Transposable Elements Contribute to the Adaptation of *Arabidopsis thaliana*

Zi-Wen Li[1,†], Xing-Hui Hou[1,2,†], Jia-Fu Chen[1,2], Yong-Chao Xu[1,2], Qiong Wu[1], Josefa González[3], and Ya-Long Guo[1,2,*]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

[3]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: yalong.guo@ibcas.ac.cn.

Accepted: August 8, 2018

Data deposition: The resequencing data of the reference genome Col-0 reported in this paper has been deposited in the NCBI Sequence Read Archive (SRA) under the accession number SRR5059243.

Abstract

Transposable elements (TEs) are mobile genetic elements with very high mutation rates that play important roles in shaping genome architecture and regulating phenotypic variation. However, the extent to which TEs influence the adaptation of organisms in their natural habitats is largely unknown. Here, we scanned 201 representative resequenced genomes from the model plant *Arabidopsis thaliana* and identified 2,311 polymorphic TEs from noncentromeric regions. We found expansion and contraction of different types of TEs in different *A. thaliana* populations. More importantly, we identified two TE insertions that are likely candidates to play a role in adaptive evolution. Our results highlight the importance of variations in TEs for the adaptation of plants in general in the context of rapid global climate change.

Key words: *Arabidopsis thaliana*, population genomics, adaptation, transposable elements (TEs).

Introduction

Transposable elements (TEs) represent an important source of genetic variation (McClintock 1984) and are highly dynamic in diverse species, such as Drosophila (Petrov et al. 2011; Kofler et al. 2015) and *Arabidopsis thaliana*-related species (Hu et al. 2011; Agren 2014; Quadrana et al. 2016; Stuart et al. 2016). TEs play crucial roles in shaping genomic architecture and phenotypic variation in diverse organisms (Finnegan 1989; Feschotte et al. 2002; Kazazian 2004; Lisch 2013).

Besides their well-known effect on genome size, the presence or absence of TEs affects various biological processes (Chuong et al. 2017). For example, inserted TEs can contribute to the coding sequences of existing genes or even form new coding genes in the genome (Lin et al. 2007; Hoen and Bureau 2015; Joly-Lopez et al. 2016). In addition, inserted TEs can regulate the expression levels of genes through either *cis*- or *trans*-regulatory elements located within TE sequences or through epigenetic modifications induced by the insertion or deletion of TEs (Naito et al. 2009; Hollister et al. 2011; Lisch 2013; Seymour et al. 2014; Stuart et al. 2016; Wei and Cao 2016). In maize, a transposon insertion located between 58.7 and 69.5 kb upstream of the well-known domestication gene *teosinte branched1* (*tb1*) acts as an enhancer of gene expression, which partially explains the increased apical dominance in maize compared with its progenitor (Studer et al. 2011). In melon, a transposon insertion located at the 3′ downstream of *CmWIP1* induced epigenetic changes, thereby regulating sex determination (Martin et al. 2009). In oil palm, loss of methylation on the *Karma* transposon in the intron of *DEFICIENS* contributed to the origin of a mantled somaclonal variant (Ong-Abdullah et al. 2015). In peppered moth, the industrial melanism mutation was induced by a transposon insertion at the first intron of *cortex*, which increased its expression level and induced melanization (Van't Hof et al. 2016). In *Drosophila melanogaster*, the activation of TEs is a contributing factor to ageing (Wood et al. 2016).

More importantly, TEs might function as agents of rapid adaptation, because they can rapidly create genetic diversity (Schrader et al. 2014; Stapley et al. 2015). Overall, TEs have emerged as important "functional elements" in the genomes of diverse organisms.

Despite the importance of TEs in shaping genome architecture and phenotypic variation, the TEs that are under positive selection in natural plant and animal populations are largely unknown. Moreover, it is important to address which genes are regulated by TE insertions to facilitate the rapid adaptation of the organism to global climate change (Rey et al. 2016). The answers to this question is largely unknown, except in *D. melanogaster*, some TEs were found to be candidate adaptive TEs based on frequency variation among populations (González et al. 2008, 2010), and/or on functional validation (Daborn 2002; Aminetzach et al. 2005; Schmidt et al. 2010; Magwire et al. 2011; Guio et al. 2014; Mateo et al. 2014; Ullastres et al. 2015; Merenciano et al. 2016).

In this study, to explore the effect of TEs on adaptation at the whole-genome level, we investigated natural populations of *A. thaliana*, as TEs in this model plant have been annotated in detail and many accessions have been sequenced, and this plant originated in Europe and northern Africa and adapted to new climates as it expanded eastward to Eastern Asia (Cao et al. 2011; Long et al. 2013; Schmitz et al. 2013; The 1001 Genomes Consortium 2016; Durvasula et al. 2017; Zou et al. 2017). To investigate the effect of TEs on adaptation in natural populations, we identified both reference and nonreference TE insertions using the read pair method. Note that nonreference TE insertions have been taken into account recently (Quadrana et al. 2016; Stuart et al. 2016).

We identified 2,311 polymorphic TEs from 201 representative *A. thaliana* accessions collected worldwide, and found the differential expansion and contraction of diverse types of TEs in different populations. We identified two TE insertions that are likely candidates to play a role in adaptive evolution. Overall, this study highlights the potential effects of TEs on adaptive evolution of *A. thaliana* in nature.

Materials and Methods

Cluster Analysis of Accessions

Raw paired-end reads of 201 accessions were used in this study, including 118 representative accessions from Europe, Central Asia, North America, and Japan, which were extracted from the 1001 Genomes Project (http://1001genomes.org/) (Cao et al. 2011; Long et al. 2013; Schmitz et al. 2013; The 1001 Genomes Consortium 2016), as well as 83 accessions from our own resequencing project (Zou et al. 2017), including 24 accessions from Northwestern China, and 59 accessions from the Yangtze River basin (supplementary table S1, Supplementary Material online). The raw reads were processed and then mapped to the *A. thaliana* reference genome (TAIR10) using BWA (Li and Durbin 2009) with default parameters. Single nucleotide polymorphisms (SNPs) were called using the Genome Analysis Toolkit (GATK) flowchart (DePristo et al. 2011) with a quality value of 25 as the threshold. Only biallelic SNP sites with minor allele frequencies greater than 0.05 were used in the principal component analysis (PCA) with EIGENSOFT (version 4.2) (Price et al. 2006). Accessions located between major clusters were filtered out.

