The Role of Small RNA-Based Epigenetic Silencing for Purifying Selection on Transposable Elements in *Capsella grandiflora*

Robert Horvath and Tanja Slotte*

Science for Life Laboratory, Department of Ecology, Environment and Plant Sciences, Stockholm University, Sweden

*Corresponding author: E-mail: tanja.slotte@su.se.

Accepted: September 27, 2017

Abstract

To avoid negative effects of transposable element (TE) proliferation, plants epigenetically silence TEs using a number of mechanisms, including RNA-directed DNA methylation. These epigenetic modifications can extend outside the boundaries of TE insertions and lead to silencing of nearby genes, resulting in a trade-off between TE silencing and interference with nearby gene regulation. Therefore, purifying selection is expected to remove silenced TE insertions near genes more efficiently and prevent their accumulation within a population. To explore how effects of TE silencing on gene regulation shapes purifying selection on TEs, we analyzed whole genome sequencing data from 166 individuals of a large population of the outcrossing species *Capsella grandiflora*. We found that most TEs are rare, and in chromosome arms, silenced TEs are exposed to stronger purifying selection than those that are not silenced by 24-nucleotide small RNAs, especially with increasing proximity to genes. An age-of-allele test of neutrality on a subset of TEs supports our inference of purifying selection on silenced TEs, suggesting that our results are robust to varying transposition rates. Our results provide new insights into the processes affecting the accumulation of TEs in an outcrossing species and support the view that epigenetic silencing of TEs results in a trade-off between preventing TE proliferation and interference with nearby gene regulation. We also suggest that in the centromeric and pericentromeric regions, the negative aspects of epigenetic TE silencing are missing.

Key words: transposable elements, methylation, purifying selection, gene expression, epigenetic silencing, *Capsella*.

Introduction

Since Barbara McClintock first described mobile elements in the maize genome in 1950 (McClintock 1950; Nanjundiah 1996), transposable elements (TEs) have been found to be very frequent in the genomes of higher plants (Hollister et al. 2011). Despite being common, the movement and accumulation of TEs are viewed as disadvantageous for the host. This is mainly because new TE insertions can disrupt genes or other functional DNA sequences, increase the risk of ectopic recombination, and because TEs are linked to a metabolic cost for the host (Langley et al. 1988; Finnegan 1992; Badge and Brookfield 1997; Hollister and Gaut 2009; Lee 2015). However, other factors, like the epigenetic modification of a TE insertion, can also potentially be harmful for the host (Hollister and Gaut 2009; Lisch 2013; Lee 2015; Hirsch and Springer 2017).

The methylation of TEs is the result of the defense mechanisms of the host plant, which can downregulate the activity of TEs, thus preventing further proliferation (Lisch 2013; Marí-Ordóñez et al. 2013; Fultz et al. 2015). In plants, epigenetic silencing of TEs is achieved mainly via the RNA-directed DNA methylation (RdDM) pathway which includes the plant-specific RNA polymerase IV (Pol IV) and 24-nucleotide (nt) small interfering RNAs (siRNAs) (Marí-Ordóñez et al. 2013; Matzke et al. 2015; Fultz et al. 2015). The Pol IV RdDM pathway primarily initiates, maintains, and reinforces the methylation of TEs, which are targeted by 24-nt siRNAs (Fultz et al. 2015). The 24-nt siRNAs are usually derived from the TE mRNA of the targeted insertion, but these 24-nt siRNAs are also known to be able to *trans*-silence other highly similar TEs (Slotkin et al. 2005; Marí-Ordóñez et al. 2013; Fultz et al. 2015).

Hollister and Gaut (2009) showed that the regulation of TE activities through epigenetic silencing could cause a conflict between reducing transposition and interfering with the expression of neighboring genes. Since epigenetically silenced TEs have a distinct chromatin state from active functional DNA sequences, the spreading of chromatin modifications,

which were initially intended to silence a TE insertion, into neighboring genes and their *cis*-regulatory regions can silence these genes (Hirsch and Springer 2017). In the selfer *Arabidopsis thaliana*, where most TE variants are rare and associated with altered gene expression and methylation (Stuart et al. 2016), the genomic distribution of TEs is thought to be mainly governed by purifying selection acting against insertions in and close to genes (Wright et al. 2003; Hollister and Gaut 2009). However, the effect of purifying selection on TE insertions close to genes, especially on epigenetically silenced TEs, is less well studied in outcrossing plants (but see, e.g., Lockton et al. 2008; Lockton and Gaut 2010).

If purifying selection affects TEs differently depending on their epigenetics, then this should result in a deficit of silenced TEs within genomic regions where silencing TEs is linked to disadvantages. In addition, purifying selection is expected to skew the insertion frequency spectrum (IFS) of silenced TEs toward an increased proportion of rare insertions (Nielsen and Slatkin 2014). However, nonequilibrium demography, TE insertion biases or nonconstant transposition rates caused by recent TE bursts can also result in similar signatures (Barrón et al. 2014; Blumenstiel et al. 2014; Maumus and Quesneville 2014). For example, recent TE bursts are expected to increase the proportion of rare TEs through a rapid accumulation of young insertions (Barrón et al. 2014; Blumenstiel et al. 2014; Maumus and Quesneville 2014). Therefore, controlling for demographic effects and transposition rate variation is crucial for assessing the influence of purifying selection on TEs.

The outcrossing crucifer *Capsella grandiflora*, which is closely related to the self-fertilizing model plants *Capsella rubella* and *A. thaliana*, is a well-suited model system to quantify selection on TEs. The main benefits of studying *C. grandiflora* are its large and relatively constant effective population size without strong population structure (Foxe et al. 2009; Slotte et al. 2010; Douglas et al. 2015), in contrast to the outcrosser *Arabidopsis lyrata*, where strong population structure and historical bottlenecks need to be accounted for when inferring selection on TEs (Lockton et al. 2008; Lockton and Gaut 2010). In *C. grandiflora*, natural selection on both protein-coding and conserved noncoding genomic regions is efficient (Slotte et al. 2010; Williamson et al. 2014; Steige et al. 2017) and, in contrast to highly selfing species where TEs could be affected by linked selection, the outcrossing nature and rapid decay of linkage disequilibrium of *C. grandiflora* enables the study of local selection effects acting on TEs. Furthermore, whole genome sequencing data are available from large *C. grandiflora* populations (Josephs et al. 2015; Steige et al. 2017).

