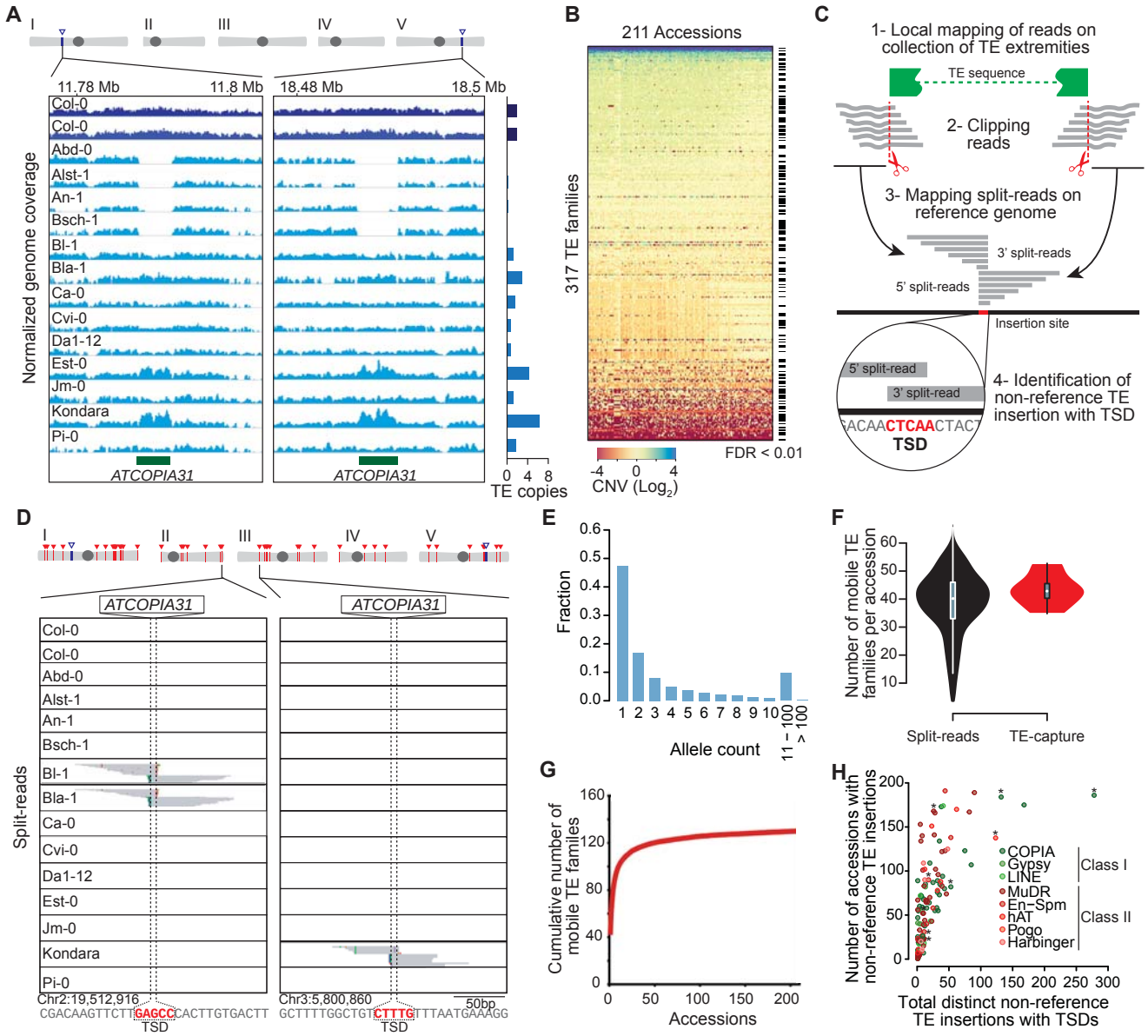
QQ-plot across the joined data set (391 accessions) for the indicated TE families. Summary statistics of associations are indicated below. MAF indicates Minor Allele Frequency in the joined dataset. Genes within GWAS intervals are indicated (*MET2a* is in bold).