Identifying Polymorphic TE Sites in Populations

Genome resequencing data, and the *A. thaliana* reference genome (Col-0 accession, TAIR10) were used to identify polymorphic TE sites in the *A. thaliana* populations. The polymorphic TE sites are TE loci in which some accessions harbor TE insertions but others do not. A method based on paired-end reads was used to identify nonreference and reference TE insertions. The paired-end reads is often used to identify polymorphic TEs (Platzer et al. 2012; Kofler et al. 2015).

To increase the accuracy of identification, mapping direction information was integrated into the identification process as previously described (Platzer et al. 2012): If the mapped read is reversely complemented, its direction is backward; if not, the direction is forward (fig. 1*B*). Mapped reads at the left and right sides of each TE site should have the same orientation (a forward-reverse arrangement in mapping result) to ensure that the detected presence/absence of TEs did not result from a chromosome inversion event.

Three steps were used to detect nonreference TE insertions. First, two reads in a pair with a mapping distance >1 kb, including one read uniquely mapped to a non-TE region and the other mapped to the annotated TE sequence similar to the predicted TE insertion sequence in the reference genome, were extracted from the mapping results for each accession. Information about the families of annotated TEs was extracted from TAIR10 annotation, which was used to predict the family types of the inserted TEs. Second, abnormally mapped reads located far from each other but within a certain range (the sequencing length of one read plus twice of the designated insertion size between paired reads) were used to set the insertion range of a polymorphic TE candidate (supplementary fig. S1*A*, Supplementary Material online). The designated insertion size between paired reads was 300 bp for accessions from our own project and 50 bp for accessions from the 1001 Project. The sequencing length of one read in the 1001 Project and our own project was 100 bp. Third, candidates from accessions in a population were merged together when they overlapped (>1 bp). During this step, candidate polymorphic TEs from different accessions were integrated at the population level. During the merging process, the number of reads supporting a candidate polymorphic TE in the population represented the average number of reads supporting this insertion in all accessions of the

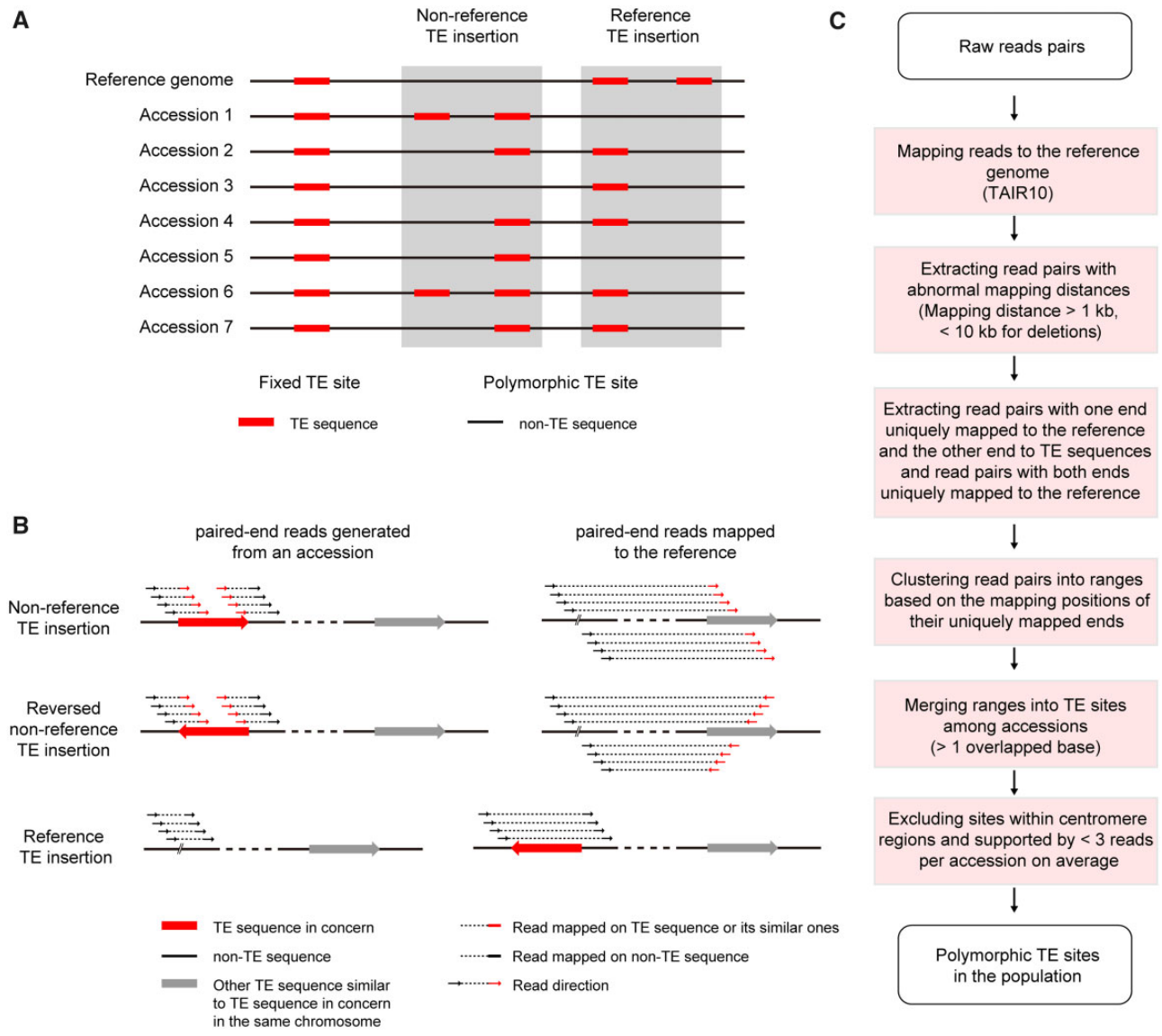

Fig. 1.—Identification of polymorphic reference and nonreference TE sites. (*A*) Diagram of polymorphic TE sites. (*B*) Diagram of the polymorphic TE identification method. (*C*) Flowchart of the polymorphic TE detection method.

population. Candidate polymorphic TEs supported by more than one read pair per accession were identified as the raw data of polymorphic TEs. Furthermore, polymorphic TEs supported by at least three reads per accession were used in the analysis.