A previous study on *C. grandiflora* showed that the presence/absence of TEs in proximity of genes was associated with *cis*-regulatory gene expression variation (Steige et al. 2017). Additionally, analyses of allele-specific gene expression have shown that silencing of TEs near genes is associated with reduced expression of the allele on the same haplotype as the TE insertion in *Capsella* (Steige et al. 2015). Hence, there is accumulating evidence that TE silencing affects gene expression in *Capsella*, but how this shapes selection against TEs has not yet been thoroughly elucidated. In this study, we used whole genome resequencing data from 166 individuals of a large *C. grandiflora* population to investigate how interference between TE silencing and gene regulation affects selection on TE insertions.

Materials and Methods

Population Genomic Data Processing

We downloaded whole genome resequencing data from 200 individuals of a single *C. grandiflora* population, generated by Josephs et al. (2015) (NCBI accession number PRJNA275635, ID: 275635). We trimmed the paired-end 100-bp raw reads with Trimmomatic 0.32 (Bolger et al. 2014), and randomly subsampled reads of each individual to a total of 54 million reads per sample (average coverage 25×) to avoid an overrepresentation of individual samples with high coverage.

We mapped the trimmed reads to a TE-merged reference with bwa bwasw 0.7.13 (Li and Durbin 2009), as recommended before analyses with PoPoolationTE2 (Kofler et al. 2016). The TE-merged reference was based on the v1.0 *C. rubella* reference (masked with RepeatMasker 4.0.7; http://www.repeatmasker.org, last accessed December 1, 2016) and we used a library of TE sequences from Slotte et al. (2013). After removing 34 samples with poor mapping coverage (supplementary table S1, Supplementary Material online), we generated a pooled *C. grandiflora* population data set by merging the mapped reads for all remaining 166 individuals with SAMtools 1.3 (Li et al. 2009; Li 2011).

In order to compare TE insertion frequencies with site frequency spectra at 4-fold degenerate synonymous sites and 0-fold degenerate nonsynonymous sites, we mapped the trimmed and subsampled reads from the 166 *C. grandiflora* individuals included in the TE analyses to the *C. rubella* reference genome using bwa mem 0.7.13 (Li 2013). We then performed variant calling using GATK 3.3.0 (McKenna et al. 2010), SAMtools 1.3 (Li et al. 2009; Li 2011), and the Picard toolkit (http://broadinstitute.github.io/picard/; last accessed February 1, 2017). We called variants using Unified Genotyper (DePristo et al. 2011; Van der Auwera et al. 2013), and applied default hard filtering on the identified SNPs in order to remove poorly called sites. In addition, we removed all sites found in repeats detected in the *C. rubella* reference genome by RepeatMasker 4.0.7, using BEDTools 2.26.0 (Quinlan and Hall 2010). We further filtered the coverage depth of each site on an individual level to remove all allele calls with coverage <10 or >200 with VCFtools 0.1.15 (Danecek et al. 2011). To polarize alleles as ancestral or derived, we followed the procedure outlined in Laenen et al. (2017) relying on a three-way whole-genome alignment of

C. rubella, *A. thaliana*, and *A. lyrata* generated as in Steige et al. (2017).

To inspect whether there was evidence for major deviations from our assumption that *C. grandiflora* has a stable effective population size with no major population structure, we computed the polarized site frequency spectrum (SFS) of all 4-fold degenerate sites found in the population and compared it with the expected SFS for a standard neutral population. All 4-fold degenerate sites, where at least 300 of the 332 alleles were successfully assessed, were used to compute the polarized SFS in R 3.3.0 (R Development Core Team 2008). The expected SFS of neutrally evolving sites in a constant population was calculated by using the Watterson estimator for genetic diversity for 4-fold sites (Watterson 1975). In addition, we generated the polarized 0-fold degenerate SFS, which represents a site class where new mutations experience substantial purifying selection in *C. grandiflora* (Slotte et al. 2010; Williamson et al. 2014; Steige et al. 2017).

Identification of TEs Targeted by 24-Nucleotide Small RNAs

In order to distinguish TE insertions that are silenced and those that are not, we determined which insertions were targeted by siRNAs. Because a high-quality *C. grandiflora* reference genome assembly was lacking, we circumvented this issue by identifying insertions targeted by siRNAs in the closely related species *C. rubella*. *C. rubella* was derived from a *C. grandiflora*-like outcrossing ancestor fairly recently, most likely <200,000 years ago (Foxe et al. 2009; Guo et al. 2009; Brandvain et al. 2013; Slotte et al. 2013), and it is the best available model for inferring such information. Additionally, a previous study found no evidence for different TE silencing efficacies between *C. rubella* and *C. grandiflora* (Steige et al. 2015), hence, we expect silenced TEs in *C. rubella* to be also silenced in *C. grandiflora*.TEs targeted by siRNAs were determined by using small RNA sequencing data from roots, seedlings, and flowers of the *C. rubella* reference accession generated by Smith et al. (2015) (NCBI Accession: PRJNA212731, ID: 455735/456437/456438). We trimmed the raw sRNA reads with Trimmomatic 0.32 (Bolger et al. 2014) and mapped them to the TE-merged-reference using STAR 2.5.1b (Dobin et al. 2013), with default settings modified to allow mapping of small RNA reads. After removing all reads with more than one nucleotide soft-clipped from their 5′ end, in order to remove long RNA reads which were not fully mapped by STAR, we considered all RNA reads, which mapped to a TE, with a length of 24 nucleotides and no mismatches to be siRNAs. We considered all TEs with a minimum of 10 mapped siRNAs to be effectively targeted by siRNAs and silenced, and labeled them as siRNA+ TEs. The remaining TEs were labeled siRNA− TEs. Within the *C. grandiflora* population, all copies of a siRNA+ and siRNA− TEs were considered as siRNA+ and siRNA− TE insertions, respectively. Differentiating between uniquely and multi-mapping siRNAs to assess siRNA+ and siRNA− TE insertions yielded concordant results with the approach described earlier (supplementary fig. S1, Supplementary Material online) and the results are also robust to a different choice of cutoff for designating TEs as siRNA+ and siRNA− (supplementary figs. S2 and S3, Supplementary Material online) and to a restriction of the data to the eight main scaffolds of *C. grandiflora* (supplementary fig. S4, Supplementary Material online).

