Widespread Interspecific Divergence in Cis-Regulation of Transposable Elements in the Arabidopsis Genus

Fei He,1,2 Xu Zhang,3 Jin-Yong Hu,1 Franziska Turck,1 Xue Dong,1 Ulrike Goebel,1 Justin O. Borevitz,3 and Juliette de Meaux*1,2

1Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany
2Institute for Evolution and Biodiversity, University of Münster, Münster, Germany
3Department of Ecology and Evolution, University of Chicago

*Corresponding author: E-mail: juliette.de.meaux@uni-muenster.de.

Abstract

Transposable elements (TEs) are so abundant and variable that they count among the most important mutational sources in genomes. Nonetheless, little is known about the genetics of their variation in activity or silencing across closely related species. Here, we demonstrate that regulation of TE genes can differ dramatically between the two closely related Arabidopsis species A. thaliana and A. lyrata. In leaf and floral tissues of F1 interspecific hybrids, about 47% of TEs show allele-specific expression, with the A. lyrata copy being generally expressed at higher level. We confirm that TEs are generally expressed in A. lyrata but not in A. thaliana. Allele-specific differences in TE expression are associated with divergence in epigenetic modifications like DNA and histone methylation between species as well as with sequence divergence. Our data demonstrate that A. thaliana silences TEs much better than A. lyrata. For long terminal repeat retrotransposons, these differences are more pronounced for younger insertions. Interspecific differences in TE silencing may have a great impact on genome size changes.

Key words: transposable elements, interspecific cis-regulatory divergence, expression variation, epigenetic modification, Arabidopsis.

Introduction

The proportion of transposable elements (TEs) varies dramatically across species. In plants, it ranges from about 80% of the genome in maize to only 10% in the particularly compact genome of Arabidopsis thaliana (Kidwell and Lisch 1997). This variation may result either from differences in the efficiency of selection for purging transposed elements or from species-specific differences in transposition rate (Le Rouzic et al. 2007). However, little is known about the molecular mechanisms controlling interspecific variation in transposition, especially in closely related species (Feschotte and Pritham 2007; Lisch 2009). In the genus Arabidopsis, the two species A. lyrata and A. thaliana differ by ~40% in genome size (Johnston et al. 2005; Oyama et al. 2008). As much as 56% of the A. lyrata genome sequence not present in A. thaliana encodes for TEs or simple repeats (Hu et al. 2011). The control of TEs in the model plant species A. thaliana has been studied in detail (Zilberman and Henikoff 2004, 2005; Weil and Martienssen 2008), and the most TEs reside in chromosomal regions that are transcriptionally inactive (Mirouze et al. 2009; Tsukahara et al. 2009).

The activity of TEs often associates with hybridization, polyploidy, or stresses (Blanc et al. 2003; Feschotte and Pritham 2007; Lockton and Gaut 2010; Ito et al. 2011; Kawakami et al. 2011). For example, ONSEN, a copia-type retrotransposon, can be activated by heat stress and synthesize extrachromosomal DNA copies in A. thaliana seedlings (Ito et al. 2011). Several studies pointed to important differences in the control of TE activity between A. thaliana and A. lyrata. In the outcrossing species A. lyrata, the strength of negative selection against new transposition has been found to be weak (Wright et al. 2001, 2003; Lockton et al. 2008; Lockton and Gaut 2009), and new transposition variants are not segregating at lower frequency than in the inbred species A. thaliana (Lockton et al. 2008; Lockton and Gaut 2010). Furthermore, the recent completion of the A. lyrata genome sequence revealed that many transposon insertions are recent in this lineage (Hu et al. 2011). In A. thaliana, cis-regulatory silencing of TEs is mediated by small interfering RNAs (siRNAs), which guide the deposition of methylation marks on homologous DNA stretches (Daxinger et al. 2009; Havecker et al. 2010; Law and Jacobsen 2010). Interspecific differences in siRNAs may cause differences in the efficiency of TE silencing (Hollister et al. 2011). Here, we monitored cis-regulatory variation in the expression 1,535 TEs and tested whether cis-regulatory silencing of TEs differs between A. lyrata and A. thaliana. We confirm that the epigenetic silencing of TEs differs drastically between the two species and explore the genomic characteristics of these differences.
Materials and Methods