Similar to the detection of nonreference TE insertions, a read pair with the following features was considered to represent a candidate polymorphic reference TE: 1) read pair with a mapping distance greater than 1 kb and less than 10 kb (90% of annotated TEs in the reference genome are shorter than 10 kb); 2) both of the paired reads uniquely mapped to the reference genome in non-TE regions; 3) annotated TEs are present between the mapped positions of the two reads in the reference genome. These read pairs with distances below a certain length (the designated insertion size between paired reads) were merged together as candidate polymorphic TEs identified by reference TE insertion method (supplementary fig. S1*A*, Supplementary Material online). Finally, candidate polymorphic nonreference and reference TEs were integrated together into the polymorphic TE data set.

Evaluating the Sensitivity and False Discovery Rate of Our Method

To estimate the detection power of our method, the reference accession Col-0 was resequenced using Illumina HiSeq 2000 with 100 bp pair-end reads, and the reads were remapped to the modified reference genomes. In detail, to evaluate nonreference TE insertion method, annotated, non-overlapping TE sequences (if overlapping, the longest one was

used) in the *A. thaliana* reference genome were deleted from the original regions and moved to the 3′ end of the chromosome they were located (supplementary fig. S1*B*, Supplementary Material online). This modified TE-deleted genome sequence (Col-0-noTEs) was the same as Col-0, except for the locations of TEs. Given that the background of resequenced reads of Col-0 was from the real reference genome with the annotated TEs in their original positions, remapping these reads to the genomes of Col-0-noTEs could uncover "nonreference TE insertions" that were deleted from the real reference genome, thereby allowing the detection efficiency of nonreference TE insertions to be calculated. For reference TE insertions, the Col-0-moreTEs genome was created by randomly inserting all TEs annotated in TAIR10 into the Col-0 genome and subjecting it to the same remapping strategy. Centromeric regions of *A. thaliana* chromosomes were defined according to a previous study (Ziolkowski et al. 2009).

The identified TE sites in the populations were validated using 20 randomly chosen TE sites (see supplementary table S2, Supplementary Material online for information about the 20 TEs and 12 accessions used for validation). Primers were designed to span the predicted TE sites (supplementary table S6, Supplementary Material online). In addition, for TE deletion sites, the PCR product of one randomly chosen accession per TE site was sequenced to confirm the TE deletion.

Detection of Adaptive TE Insertions

In iHS (integrated haplotype score) statistic analysis, biallelic homozygous SNP sites and polymorphic TE loci with minor allele frequencies greater than 0.05 were used to calculate iHS using the selscan program (Szpiech and Hernandez 2014). The absolute value of iHS (|iHS|) was used to detect selective sweeps. A significant high |iHS| value is an indicator of the location of a TE locus in a selective sweep region. To estimate the significance of observed iHS values, we computed iHS values in permutated data (100 permutations) for TEs with the top 5% highest |iHS| values. At each TE locus and in each permutation, a number of accessions equally to the original number of TE insertion accessions were randomly sampled without repetition from a population, which were considered as permutated accessions with TE insertion alleles; correspondingly, other alleles were considered as non-TE alleles. These permutations were performed 100 times for each TE locus in popE, popN, and popY, respectively. After permutations, TEs with observed |iHS| values higher than the permutated values were further analyzed.

The f_{TE} statistic was estimated according to the method reported in a previous study (González et al. 2008). Permutation analyses on f_{TE} were also performed for each polymorphic TE locus in each population (100 permutations). A significant low f_{TE} is an indicator of positive selection on TE insertion alleles. TEs with significantly low f_{TE} values were considered as putatively candidates. Finally, iHS was also estimated for all SNPs in 20 kb regions flanking the candidate adaptive TEs. Extended haplotype homozygosity (EHH) statistic was performed by Rehh package based on polymorphic data set used in iHS calculation.

Statistical Analysis

All statistical analyses were performed using the R package v3.1.3 (R Core Team 2014). All *P*-values in multiple testing were adjusted using the "fdr" option in the "*p.adjust*" function in R (Benjamini and Hochberg 1995).

Results

Identification of Polymorphic TEs in Various Populations

We used paired-end reads to identify polymorphic TE sites, including both reference and nonreference polymorphic TEs, based on the inconsistency between the mapping distances and the insertion sizes of read pairs (fig. 1*A* and *B*; supplementary fig. S1*A*, Supplementary Material online). Nonreference TE insertions are TEs present in at least one accession that do not exist in the reference genome at the syntenic region. Reference TE insertions are TE insertions that exist in the reference genome but are absent in at least one other accession (fig. 1*A*).

The modified reference genome sequence (Col-0), with TE sequences removed from their original positions in the reference genome (supplementary fig. S1*B*, Supplementary Material online), was used to evaluate the detection power of our method. Apparently, the use of mapping direction information was an efficient way to reduce the false discovery rate (FDR; from 11.26% to 1.39%) while roughly maintaining the sensitivity of detection (from 90.59% to 89.05%; table 1). In addition, given that TEs are enriched in centromeric regions, TEs identified in these regions likely have a much higher FDR due to the difficulty in mapping short reads to such regions. Consistently, TEs detected across the whole genome, including centromeric regions, had a higher FDR (1.39% vs 1.04%) and a lower sensitivity (89.05% vs 90.45%) than the modified genome without centromeric regions (table 1). Furthermore, TEs located outside of the centromere that are supported by at least three reads (on average) across all accessions had an even lower FDR (0.93%; table 1). Therefore, in subsequent analyses, we focused on the polymorphic TEs outside of centromere (fig. 1*C*).

Given that the estimated sensitivity of our method is approximately 90%, 10% of the TEs were not detected. These "missing" TEs appear to share common characteristics: Most are 100 bp or even shorter, and the distances from nearby TEs are frequently less than 1 kb (supplementary fig. S2*A* and *B*, Supplementary Material online). Consistently, among different TE families, the proportions of the TEs identified ranged from 42% to 98% (supplementary fig. S2*C*, Supplementary Material online). For example, for TE families RathE1, RathE2,

Table 1
Sensitivity and FDR of Various Identification Methods

Identification Method	Number of Identified TEs	Sensitivity (%)	FDR (%)
Raw data	10,910	90.59	11.26
Filtering with direction information	9,608	89.05	1.39
Filtering with direction information in noncentromeric region	8,883	90.45	1.04
Filtering with direction information in noncentromeric region and at least three reads	8,853	90.32	0.93

and RathE3, only half of the total TEs were detected, largely due to their shorter lengths or being located too close to nearby TEs (supplementary fig. S2*D* and E, Supplementary Material online). Therefore, TEs longer than 100 bp and more than 1 kb away from other TEs could be identified by our method.