Purifying Selection on TEs

We used PoPoolationTE2 (Kofler et al. 2016) to identify and estimate the frequencies of TE insertions present in the *C. grandiflora* population. This method does not rely on the annotated TEs present in the reference genome and is therefore able to identify both novel and annotated TEs, as well as estimate their population frequencies. We analyzed our pooled population data set, following recommendations for PoPoolationTE2 analyses (Kofler et al. 2016) with slight modifications. Specifically, we conducted hard filtering of the TE insertions based on a minimum average physical coverage of 10, minimum average coverage of 10, maximum allowed frequency of other TEs of 0.2, and maximum allowed frequency of structural variants of 0.2. We used the results from PopoolationTE2 to generate insertion frequency spectra (IFS) of the TE insertions detected in the pooled *C. grandiflora* population data set. Distinguishing between retrotransposons (class I TEs) and DNA transposons (class II TEs) resulted in similar patterns for the IFS (supplementary figs. S5 and S6, Supplementary Material online), hence, we present results for all TEs here.

Interference of TE silencing with nearby gene regulation is expected to depend on the genomic region in which the TE is inserted. We therefore distinguished between TEs inserted in centromeric and pericentromeric regions, which are likely highly heterochromatic, and those inserted in the chromosome arms, which are expected to have a higher proportion of open chromatin, and where insertion of a silenced TE might be more likely to interfere with gene regulation. If siRNA+ TEs in chromosome arms are subject to stronger purifying selection than those in centromeric/pericentromeric regions, we expect the proportion of siRNA+ TEs to be lower in chromosome arms than in centromeric/pericentromeric regions. Genomic regions were assigned as likely centromeric/pericentromeric or representing chromosome arms (supplementary table S2, Supplementary Material online) as in Steige et al. (2015), and we tested for a difference in the proportion of siRNA+ TEs between these regions using binomial tests in R 3.3.0 (R Development Core Team 2008).

A dearth of siRNA+ TEs in chromosome arms could potentially be an effect of preferential insertion of TEs in centromeric and pericentromeric regions. We therefore used insertion frequency spectra to test whether there was evidence for stronger purifying selection on siRNA+ TEs near

genes. If the impact of TE silencing on gene expression results in stronger purifying selection on TEs, we expect siRNA+ TEs close to genes to have an increased proportion of rare insertions (Nielsen and Slatkin 2014). As the effects of TE silencing on nearby gene expression tend to be local and dissipate over a distance of > 2 kb (Hollister et al. 2011; Steige et al. 2015), we classified TEs as follows: 1) insertions within genes, 2) TE insertions flanking genes but not more than 1 kb away from genes, 3) TE insertions within 1–2 kb from a gene, and 4) TE insertions further than 2 kb away from genes (excluding all TE insertions on scaffolds without genes). To test for differences in purifying selection on TEs, we compared insertion frequency spectra (IFS) for siRNA+ and siRNA− TE insertions in different genomic locations. We conducted tests of differences in the proportion of rare (frequency < 0.02) insertions depending on TE epigenetic status, based on IFS of TEs found within, flanking, 1–2, 2–3, 3–5 and >5 kb from genes using a Wilcoxon rank sum test in R 3.3.0 (R Development Core Team 2008).

The Age-of-Allele Neutrality Test for TE Insertions

To account for effects of a nonconstant transposition rate on our IFS, we performed the age-of-allele test of neutrality for TE insertions proposed by Blumenstiel et al. (2014), on a subset of siRNA+ TE insertions found in the *C. grandiflora* population. Briefly, this method first infers the age of each TE insertion, using numbers of unique substitutions, and then uses information on TE age to infer the probability distribution of TE frequency in a neutrally evolving population. Results from this method can therefore be used to test for neutrality of TE insertions, without assuming a constant transposition rate.

For the age-of-allele neutrality test, we chose the retrotransposons *Gypsy 2395* and *Gypsy 2500*, which had high copy numbers (212 and 242, respectively) within the *C. rubella* reference genome and were both designated as siRNA+ TEs. We note that the use of the *C. rubella* reference genome limits our ability to examine selection on very recently inserted TEs, which are not likely to be shared by *C. grandiflora* and *C. rubella.* Nevertheless, these analyses are useful to assess whether our inference of purifying selection on TEs in *C. grandiflora* is robust when relaxing the assumption of a nonconstant transposition rate. We extracted the sequence of all *Gypsy 2395* and *Gypsy 2500* copies from the *C. rubella* reference genome, including an additional 500 base pair (bp) window on both sides of the insertions. The sequences were numbered, combined into a FASTA file and mapped to the *Gypsy 2395* and *Gypsy 2500* reference sequence provided in the library of TE sequences published by Slotte et al. (2013) with bwa mem 0.7.13 (Li 2013) and SAMtools 1.3 (Li et al. 2009; Li 2011), using the default settings. We sorted the mapped sequences by coordinates with the Picard toolkit (http://broadinstitute.github.io/picard/; last accessed February 1, 2017) and extracted the aligned sequence of each TE copy based on the CIGAR string of the corresponding mapped sequence. For sequences which mapped at multiple positions, we only considered the longest mapped sequence. The alignments were analyzed in R 3.3.0 (R Development Core Team 2008), and we assigned to each TE copy the number of unique sequence differences found only in that specific copy. For designating shared and unique sequence differences, we requested a minimum of five aligned copies per site. The number of unique sequence differences per *Gypsy 2395* and *Gypsy 2500* copy was used to carry out the age-of-allele test of neutrality for TE insertions following the procedure of Blumenstiel et al. (2014), in R 3.3.0 (R Development Core Team 2008). We only included TE copies which were present in the *C. grandiflora* population and where we could align at least 1,000 bp to the respective TE reference sequence. We used a mutation rate of 7×10^{-9} (Ossowski et al. 2010) and an effective population size of 500,000 (Douglas et al. 2015) in these analyses.