Plant Material
Seeds of A. thaliana parental accessions Be-0 (accession number CS28062) were obtained from the Arabidopsis Biological Resource Center (ABRC, USA). The A. lyrata ssp. petraea parental genotype LF-10 (collected from Lilienfeld, Austria) was provided by Maria J. Claus (Max Planck Institute for Chemical Ecology, Jena, Germany). Parental lines were crossed with A. thaliana as mother and A. lyrata as father, as described in de Meaux et al. (2006). Reciprocal crosses using pollen from A. thaliana father, as described in de Meaux et al. (2006). Reciprocal crosses using pollen from A. thaliana father, as described in de Meaux et al. (2006). Reciprocal crosses using pollen from A. thaliana father, as described in de Meaux et al. (2006).

Sample Preparation and Microarray Hybridization
Three leaves from 4-week-old plants (the three youngest leaves greater than 0.5 cm), 3–4 open flowers, and 3–4 closed flower buds were collected from each of four individual F1 hybrids. As A. lyrata ssp. petraea is outcrossing, all four F1 individuals are not genetically identical. Genomic DNA and total RNA were isolated using the corresponding plant mini kits (Qiagen). Double-stranded cDNA was generated from RNA samples, and each DNA and cDNA sample was hybridized onto the AtSNPtile 1 array as described by Zhang and Borevitz (2009). This array contains approximately 250,000 sets of four probes designed to interrogate each one single nucleotide polymorphism (SNP) (Atwell et al. 2010; Zhang et al. 2011). Each SNP probe set comprises a sense and an antisense pair of 25 bp allelic probes differing in their central nucleotide. In other words, the four probes can be described as: allele1-sense, allele1-antisense, allele2-sense, and allele2-antisense, with allele 1 and 2 differing by one nucleotide in position 13 of the 25 bp oligonucleotide.

Identification of Interspecific Differences in TE Cis-Regulation
All scripts were written in R and are available upon request. Raw probe intensities were corrected for background noise and normalized as previously described (Borevitz et al. 2003; Zhang and Borevitz 2009). For each array hybridization, two SNP probe ratios (SPRs) were calculated for each probe set by calculating the log2 ratio of intensities of the sense probe of allele1 to the sense probe of allele2 and of the antisense probe of allele1 to the antisense probe of allele2. In the analysis reported here, the sense and antisense ratios were considered as independent measures of allelic expression differences. Analyses restricted to either the sense or the antisense probes gave the same results (ca. 90% overlap, not shown).

To identify SNPs informative to discriminate A. thaliana and A. lyrata alleles, we selected SNP probe sets for which each Be-0 and LF-10 DNA hybridized preferentially with alternative allelic probes. For this, probe sets have to fulfill the following conditions: SPR_{Be-0} < SPR_{f1} < SPR_{LF-10} and SPR_{Be-0} < 0 < SPR_{LF-10}, in which case Be-0 preferentially hybridizes with allele2 and LF-10 with allele1; or SPR_{Be-0} > SPR_{f1} > SPR_{LF-10} and SPR_{Be-0} > 0 > SPR_{LF-10}, in which case Be-0 preferentially hybridizes with allele1 and LF-10 with allele2. Excluding cases in which the SPR of parental DNA had the same sign ensures that parental alleles display a preference each for a distinct allelic probe in an SNP probe set. Differences were tested with the method developed by Tusher et al. (2001), and the false discovery rate (FDR) rate was set to 0.05. In this way, 63,395 of the 250,000 SNP probe sets contained on the microarray were identified as informative SNP probe sets. Of these, 3,994 SNP probe sets were located in 1,535 transcribed regions annotated as TEs (TAIR9 annotation, www.arabidopsis.org).