Polymorphic TEs in Three *A. thaliana* Populations

We used resequenced genomes of 201 representative accessions, mainly from Eurasia (supplementary fig. S3 and table S1, Supplementary Material online). The resequencing depths of these accessions were all greater than 15×, and those of 178 accessions were greater than 20×. PCA based on SNPs revealed that accessions from the Yangtze River basin clustered into an independent group; in contrast, accessions from Northwestern China formed a cluster with several Central Asian accessions that joined with the Europe accessions cluster (fig. 2). Eleven accessions that were roughly isolated from the three main clusters were excluded from subsequent analysis (fig. 2, accession names shown in red). We ultimately selected 191 accessions, 59 from the Yangtze River basin, 24 from Northwestern China, and 108 (mainly) from Europe including Col-0 reference genome for polymorphic TEs analysis, which were considered to form three large populations (hereafter referred to as popY, popN, and popE, respectively).

Using the paired-end method and excluding TEs present in only one accession or supported by only one paired-end read per accession, we identified a total of 4,305 polymorphic TE loci in the 3 populations (table 2), including 3,856 in noncentromeric regions, 2,311 of which were supported by at least 3 reads pairs per accession. Permutation analyses revealed that the number of detected polymorphic TEs does not increase linearly: 95 randomly selected accessions contain nearly 90% of all the polymorphic TEs (supplementary fig. S4*A*, Supplementary Material online). Therefore, the number of TEs tends to reach saturation rather than grow continually when the number of accessions increases. Validation of 20 identified TE loci by PCR using 12 representative accessions suggested that the FDR of the predicted TE present/absent events for the 2,311 polymorphic TE loci was 5.42% (supplementary table S2, Supplementary Material online).

One measure of the success of our identification method is that the allele frequency distribution was U-shaped (fig. 3*A*). The allele frequency distribution of polymorphic TE sites at the genome level is usually U-shaped (Kofler et al. 2015), whereas the allele frequency distribution of polymorphic TEs taking into account only the nonreference TE insertions or only the reference TE insertions in popE obtained in the present study is L-shaped (supplementary fig. S5*A* and *B*, Supplementary Material online). The abnormal frequency distribution of TEs in popE may largely result from the different genetic distances between the studied accessions and the reference accession. This bias became stronger with increasing genetic distance between the studied accessions and the reference accession, especially for accessions within popE, which are closely related to the reference genome (Spearman's rank correlation coefficient [τ] =0.66, $P < 1.26 \times 10^{-14}$ for nonreference TE insertions; supplementary fig. S6*A*, Supplementary Material online; $\tau = -0.43$, $P < 4.54 \times 10^{-6}$ for reference TE insertions; supplementary fig. S6*B*, Supplementary Material online). However, after merging nonreference and reference polymorphic TEs, there was no significant correlation between genetic distances and the number of polymorphic TE insertions ($\tau = -0.11$, $P = 0.26$) (supplementary fig. S6*C*, Supplementary Material online), and the frequency distribution of polymorphic TEs in popE was U-shaped (supplementary fig. S5*C*, Supplementary Material online). Overall, combining results from nonreference and reference TE insertions is a much more robust way to reveal the evolutionary pattern of TEs than polymorphic TEs only identified by nonreference or by reference TE insertions.

Among the 2,311 loci, 1,339 loci were nonreference TE insertions and the other 972 loci were reference TE insertions. For each population, we identified 2,064, 1,434, and 1,403 polymorphic TE loci in popE, popN, and popY, respectively. In addition, 1,047 TE loci were polymorphic in all 3 populations. In contrast, we identified 618, 81 and 69 population-specific TE loci in popE, popN, and popY, respectively. Apparently, popE had the largest number of polymorphic TEs and population-specific TEs. After the effect of sample size was ruled out via a permutation test, popE still had the largest number of population-specific TEs (supplementary fig. S4*B*, Supplementary Material online), and also the largest number of inserted TEs (table 2; supplementary fig. S4*C* and *D*, Supplementary Material online). Most TEs are distributed at intergenic regions at either the species level or the specific population level, but, still, nearly 20% of TEs exist at genic regions in either coding sequences or introns (fig. 3*B*). Coding regions (CDS) of 242 genes contain 245 polymorphic TE insertions (supplementary table S3, Supplementary Material online). The functions of these genes are enriched in defense response

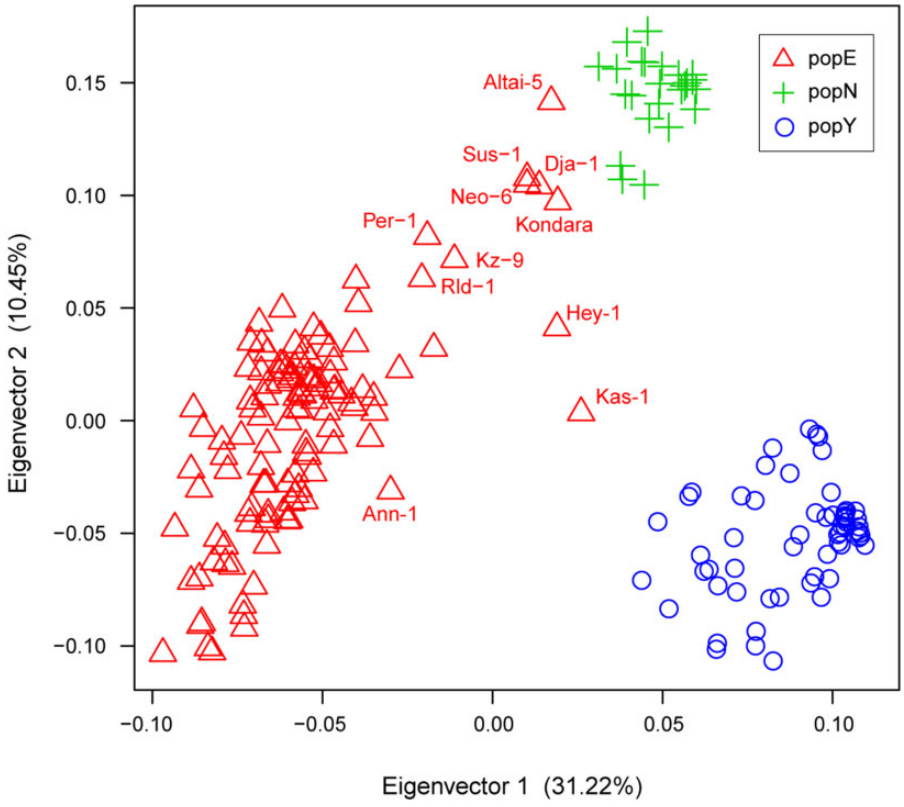

Fig. 2.—PCA based on SNP sites in 201 accessions. Accessions filtered out of the three populations are marked with accession names. popE, popN, and popY represent accessions mainly from Europe, accessions from Northwestern China and accessions from the Yangtze River basin, respectively.