Results

Identification of TE Insertions in the *C. grandiflora* Population and Assessment of 4-Fold Site Frequency Spectra

We detected a total of 14,728 TE insertions in the pooled population resequencing data from 166 *C. grandiflora* individuals. Most of these TE insertions were rare (supplementary table S3, Supplementary Material online), and there were more siRNA+ than siRNA− insertions (12,958 siRNA+ and 1,770 siRNA− TE insertions). TEs from the superfamilies *Copia* and *Gypsy* were the most common, with 4,998 and 4,281 TE insertions, respectively (supplementary table S4, Supplementary Material online). There were a total of 7,798 TE insertions in centromeric/pericentromeric regions and 6,930 TE insertions in chromosome arms. In the centromeric/pericentromeric regions, we observed on an average a higher density of TE insertions than in the chromosome arms (153.1 vs. 82.6 TE insertions/Mbp, respectively).

To assess whether there was evidence for a deviation from demographic stability which could affect the interpretation of TE insertion frequency spectra, we computed the 4-fold degenerate polarized SFS and compared it with the expectation under the standard neutral model. As expected given previous inference of a relatively constant effective population size in *C. grandiflora* (Foxe et al. 2009; Slotte et al. 2013; Douglas et al. 2015; Steige et al. 2017), we observe a very good fit between the expected and observed SFS at 4-fold degenerate sites under neutrality and for a constant population size (supplementary fig. S7, Supplementary Material online).

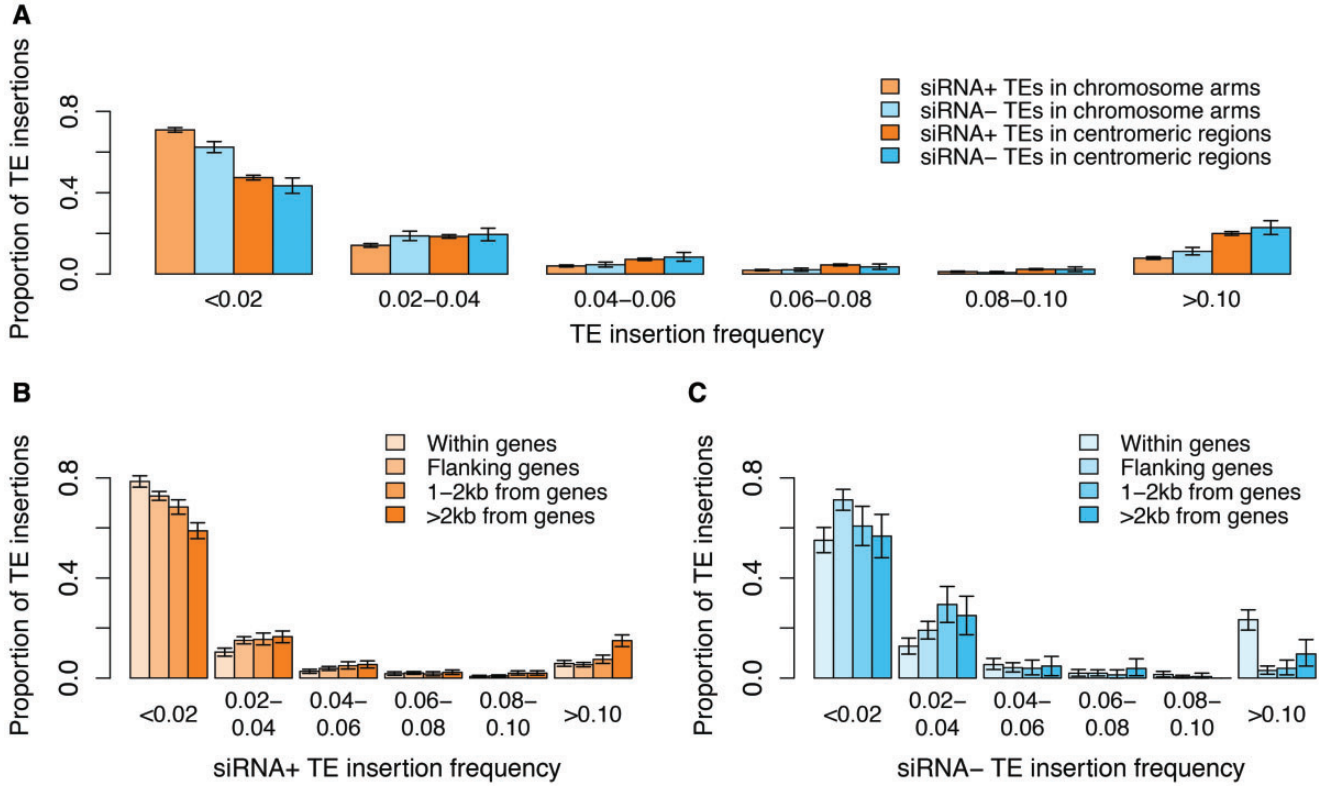

Fig. 1.—Insertion frequency spectrum (IFS) of siRNA + (orange) and siRNA − (blue) TE insertions. (*A*) IFS of siRNA+ and siRNA− TEs in centromeric/pericentromeric regions and chromosome arms. Error bars are 95% confidence intervals derived from 1,000 bootstrap replicates of each TE category. (*B*) ISF of siRNA+ TE insertions split into four different groups based on their position on the chromosome arms (within genes, flanking genes, 1–2 kb away from genes and >2 kb away from genes). (*C*) ISF of siRNA− TE insertions split into the same four groups. Error bars are 95% confidence intervals derived from 1,000 bootstrap replicates of each TE category.