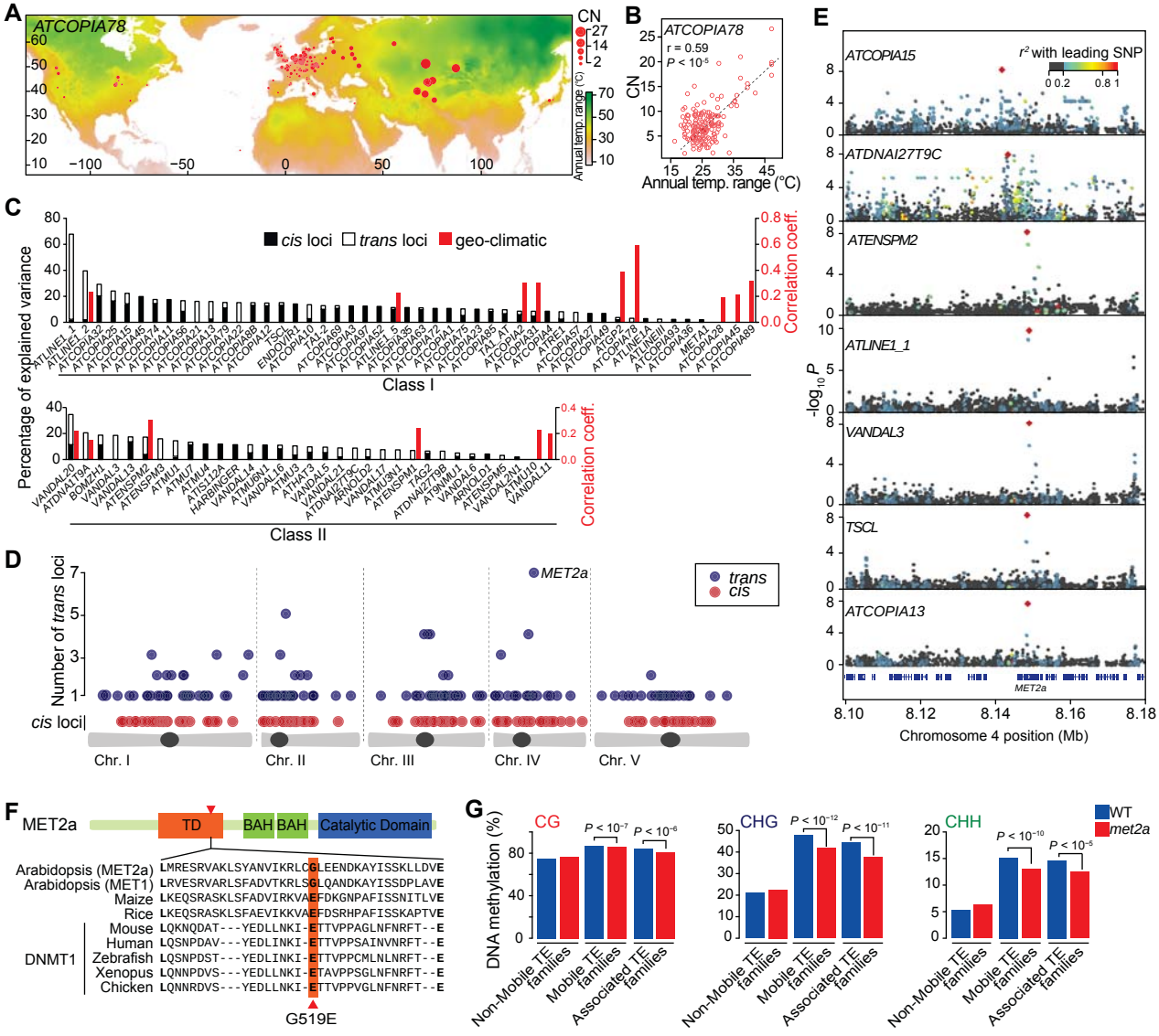
Supplementary file 2. DNA sequence motifs at insertions sites. Sequence logo of the overrepresented DNA sequence motifs at insertion sites (± 30bp) is shown for the 79 mobile TE families with at least 10 non-reference TE insertions with TSDs. The number of sequences used in each case is indicated.