Table 2
The Numbers of TE Loci in the Three Populations

	popE	popN	popY	Three Populations
Raw TE loci	3,930 (2,046/1,884)	2,891 (1,258/1,633)	2,932 (1,324/1,608)	4,305 (2,421/1,884)
TE loci in noncentromeric region	3,504 (1,756/1,748)	2,556 (1,052/1,504)	2,556 (1,077/1,479)	3,856 (2,108/1,748)
TE loci in noncentromeric region with at least three reads per accession on average	2,064 (1,092/972)	1,434 (704/730)	1,403 (700/703)	2,311 (1,339/972)
Inserted TEs	101,085 (16,443/84,642)	19,684 (7,910/11,774)	47,869 (20,282/27,587)	168,638 (44,635/124,003)
Inserted TEs per accession (95% confidence intervals)	944.7±7.7 (153.7±6.9/791.0±12.4)	820.2±5.5 (329.6±5.0/490.6±4.9)	811.3±2.9 (343.8±2.3/467.6±3.1)	887.6±10.3 (234.9±13.8/652.6±23.5)
Population-specific TE loci	618 (450/168)	81 (81/0)	69 (69/0)	768 (600/168)

The first number in parentheses is the number of polymorphic TEs identified by nonreference TE insertion, and the second is the number of polymorphic TEs identified by reference TE insertion.

(GO enrichment analysis, FDR =0.000075) and immune response (FDR =0.015). Polymorphic TEs inserted in CDS regions and untranslated regions (UTRs) were significantly biased toward low frequencies (frequency ≤0.1) compared with TEs in intergenic regions (fig. 3C; Fisher's exact test, multiple testing corrected $P = 0.0013$ for TEs in CDS regions and 0.0063 in UTR regions). These results indicate the spreading of TE insertions in the regions of CDS and UTR is constrained by purifying selection.

Of the 2,311 polymorphic TEs, 1,445 could be classified into specific TEs types. The proportion of DNA-type TEs (35.8%) is significantly higher than that of Helitron-type TEs (26.0%; χ^2 test, $P = 1.36 \times 10^{-8}$) and LTR-type (long terminal repeat) TEs (30.7%; $P = 0.0039$). Furthermore, popE, popN, and popY are roughly consistent in the composition of polymorphic TE types, as well as for polymorphic TEs shared among the three populations (fig. 3D; supplementary table S4, Supplementary Material online). However, the

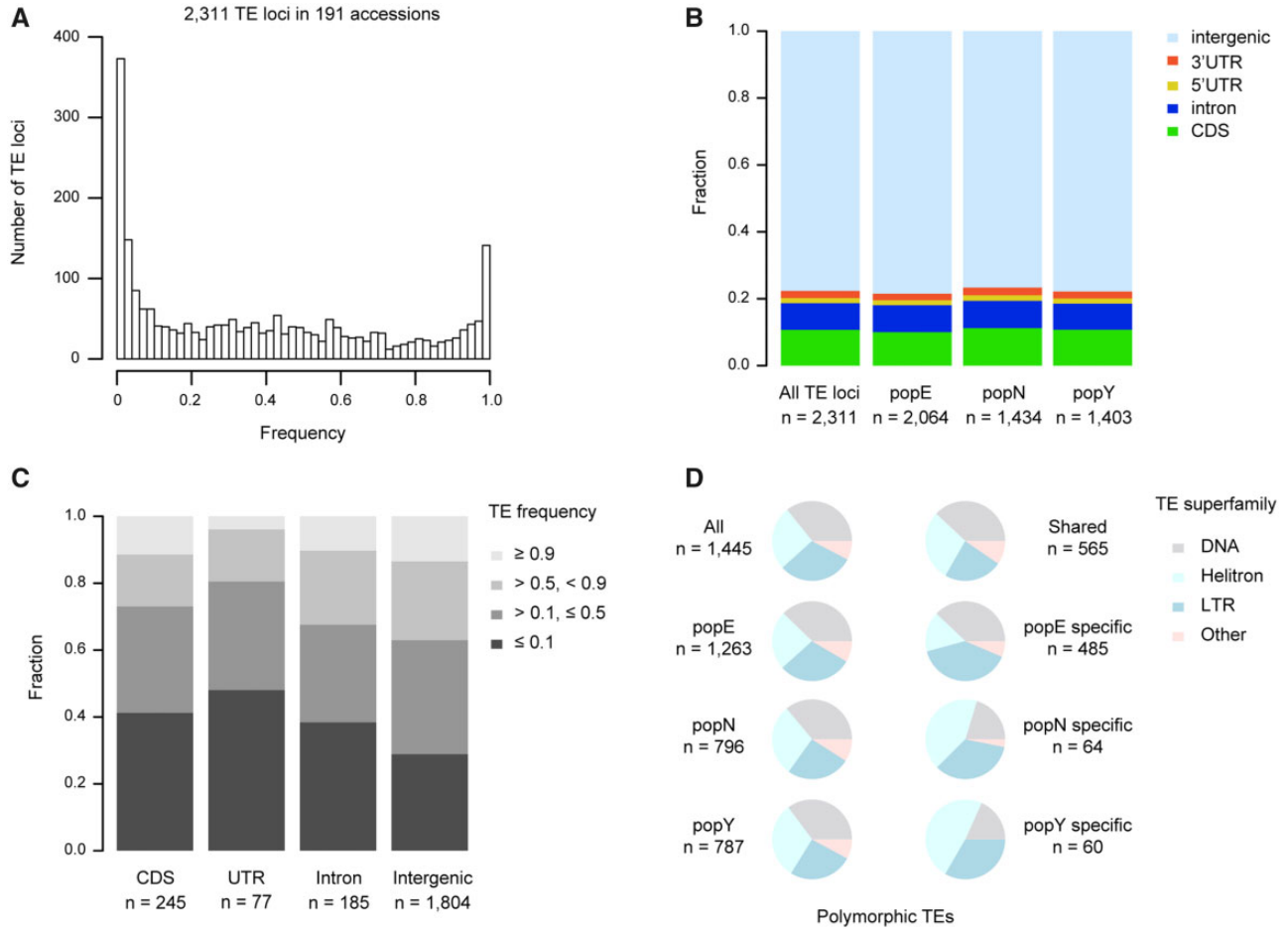

Fig. 3.—Frequency and distribution of polymorphic TEs in *A. thaliana* populations. (*A*) Frequency of polymorphic TEs in *A. thaliana*. (*B*) Distribution of polymorphic TEs in the genome. (*C*) Frequency of polymorphic TEs in different genomic regions. (*D*) Composition of polymorphic TEs in different populations.