TE Silencing Is Associated with Differences in the Genomic Distribution of TEs

Interference of TE silencing with nearby gene regulation could be expected to depend on the genomic region where the insertion occurs. Specifically, insertions in chromosome arms are more likely to result in interference with gene regulation than those in heterochromatic centromeric/pericentromeric regions. A signature of purifying selection against TEs due to their effects on gene expression is therefore expected to be a dearth of siRNA+ TEs in chromosome arms. Among all TE insertions, 7,151 siRNA+ and 647 siRNA− as well as 5,807 siRNA+ and 1,123 siRNA− TE insertions were located in the centromeric/pericentromeric regions and the chromosome arms, respectively. Overall, 55.2% of siRNA+ TE insertions (prop = 0.552, SE = 0.00437) were found in centromeric/pericentromeric regions, and this was significantly higher than the percentage of siRNA− TE insertions found in centromeric/pericentromeric regions (36.6%) (two-sided binomial test, P value < 0.001, $n = 12{,}958$). This agrees with the expectation if siRNA+ TEs in chromosome arms are under stronger purifying selection than siRNA− TEs, but could also be a result of TE insertion biases.

TE Silencing Is Associated with an Excess of Rare TE Insertions in Chromosome Arms

To further distinguish between possible TE insertion biases and varying purifying selection strength as causes of the genomic location of TEs, we computed the IFS of siRNA+ and siRNA− TE insertions in the two different genomic regions.

For TEs under stronger purifying selection, we expect an excess of rare insertions, and the excess of rare insertions should be positively correlated with the purifying selection strength affecting the TEs (Nielsen and Slatkin 2014). Both siRNA+ and siRNA− TEs have a significantly higher proportion of rare insertions in the chromosome arms than in the centromeric/pericentromeric regions (fig. 1*A*; two-sided empirical probability estimation approach, Benjamini–Hochberg adjusted P value <0.001 for both insertion types). This suggests that purifying selection against TEs is stronger in chromosome arms than in centromeric/pericentromeric regions for both types of TEs.

We further found a significantly higher proportion of rare siRNA+ TE insertions compared with siRNA− TE insertions in the chromosome arms (fig. 1*A*; two-sided empirical probability estimation approach, Benjamini–Hochberg adjusted

Table 1
Number and Proportion of siRNA+ and siRNA− TE Insertions Sorted by Their Distance to Genes within the Chromosome Arms

	Within Genes	Flanking Genes (<1 kb from Genes)	1–2 kb Away from Genes	>2 kb Away from Genes
siRNA+ TE insertions				
Number of insertions	1,471	2,397	904	962
Proportion (%)	25.6	41.8	15.8	16.8
siRNA− TE insertions				
Number of insertions	407	455	153	104
Proportion (%)	36.4	40.6	13.7	9.3

P value <0.001), whereas this was not the case in centromeric/pericentromeric regions. This agrees with our expectation if purifying selection is stronger on siRNA+ TEs than on siRNA− TEs specifically in chromosome arms, but not in centromeric/pericentromeric regions. Indeed, for siRNA+ TE insertions in chromosome arms, the proportion of rare alleles was higher than for 0-fold degenerate SNPs (supplementary fig. S7, Supplementary Material online), which are under strong purifying selection in *C. grandiflora* (Slotte et al. 2010; Williamson et al. 2014; Steige et al. 2017).

Increased Proportion of Rare siRNA+ TE Insertions in the Proximity of Genes

After identifying significant differences in the proportion of rare siRNA+ and siRNA− TE insertions in the chromosome arms, we further examined how the proportion of TE insertions change with the distance to the next gene on chromosome arms (see Materials and Methods).

We found 1,471, 2,397, 904, and 962 siRNA+ as well as 407, 455, 153, and 104 siRNA− TE insertions within the four different groups sorted by increasing distance to the next gene (table 1). There were significant differences in the observed proportion of siRNA+ and siRNA− TE insertions in genes, 1–2 kb, and >2 kb from genes (SE = 0.0048–0.0058, $n = 5734$, two-sided binomial test, Benjamini–Hochberg adjusted *P* value < 0.001), with siRNA+ TEs generally being less common in genes than siRNA− TEs (table 1). However, the proportion of TE insertions flanking genes did not differ significantly between siRNA+ and siRNA− TE insertions (SE = 0.0065, $n = 5734$, two-sided binomial test, NS).

Comparing the IFS of the four TE insertion groups revealed that siRNA+ TEs closer to genes had a higher proportion of rare insertions (fig. 1*B*; two-sided empirical probability estimation approach, Benjamini–Hochberg adjusted *P* value = 0.02 comparing siRNA+ TE insertions flanking and 1–2 kb from genes and Benjamini–Hochberg adjusted *P* value <0.001 for all other comparisons), as expected if siRNA+ insertions closer to genes are more strongly selected against than those further from genes. However, for siRNA− TEs, a similar pattern was found, with flanking TEs having more rare insertions than more distant TEs (fig. 1*C*; two-sided empirical probability estimation approach, Benjamini–Hochberg adjusted *P* value <0.001, *P* value =0.02, and *P* value =0.003 for siRNA− TE insertions in genes, 1–2 kb and >2 kb from genes, respectively), suggesting that other forces than TE silencing also affect purifying selection against TEs.

Differences in the Proportion of Rare siRNA+ and siRNA− TE Insertions in Genes

To investigate how purifying selection on TEs is mediated by the epigenetics of the TE insertions, we compared the proportion of rare insertions of siRNA+ and siRNA− TE insertions at varying distances from genes. If purifying selection is stronger on siRNA+ than on siRNA− TE insertions near genes, we expect to observe a greater proportion of rare siRNA+ TE insertions than the proportion of rare siRNA− TE insertions.