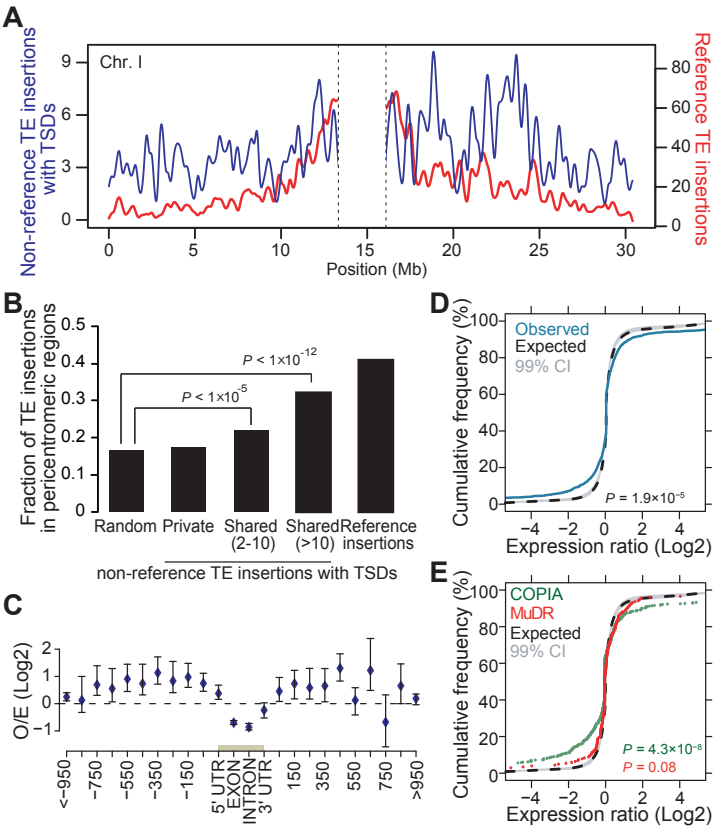
Supplementary file 3. PCR validation of non-reference TE insertions with TSDs and list of primer sequences used in this study.

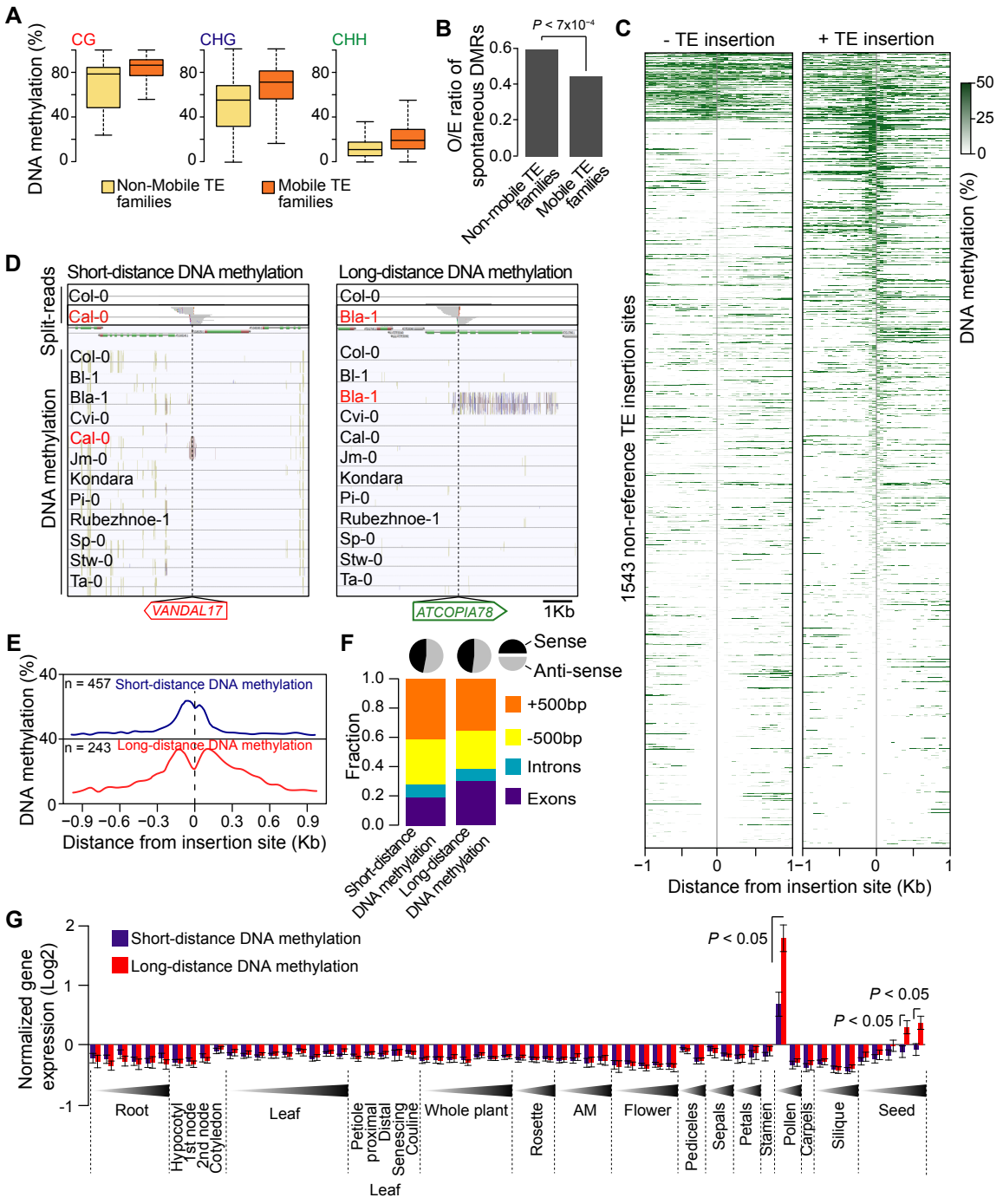
Supplementary file 4. **List of TE-capture targets**.