compositions of population-specific TEs in each population were different. DNA-type and LTR-type TEs (37.9% and 39.4%) are enriched in popE-specific polymorphic TEs, whereas Helitron-type TEs are the major components in popN- and popY-specific polymorphic TEs (42.2% and 48.3%, respectively; fig. 3*D*). Overall, these results suggest that when *A. thaliana* spread across the world, the expansion and contraction of different types of TEs differentiated among *A. thaliana* populations.

Detecting Adaptive TE Insertions

Given the functional importance of TEs, TE loci with a selective advantage in specific environments could spread in a population and speed up the adaptation of an organism. We aimed to identify TEs that might have contributed to the adaptation of *A. thaliana* as it expanded out of Europe (The 1001 Genomes Consortium 2016; Zou et al. 2017). We screened for adaptive TEs in each population (popE, popN, popY) in two steps. First, we identified adaptive TE candidates located in selective sweep regions using the integrated haplotype score (iHS statistic), and the nucleotide diversity of TE insertion alleles compared with that of the background genome (f_{TE} statistic; Voight et al. 2006; González et al. 2008). Second, we estimated whether the adaptive TE candidates were the actual targets of positive selection by comparing iHS values of the adaptive TE candidates with its surrounding SNP sites in 20 kb regions (fig. 4*A*). iHS is a commonly used method in detecting genetic loci under positive selection (Voight et al. 2006; Colonna et al. 2014; Nedelec et al. 2016), as well as in identifying adaptive TEs (González et al. 2008). Here, we performed iHS analysis in each population (popE, popN, and popY) on polymorphic TE sites and genome wide SNP sites with minor allele frequency (MAF) larger than 0.05. We identified 49, 46, and 38 TE sites having the top 5% highest |iHS| values among polymorphic TEs, in popE (|iHS| thresholds is 2.33), popN (2.47), and popY (1.94), respectively (fig. 4*B*). To quantify the significance of the top 5% highest |iHS| values, permutations were performed 100 times at each of the above TE sites. Consequently, 49, 17, and 31