We found a significantly higher proportion of rare siRNA+ than siRNA− TE insertions (fig. 2, two-sided paired Wilcoxon rank sum test, *P* value =0.03), suggesting that siRNA+ TE insertions are overall under stronger purifying selection than siRNA− TEs in the chromosome arms. However, the largest difference was found for TEs within genes, not flanking genes.

siRNA+ TEs Are Further from Genes than siRNA− TEs

If siRNA+ TE insertions near genes are removed by purifying selection at a higher rate than siRNA− TE insertions, then we expect siRNA+ TE insertions in the chromosome arms to be further away from the next gene than siRNA− TE insertions. This was indeed the case (two-sided Wilcoxon rank sum test, *P* value <0.001), and the median distance between siRNA+ and siRNA− TE insertions and the next gene was 819 and 659 bp, respectively (supplementary fig. S8, Supplementary Material online).

Testing for Purifying Selection under Relaxed Transposition Rate Assumptions

To examine whether a nonequilibrium transposition rate could be exclusively responsible for our observations, we conducted a neutrality test that accounts for TE age and is thus robust to varying transposition rates. We focused on two siRNA+ retrotransposons, *Gypsy 2395* and *Gypsy 2500*, which were common in the *C. rubella* reference genome.

For these TEs, the age-of-allele test of neutrality for TE insertions (Blumenstiel et al. 2014) supported purifying selection against some old TE insertions (fig. 3) and purifying selection contributing to the IFS skews. Indeed, for the 28 TE insertions examined, 9 insertions had a frequency significantly smaller (*P* value <0.05) than expected under neutrality (fig. 3).

Assessing the Potential Impact of Different Ectopic Recombination Probabilities

Skews toward rare insertions in the IFS could also be the result of purifying selection acting against the accumulation of longer TEs, which are more prone to undergo ectopic recombination and lead to deleterious chromosomal rearrangements (Hollister and Gaut 2009; Lee 2015). However, siRNA+ TEs were significantly shorter than siRNA− TEs (two-sided Wilcoxon rank sum test, *P* value <0.001) and therefore the observed excesses of rare siRNA+ TEs cannot be explained by the length differences between siRNA+ and siRNA− TEs (supplementary fig. S9, Supplementary Material online).

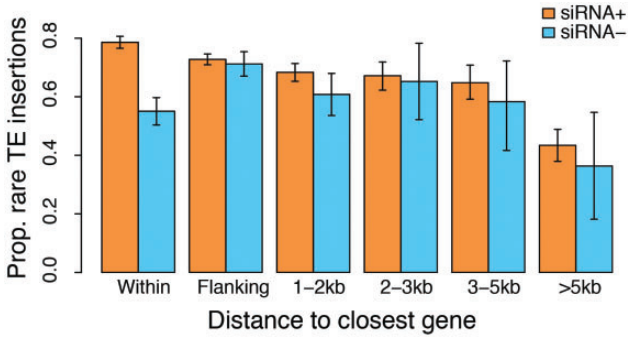

Fig. 2.—Proportion of the rarest (frequency < 0.02) siRNA + (orange) and siRNA − (blue) TE insertions found within, flanking, 1–2, 2–3, 3–5 and >5 kb away from genes. Error bars are 95% confidence intervals derived from 1,000 bootstrap replicates of each TE category.

Discussion

Here, we have investigated how interference between TE silencing and gene regulation affects the impact of purifying selection on TEs in a large population of the outcrossing plant *C. grandiflora*. By analyzing whole genome resequencing data in conjunction with small RNA data, we have investigated the genomic distribution and insertion frequency spectra of siRNA+ and siRNA− TEs. We found that overall, most TEs were rare, and that for chromosome arms, but not for centromeric and pericentromeric regions, there was an excess of low-frequency TE insertions specifically for siRNA+ TEs. This is in line with the hypothesis that in highly methylated and heterochromatic centromeric/pericentromeric regions, targeting of TEs for silencing by 24-nucleotide siRNAs is not likely to interfere with gene expression. Therefore, small RNA-based TE silencing is less likely to have negative side effects in this genomic region.

Within the chromosome arms, siRNA+ TE insertions were further away from genes than siRNA− TE insertions (supplementary fig. S8, Supplementary Material online), and siRNA+ TEs that were closer to genes had a higher proportion of rare insertions. In contrast, siRNA− TEs showed a more complicated pattern with respect to skews in the IFS (fig. 1*C*). Our results agree in general with expectations if the impact of purifying selection on siRNA+ TEs is stronger close to genes. Although different efficacies of removal of TEs by ectopic recombination may contribute to differences in the TE content of chromosome arms and centromeric/pericentromeric

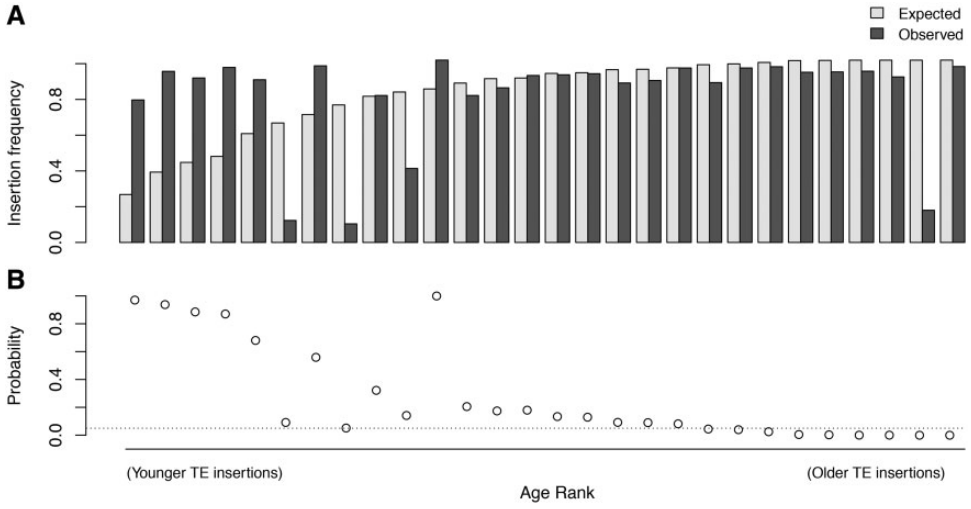

Fig. 3.—Results of the age-of-allele test of neutrality for TE insertions (*A*) Observed and expected TE insertion frequencies of the 28 evaluated TE copies in the *Capsella grandiflora* population, ranked by increasing insertion age. (*B*) Probability of observing an identical or lower TE insertion frequency, if the TE insertions are evolving neutrally. The dotted line represents the significance cutoff of 0.05.