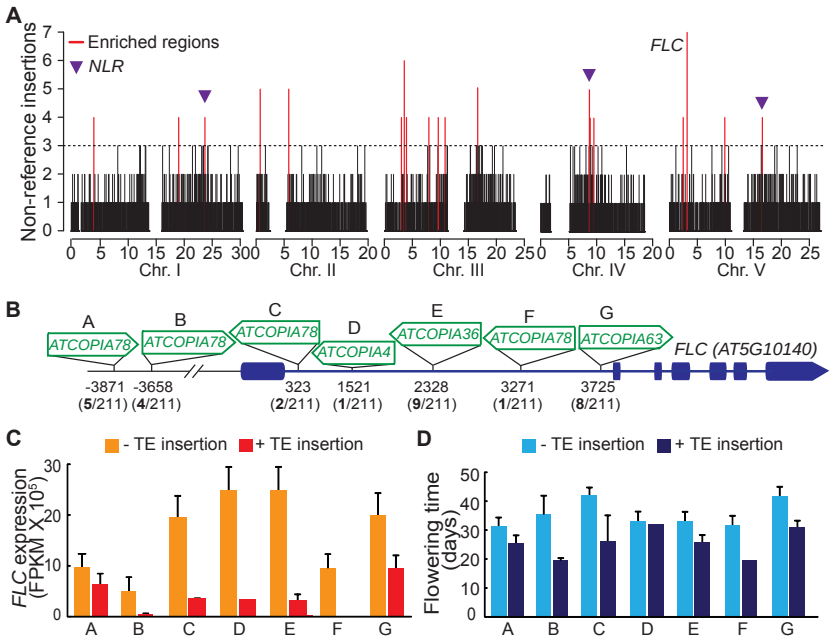
Supplementary file 5. **Historical population-scaled recombination rate map for *A. thaliana***


A
I
II
III
IV
V
11.78 Mb
11.8 Mb
18.48 Mb
18.5 Mb
Col-0
Col-0
Abd-0
Alst-1
An-1
Bsch-1
Bl-1
Bla-1
Ca-0
Cvi-0
Da1-12
Est-0
Jm-0
Kondara
Pi-0
Normalized genome coverage
ATCOPIA31
ATCOPIA31
0 4 8
TE copies
B
211 Accessions
317 TE families
FDR < 0.01
-4 0 4
CNV (Log2)
C
1- Local mapping of reads on collection of TE extremities
TE sequence
2- Clipping reads
3- Mapping split-reads on reference genome
3' split-reads
5' split-reads
Insertion site
5' split-read
3' split-read
ACAACTCAACTACT
TSD
4- Identification of non-reference TE insertion with TSD
D
I
II
III
IV
V
ATCOPIA31
ATCOPIA31
Split-reads
Col-0
Col-0
Abd-0
Alst-1
An-1
Bsch-1
Bl-1
Bla-1
Ca-0
Cvi-0
Da1-12
Est-0
Jm-0
Kondara
Pi-0
Chr2:19,512,916
CGACAAGTTCTTGAGCCCACTTGTGACTT
TSD
Chr3:5,800,860
GCTTTTGGCTGTCTTTGTTTAATGAAAGG
TSD
50bp
E
Fraction
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 – 100 > 100
Allele count
F
Number of mobile TE families per accession
60
50
40
30
20
10
Split-reads
TE-capture
G
Cumulative number of mobile TE families
160
120
80
40
0
0 50 100 150 200
Accessions
H
Number of accessions with non-reference TE insertions
200
150
100
50
0
COPIA
Gypsy
LINE
MuDR
En–Spm
hAT
Pogo
Harbinger
Class I
Class II
0 50 100 150 200 250 300
Total distinct non-reference TE insertions with TSDs