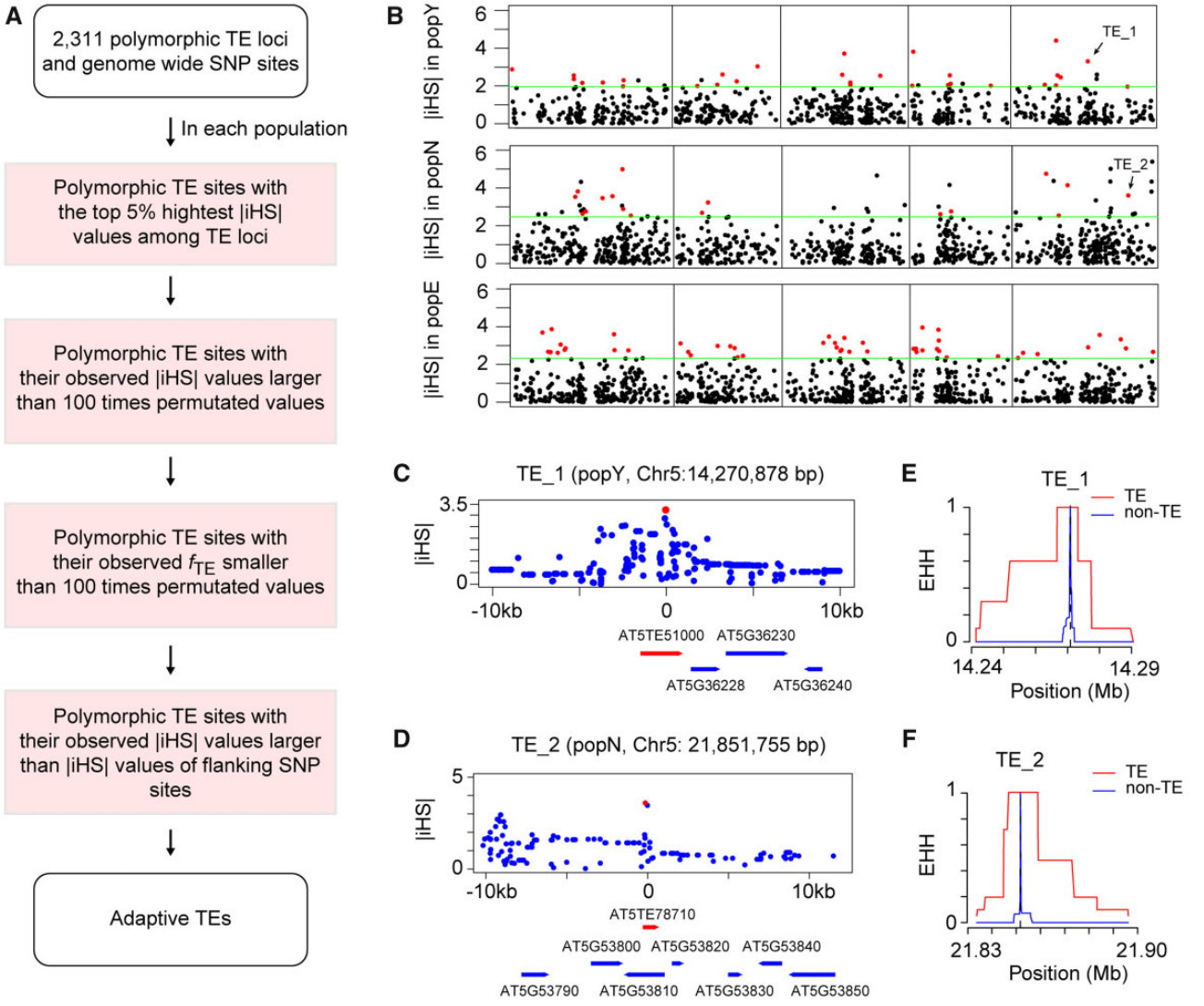

Fig. 4.—Identification of adaptive TEs. (*A*) Flowchart of the adaptive TE detection method. (*B*) The absolute values of integrated haplotype scores (|iHS|) at polymorphic TE sites in popY, popN, and popE, respectively. Each point indicates a polymorphic TE site (MAF>0.05). The green line represents the threshold of top 5% highest |iHS| values in each population. Red points are polymorphic TE sites with significantly high |iHS| values in the permutation analysis, whereas polymorphic TE sites with nonsignificant |iHS| values are marked as black points. The arrows indicate the two adaptive TEs. (*C*, *D*) |iHS| values of TE_1 and TE_2 (the red points) and SNP sites (the blue points) in 20 kb flanking regions, respectively. Gene models and the TE locus are corresponding to genome positions used in |iHS| plot. (*E*, *F*) Extended haplotype homozygosity (EHH) in adaptive TEs and its flanking regions.

TE sites with observed |iHS| values larger than the permutated values remained significant in popE, popN, and popY, respectively (fig. 4*B*).

Genome regions under positive selection usually have lower nucleotide polymorphisms than background genomes. To confirm the result of adaptive TE candidates identified by the iHS method, we computed the proportion of nucleotide polymorphisms of TE insertion accessions (π_{TE}, 2 kb around the predicted insertion site, namely, 1 kb upstream and 1 kb downstream) to the total nucleotide diversity of all the accessions ($\pi_{TE} + \pi_{non\text{-}TE}$, in the same 2 kb region) for each polymorphic TE locus in each population (f_{TE} statistics). Permutation analyses on f_{TE} values were performed according to the same method used for |iHS| permutation, at each TE site in each population (see Material and methods for details). TE sites with f_{TE} values lower than their 100 times permutated values were considered significant, as they are more likely located in genome regions affected by positive selection. Among TE sites with significantly high |iHS| values, 33, 7, and 13 TE sites have significant low f_{TE} values in popE, popN, and popY, respectively. Adaptive TE candidates with low f_{TE} values have low nucleotide diversity in TE insertion alleles, which suggest the positive selection targets may be the TE insertion alleles. Thus, overall we identified 33, 7, and

Table 3
List of Two Adaptive TEs

Adaptive TE	The Predicted Central Position of TE	The Upstream Gene		The Related TE		The Downstream Gene		TE Insertion Frequency in popE/N/Y	|iHS|	f_{TE}
		Gene ID	Annotation	TE ID	Annotation	Gene ID	Annotation			
TE_1	Chr5: 14270878	AT5G36220	CYP81D1	AT5TE51000	LINE element	AT5G36228	Nucleic acid-binding protein	76/1/5	3.29	0.036
TE_2	Chr5: 21851755	AT5G53800	Nucleic acid-binding protein	AT5TE78710 in the intron of AT5G53810	Copia element; AT5G53810 is an *O*-methyltransferase family protein.	AT5G53820	Late embryogenesis abundant protein (LEA) family protein	101/7/0	3.60	0

13 adaptive TE candidates in the 3 populations, respectively (supplementary table S5, Supplementary Material online).

To further confirm that TEs identified as adaptive candidates are the targets of positive selection, we screened |iHS| values of SNP sites in 20 kb regions surrounding each TEs (10 kb upstream and 10 kb in downstream of each TE). Finally, 2 adaptive TE candidates have the highest |iHS| values in their flanking 20 kb regions (fig. 4*C* and *D*; table 3). The two adaptive TEs show higher haplotype homozygosities in TE insertion alleles than alleles without the TEs, respectively (fig. 4*E* and *F*). These two adaptive TE candidates are more likely to be the actual targets of positive selection in the tested populations.

Discussion

TEs are a major source of genomic mutations, which, like any environmental mutagen, occasionally lead to beneficial changes (Lynch 2007). After its discovery in maize (McClintock 1950), TEs have been investigated comprehensively, including their identification and classification (Lisch 2013), evolutionary dynamics (Petrov et al. 2011; González and Petrov 2012; Agren 2014; Barrón et al. 2014; Bousios et al. 2016; Pietzenuk et al. 2016), and structural and functional effects (Kazazian 2004; Rey et al. 2016; Wei and Cao 2016). From an evolutionary perspective, it is important to study two aspects of TEs, that is, their evolutionary dynamics in the genome and the contribution of TEs to adaptation in the context of global climate change. To address these questions, we need to identify TEs based on resequencing data sets from natural populations. Most previous studies of TEs in populations have only focused on polymorphic TEs absent from the reference genome and have ignored reference TE insertions, thereby failing to address a significant portion of TE polymorphisms. In this study, we merged data from reference and nonreference TE insertions, and demonstrated that combining nonreference and reference TE insertions is a more robust way to reveal the evolutionary pattern of TEs.

As TEs represent an important source of genetic variation, they can contribute to the evolution of an organism in diverse ways, such as acquiring coding ability and altering the coding sequence (Cowan et al. 2005; Joly-Lopez et al. 2012; Sun et al. 2014), and the expression level of a gene (Kobayashi et al. 2004; González et al. 2009; Butelli et al. 2012). More importantly, TE mutations could affect adaptation to the environment (Casacuberta and González 2013; Quadrana et al. 2016; Van't Hof et al. 2016), and TE mutations with beneficial effects on adaptation in natural populations could become fixed. In this study, we found that at least two of 2,311 TE loci are likely to be targets of positive selection, and thus have contributed to the adaptation of *A. thaliana*. Overall, our findings highlight that TEs could play important roles in the adaptation of organisms to global climate change.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank the anonymous reviewers to help improve the manuscript, and thank Song Ge for insightful comments and members of the Guo laboratory for discussions and technical assistance.

Funding

This work was supported by National Natural Science Foundation of China grants (91231104, 31470331, and 31222006 to Y.-L.G.) and the 100 Talents Program of the Chinese Academy of Sciences (Y.-L.G.). J.G. is funded by the European Commission (H2020-ERC-2014-CoG-647900) and by the Spanish Ministry of Science, Innovation, and Universities (BFU2014-57779-P and BFU2017-82937-P).

Literature Cited

Agren JA. 2014. Mating system shifts and transposable element evolution in the plant genus *Capsella*. BMC Genomics 15(1):602.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309(5735):764–767.