regions, we found that siRNA− TEs were longer than siRNA+ TEs, hence, stronger purifying selection against siRNA+ TEs as a result of a higher risk of ectopic recombination for longer TEs can be excluded (supplementary fig. S9, Supplementary Material online). Thus, we conclude that the impact of TE silencing on neighboring gene regulation seems to be an important factor shaping selection against TEs in *C. grandiflora*.

Our results are in very good general agreement with previous findings described in *A. thaliana* (Hollister and Gaut 2009) and *Drosophila melanogaster* (Lee 2015) as well as with the general view that the location of a silenced insertion is the most important factor when it comes to the side effects of the host initiated TE silencing process (Lisch 2013; Sigman and Slotkin 2016; Hirsch and Springer 2017). Consistently with previous studies on *A. thaliana* (Wright et al. 2003; Hollister and Gaut 2009), 87.2% of all TEs found in *C. grandiflora* were outside of genes. The genome-wide proportion of siRNA− TEs was 12% in *C. grandiflora*, which is similar to the proportion in *A. thaliana* (15%; Hollister and Gaut 2009). Like in *A. thaliana* (Hollister and Gaut 2009), the proportion of siRNA− TE insertions relative to both TE types within a specific genomic region was the highest within genes (21.7%) and decreased with an increasing distance to genes until reaching approximately genome-wide proportions >2 kb away from genes (9.8%). Additionally, the distribution of siRNA+ TE insertions revealed a significant underrepresentation of siRNA+ TEs in genes (table 1), like in *A. thaliana* (Hollister and Gaut 2009). Finally, the median distance to the next gene was significantly higher for siRNA+ than for siRNA− TE insertions, which was also the case in *A. thaliana* (Hollister and Gaut 2009). Although we did not directly assess local methylation status, in *A. thaliana* it has been shown that small RNA targeting is a good proxy for methylation status (Hollister and Gaut 2009).

We found no strong support for major deviations from demographic equilibrium based on comparisons of expected and observed 4-fold synonymous site frequency spectra under a standard neutral model, suggesting that demographic changes should have a limited effect on our TE insertion frequency spectra (supplementary fig. S7A, Supplementary Material online). We also investigated the possibility of purifying selection affecting our IFS under a relaxed transposition rate assumption by a detailed analysis using a method that accounts for TE age. Although these results should be seen as tentative, as they are based on a low number of TE insertions, they suggest that negative selection is more important than variation in TE transposition rates in explaining TE insertion frequencies in *C. grandiflora* (fig. 3).

To investigate the impact of our bioinformatic procedure to classify TEs as siRNA+ or siRNA−, we increased the mapped siRNA cutoff, but the results remained similar (supplementary figs. S2 and S3, Supplementary Material online). In particular, one unexpected result remained regardless of the cutoff used, namely the excess of rare siRNA− TE insertions flanking genes compared with the other three siRNA− TE insertion groups (fig. 1*C*). This would suggest that for siRNA− chromosome arm TE insertions, purifying selection has a similar impact on insertions in genes, 1–2 kb from genes and >2 kb from genes but that siRNA− TE insertions flanking genes are under stronger purifying selection. Although it is possible that gene flanking insertions are more likely to disrupt functional DNA sequences like *cis*-regulatory regions than insertions further away from genes, this does not explain the observed difference in IFS between insertions flanking genes and those in genes. It is however possible that we missed or falsely assigned some silenced TE insertions, because we used small RNA data from *C. rubella* to discriminate siRNA+ and siRNA− TE insertions in *C. grandiflora*. Additionally, false positive/negative TE identifications and inaccurate insertion frequency estimations are possible as a result of challenges resulting from identifying nonreference TE insertions (Kofler et al. 2016). Although simulations have shown that PoPoolationTE2 can have an error rate of up to 4.8% for TE identification, the accuracy of estimates of TE frequencies was in general higher for TEs with population frequencies ∼0.1 or 0.9 (mean estimation deviation ∼0.01) (Kofler et al. 2016). However, we do not expect such errors to depend on the epigenetic state of TEs, and our results were further robust to different siRNA+/siRNA− assessments and to reanalyses using subsampled data (supplementary figs. S1–S6, Supplementary Material online). Thus, errors in TE identification likely have a minor effect on our results. Future studies should generate small RNA data directly from the studied *C. grandiflora* individuals and improvement of TE identification software should mitigate such effects.

The TE age assessment in *C. grandiflora* based on *C. rubella* could also be a source of error, because losses or gains of TE copies in in *C. rubella* would lead to over- or underestimation of the age of TE copies (Blumenstiel et al. 2014). Additionally, not all copies of the *Gypsy 2395* and *Gypsy 2500* retrotransposons could be properly aligned to their respective references, which could also lead to incorrect age estimates. However, all TE insertions which were used for the neutrality test were relatively old, with age estimates ranging from circa 240,000 to 6,800,000 years. This was expected since we only used TE insertions which were present in *C. grandiflora* and *C. rubella*, therefore, we expect the last common ancestor of these two species to have harbored these insertions. We argue that determining the age of the TE insertions found in *C. grandiflora* by estimating their age in *C. rubella* yielded, despite some alignment issues, reliable results for these particular TEs. Additionally, because we excluded all insertions found in *C. rubella* but absent in *C. grandiflora*, we probably failed to include some TE insertions which were completely removed by selection from the *C. grandiflora* population. However, we decided not to include these insertions in our analyses because we could not distinguish between insertions which are effectively absent in the *C. grandiflora* population and

insertions that we failed to detect. As higher quality *C. grandiflora* assemblies based on long-read technology become available, permitting detailed analyses of a broader range of TEs (as in, e.g., Jiao et al. 2017), it will be important to revisit these results and assess how generally they apply to other TE families.