A
ATCOPIA78
CN
27
14
2
Annual temp. range (°C)
70
50
30
10
B
ATCOPIA78
r = 0.59
P < 10-5
CN
Annual temp. range (°C)
C
cis loci
trans loci
geo-climatic
Percentage of explained variance
Correlation coeff.
Class I
Class II
D
Number of trans loci
cis loci
MET2a
trans
cis
Chr. I
Chr. II
Chr. III
Chr. IV
Chr. V
E
ATCOPIA15
r2 with leading SNP
ATDNAI27T9C
ATENSPM2
ATLINE1_1
VANDAL3
TSCL
ATCOPIA13
MET2a
-log10 P
Chromosome 4 position (Mb)
F
MET2a
TD
BAH
BAH
Catalytic Domain
Arabidopsis (MET2a)
Arabidopsis (MET1)
Maize
Rice
DNMT1
Mouse
Human
Zebrafish
Xenopus
Chicken
G519E
G
CG
CHG
CHH
WT
met2a
DNA methylation (%)
Non-Mobile TE families
Mobile TE families
Associated TE families


A
Chr. I
Non-reference TE insertions with TSDs
Reference TE insertions
Position (Mb)
B
Fraction of TE insertions in pericentromeric regions
P < 1×10-12
P < 1×10-5
Random
Private
Shared (2-10)
Shared (>10)
Reference insertions
non-reference TE insertions with TSDs
C
O/E (Log2)
<−950
−750
−550
−350
−150
5' UTR
EXON
INTRON
3' UTR
150
350
550
750
>950
D
Observed
Expected
99% CI
Cumulative frequency (%)
P = 1.9×10−5
Expression ratio (Log2)
E
COPIA
MuDR
Expected
99% CI
Cumulative frequency (%)
P = 4.3×10−8
P = 0.08
Expression ratio (Log2)


A
DNA methylation (%)
CG
CHG
CHH
Non-Mobile TE families
Mobile TE families
B
O/E ratio of spontaneous DMRs
P < 7x10⁻⁴
Non-mobile TE families
Mobile TE families
C
- TE insertion
+ TE insertion
1543 non-reference TE insertion sites
DNA methylation (%)
Distance from insertion site (Kb)
D
Short-distance DNA methylation
Long-distance DNA methylation
Split-reads
DNA methylation
Col-0
Cal-0
Bla-1
Bl-1
Cvi-0
Jm-0
Kondara
Pi-0
Rubezhnoe-1
Sp-0
Stw-0
Ta-0
VANDAL17
ATCOPIA78
1Kb
E
DNA methylation (%)
n = 457 Short-distance DNA methylation
n = 243 Long-distance DNA methylation
Distance from insertion site (Kb)
F
Sense
Anti-sense
Fraction
+500bp
-500bp
Introns
Exons
Short-distance DNA methylation
Long-distance DNA methylation
G
Normalized gene expression (Log2)
Short-distance DNA methylation
Long-distance DNA methylation
P < 0.05
P < 0.05
P < 0.05
Root
Hypocotyl
1st node
2nd node
Cotyledon
Leaf
Petiole
proximal
Distal
Senescing
Couline
Leaf
Whole plant
Rosette
AM
Flower
Pediceles
Sepals
Petals
Stamen
Pollen
Carpels
Silique
Seed


A
Enriched regions
NLR
FLC
Non-reference insertions
Chr. I
Chr. II
Chr. III
Chr. IV
Chr. V
B
ATCOPIA78
ATCOPIA78
ATCOPIA78
ATCOPIA4
ATCOPIA36
ATCOPIA78
ATCOPIA63
FLC (AT5G10140)
-3871 (5/211)
-3658 (4/211)
323 (2/211)
1521 (1/211)
2328 (9/211)
3271 (1/211)
3725 (8/211)
C
- TE insertion
+ TE insertion
FLC expression (FPKM X 10^5)
D
- TE insertion
+ TE insertion
Flowering time (days)