Barrón MG, Fiston-Lavier AS, Petrov DA, Gonzalez J. 2014. Population genomics of transposable elements in Drosophila. Annu Rev Genet. 48(1):561–581.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to mutliple testing. J R Stat Soc B 57:289–300.

Bousios A, et al. 2016. A role for palindromic structures in the cis-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response. Genome Res. 26(2):226–237.

Butelli E, et al. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24(3):1242–1255.

Cao J, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 43(10):956–963.

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22(6):1503–1517.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18(2):71–86.

Colonna V, et al. 2014. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 15(6):R88.

Cowan RK, Hoen DR, Schoen DJ, Bureau TE. 2005. *MUSTANG* is a novel family of domesticated transposase genes found in diverse angiosperms. Mol Biol Evol. 22(10):2084–2089.

Daborn PJ. 2002. A single p450 allele associated with insecticide resistance in Drosophila. Science 297(5590):2253–2256.

DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491–498.

Durvasula A, et al. 2017. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 114(20):5213–5218.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 3(5):329–341.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5(4):103–107.

González J, Karasov TL, Messer PW, Petrov DA. 2010. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genet. 6(4):e1000905.

González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. 2008. High rate of recent transposable element-induced adaptation in *Drosophila melanogaster*. PLoS Biol. 6(10):e251.

González J, Macpherson JM, Petrov DA. 2009. A recent adaptive transposable element insertion near highly conserved developmental loci in *Drosophila melanogaster*. Mol Biol Evol. 26(9):1949–1961.

González J, Petrov DA. 2012. Evolution of genome content: population dynamics of transposable elements in flies and humans. Methods Mol Biol. 855:361–383.

Guio L, Barron MG, Gonzalez J. 2014. The transposable element Bari-Jheh mediates oxidative stress response in Drosophila. Mol Ecol. 23(8):2020–2030.

Hoen DR, Bureau TE. 2015. Discovery of novel genes derived from transposable elements using integrative genomic analysis. Mol Biol Evol. 32(6):1487–1506.

Hollister JD, et al. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 108(6):2322–2327.

Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43(5):476–481.

Joly-Lopez Z, Forczek E, Hoen DR, Juretic N, Bureau TE. 2012. A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in *Arabidopsis thaliana*. PLoS Genet. 8(9):e1002931.

Joly-Lopez Z, Hoen DR, Blanchette M, Bureau TE. 2016. Phylogenetic and genomic analyses resolve the origin of important plant genes derived from transposable elements. Mol Biol Evol. 33(8):1937–1956.

Kazazian HH. Jr 2004. Mobile elements: drivers of genome evolution. Science 303(5664):1626–1632.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. Science 304(5673):982.

Kofler R, Nolte V, Schlotterer C. 2015. Tempo and mode of transposable element activity in Drosophila. PLoS Genet. 11(7):e1005406.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Lin R, et al. 2007. Transposase-derived transcription factors regulate light signaling in Arabidopsis. Science 318(5854):1302–1305.

Lisch D. 2013. How important are transposons for plant evolution? Nat Rev Genet. 14(1):49–61.

Long Q, et al. 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet. 45(8):884–890.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. 2011. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet. 7(10):e1002337.

Martin A, et al. 2009. A transposon-induced epigenetic change leads to sex determination in melon. Nature 461(7267):1135–1138.

Mateo L, Ullastres A, Gonzalez J. 2014. A transposable element insertion confers xenobiotic resistance in Drosophila. PLoS Genet. 10(8):e1004560.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 36(6):344–355.
McClintock B. 1984. The significance of responses of the genome to challenge. Science 226(4676):792–801.
Merenciano M, Ullastres A, de Cara MAR, Barrón MG, González J. 2016. Multiple independent retroelement insertions in the promoter of a stress response gene have variable molecular and functional effects in Drosophila. PLoS Genet. 12(8):e1006249.
Naito K, et al. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461(7267):1130–1134.
Nedelec Y, et al. 2016. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell 167(3):657–669.e621.
Ong-Abdullah M, et al. 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525(7570):533–537.
Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, Gonzalez J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. Mol Biol Evol. 28(5):1633–1644.
Pietzenuk B, et al. 2016. Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. Genome Biol. 17(1):209.
Platzer A, Nizhynska V, Long Q. 2012. TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. Biology (Basel) 1(2):395–410.
Price AL, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38(8):904–909.
Quadrana L, et al. 2016. The *Arabidopsis thaliana* mobilome and its impact at the species level. Elife 5:e15716.
R Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. 2016. Adaptation to global change: a transposable element-epigenetics perspective. Trends Ecol Evol. 31(7):514–526.
Schmidt JM, et al. 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genet. 6(6):e1000998.
Schmitz RJ, et al. 2013. Patterns of population epigenomic diversity. Nature 495(7440):193–198.
Schrader L, et al. 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun. 5(1):5495.
Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. 2014. Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. PLoS Genet. 10(11):e1004785.
Stapley J, Santure AW, Dennis SR. 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol. 24(9):2241–2252.
Stuart T, et al. 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. Elife 5:e20777.
Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet. 43(11):1160–1163.
Sun W, Shen YH, Han MJ, Cao YF, Zhang Z. 2014. An adaptive transposable element insertion in the regulatory region of the EO gene in the domesticated silkworm, *Bombyx mori*. Mol Biol Evol. 31(12):3302–3313.
Szpiech ZA, Hernandez RD. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 31(10):2824–2827.
The 1001 Genomes Consortium. 2016. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491.
Ullastres A, Petit N, Gonzalez J. 2015. Exploring the phenotypic space and the evolutionary history of a natural mutation in *Drosophila melanogaster*. Mol Biol Evol. 32(7):1800–1814.
Hof A. E v, et al. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534(7605):102–105.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4(3):e72.
Wei L, Cao X. 2016. The effect of transposable elements on phenotypic variation: insights from plants to humans. Sci China Life Sci. 59(1):24–37.
Wood JG, et al. 2016. Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in Drosophila. Proc Natl Acad Sci U S A. 113(40):11277–11282.
Ziolkowski PA, Koczyk G, Galganski L, Sadowski J. 2009. Genome sequence comparison of Col and Ler lines reveals the dynamic nature of Arabidopsis chromosomes. Nucleic Acids Res. 37(10):3189–3201.
Zou YP, et al. 2017. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. Genome Biol. 18(1):239.

Associate editor: Yves Van De Peer