In this study, we first showed that in centromeric and pericentromeric regions, TE silencing is not affecting purifying selection against TEs. Second, silenced TEs are under significantly stronger purifying selection than other TEs in the chromosome arms, and silenced TEs close to genes are under significantly stronger purifying selection than those further away from genes, and are also on an average found farther away from genes than TEs that are not silenced. Previously, Steige et al. (2017) showed that the presence/absence of TEs in the proximity of genes was associated with *cis*-regulatory gene expression variation in *C. grandiflora* and silenced TEs were associated with lower allele-specific gene expression in *Capsella* F1 hybrids (Steige et al. 2015). Based on these results and our findings, we conclude that gene expression variation caused by silenced TEs results in decreased fitness leading to stronger purifying selection against silenced TEs near genes. These findings contribute to our understanding of the evolutionary process governing TE accumulation in outcrossing plants and suggest that increased selection against silenced TEs in chromosome arms results from a tradeoff between reduced TE activity and nearby gene expression silencing, as hypothesized by Hollister and Gaut (2009).

Supplementary Material

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

Acknowledgments

The computational analyses of this work were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b2014146. This work was supported by grants from the Swedish Research Council and the Science for Life Laboratory to T.S. The authors thank Benjamin Laenen, Douglas Scofield, and Kim Steige for help, comments, and discussions.

Literature Cited

Badge RM, Brookfield JFY. 1997. The role of host factors in the population dynamics of selfish transposable elements. J Theor Biol. 187(2):261–271.

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

Blumenstiel JP, Chen X, He M, Bergman CM. 2014. An age-of-allele test of neutrality for transposable element insertions. Genetics 196(2):523–538.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G. 2013. Genomic identification of founding haplotypes reveals the history of the selfing species *Capsella rubella*. PLoS Genet. 9(9):e1003754.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

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

Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21.

Douglas G, et al. 2015. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. Proc Natl Acad Sci U S A. 112(9):2806–2811.

Finnegan DJ. 1992. Transposable elements. Curr Opin Genet Dev. 2(6):861–867.

Foxe JP, et al. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. Proc Natl Acad Sci U S A. 106(13):5241–5245.

Fultz D, Choudury SG, Slotkin RK. 2015. Silencing of active transposable elements in plants. Curr Opin Plant Biol. 27:67–76.

Guo YL, et al. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc Natl Acad Sci U S A. 106(13):5246–5251.

Hirsch CD, Springer NM. 2017. Transposable element influences on gene expression in plants. Biochim Biophys Acta 1860(1):157–165.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19(8): 1419–1428.

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.

Jiao Y, et al. 2017. Improved maize reference genome with single-molecule technologies. Nature 546(7659):524–527.

Josephs EB, Lee YW, Stinchcombe JR, Wright SI. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proc Natl Acad Sci U S A. 112(50):15390–15395.

Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. Mol Biol Evol. 33(10):2759–2764.

Laenen B, et al. 2017. Demography and mating system shape the genome-wide impact of purifying selection in *Arabis alpina*. bioRxiv. doi: https://doi.org/10.1101/127209.

Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B. 1988. On the role of unequal exchange in the containment of transposable element copy number. Genet Res. 52(3):223–235.

Lee YCG. 2015. The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster*. PLoS Genet. 11(6):1–24.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN], uploaded on May 26, 2013.

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

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.

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

Lockton S, Gaut BS. 2010. The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC Evol Biol. 10:10.

Lockton S, Ross-ibarra J, Gaut BS. 2008. Demography and weak selection drive patterns of transposable element diversity in natural populations of *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 105(37):13965–13970.

Marí-Ordóñez A, et al. 2013. Reconstructing de novo silencing of an active plant retrotransposon. Nat Genet. 45(9):1029–1039.

Matzke MA, Kanno T, Matzke AJM. 2015. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. Annu Rev Plant Biol. 66:1–25.

Maumus F, Quesneville H. 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. Nat Commun. 5:4104.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36(6):344–355.

McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.

Nanjundiah V. 1996. Barbara McClintock and the discovery of jumping genes. Resonance 1(10):56–62.

Nielsen R, Slatkin M. 2014. An introduction to population genetics: theory and applications. Sunderland (MA): Sinauer Associates, Inc. Publishers.

Ossowski S, et al. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327(5961):92–94.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Sigman MJ, Slotkin RK. 2016. The first rule of plant transposable element silencing: location, location, location. Plant Cell 28(2):304–313.

Slotkin RK, Freeling M, Lisch D. 2005. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nat Genet. 37(6):641–644.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol. 27(8):1813–1821.

Slotte T, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 45(7):831–835.

Smith LM, et al. 2015. Rapid divergence and high diversity of miRNAs and miRNA targets in the Camelineae. Plant J. 81(4):597–610.

Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T. 2017. Genomic analysis reveals major determinants of *cis*-regulatory variation in *Capsella grandiflora*. Proc Natl Acad Sci U S A. 114:1087–1092.

Steige KA, Reimegård J, Koenig D, Scofield DG, Slotte T. 2015. Cis-regulatory changes associated with a recent mating system shift and floral adaptation in *Capsella*. Mol Biol Evol. 32(10):2501–2514.

Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. 2016. Population scale mapping of transposable elements reveals links to gene regulation and epigenomic variation. eLife 5:e20777.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, et al. 2013. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–11.10.33.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7(2):256–276.

Williamson RJ, et al. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. PLoS Genet. 10(9):e1004622.

Wright SI, Agrawal N, Bureau TE. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res. 13(8):1897–1903.

Associate editor: Susanne Renner