Genomics Analysis
Lists of genes enriched in various histone or DNA methylation marks were extracted from reports published by Turck et al. (2007) and Lister et al. (2008), respectively. Overlaps were tested by hypergeometric tests. For the analysis of sequence divergence patterns associating with ASE-TEs, sequences of the 1,535 TE genes were downloaded from TAIR9 and blasted against A. thaliana genome and the assembled A. lyrata genome (Hu et al. 2011). In total, 885 genes annotated as TEs presented extended homologies between species. Analysis of sequence divergence patterns associating with ASE-TEs were performed on this subset of sequences. GC content was calculated in either the coding region or the promoter (1,000 bp upstream to transcription starting site, TSS). CG enriched regions were defined as 100 bp bins with over 50% of GC and with observed/expected CG dinucleotide ratio of 60%. Differences in the number of CG enriched regions was tested by Fisher’s exact test, and the Kolmogorov–Smirnov test was applied to establish the significance of differences in the distribution of CG enriched regions upstream from the TSS. To identify nucleotide motifs associating with ASE-TEs, we downloaded 469 cis-regulatory elements of plant from Cis-PtACE (A Database of Plant Cis-acting Regulatory Elements, 2004) and the publicly available Arabidopsis genome sequences (TAIR9, 2007) for each TE-box in the 469 cis-regulatory elements.
DNA Elements, Higo et al. 1999) and counted their occurrence in the 1,000 bp window upstream to TSS A. thaliana TE genes. Fisher’s exact test was used to establish significant differences in motif counts. New motif search was performed by the motif-finding program Multiple Em for Motif Elicitation (MEME) (Bailey and Elkan 1994).

Copy number estimates were compiled based on local Blast results using the following parameters: a minimum length of 100 bp, ≥20% of the length of the query, >80% similarity between query and hit, and E value cut off 1 × 10⁻30. Orthologous sequence was identified if the TE and its two neighbor genes were collinear in A. thaliana and A. lyrata. To analyze the relationship between TE expression divergence and age of insertions, we focused on long terminal repeat (LTR) retrotransposons because they are the most abundant TEs in Arabidopsis genomes and their age can be reliably estimated by divergence between flanking terminal repeats. Age estimates in A. lyrata and A. thaliana were taken from Hollister et al. (2011), and we restricted the data set to TEs showing extended homologies between species. Age of ASE and non-ASE TEs were compared only within each genome because age estimates of orthologous TEs were much greater in A. lyrata than in A. thaliana. This is an indication that interspecific differences alter the divergence rate of terminal repeats and that a comparative analysis of TE age estimates is more robust within genome.

Expression of TEs
Allele-specific relative expression levels were quantified at eight TE genes using the quantitative properties of pyrosequencing following previously described protocols (Neve et al. 2002; de Meaux et al. 2005). Primers for the pyrosequencing assay were designed based on Sanger sequencing of the parental alleles. Loci are described in supplementary table 1 (Supplementary Material online). At least eight independent biological RNA samples were collected from closed buds or open flowers in progenies of the same parental lines. RNA extraction and cDNA synthesis were performed using standard protocols we have described previously (de Meaux et al. 2005).

The expression level of TEs in parental lines and F1 progeny was examined by semiquantitative reverse transcription—polymerase chain reaction (RT-PCR). Total RNA was extracted from flower tissues with TRIzol (Invitrogen) and treated with DNA-free kit (Ambion, Applied Biosystems). Two micro gram of total RNA was used to synthesize first strand cDNA by random primers with SuperScript III Reverse Transcriptase (Invitrogen). A one-tenth aliquot of the reverse transcription reaction was used as a template for polymerase chain reaction (PCR) (total 10 μl). The PCR conditions were described as follows: 94 °C for 2 min, then 35 cycles of 94 °C for 20 s, 58 °C for 20 s, and 72 °C for 45 s, and then 72 °C for 3 min. The PCR product was then separated by electrophoresis on 1.5% agarose. The primer pairs used for RT-PCR and pyrosequencing are shown in supplementary tables 1 and 2 (Supplementary Material online).

Methylation Analysis
Genomic DNA was extracted from three leaves of 4-week-old plants (three youngest leaves greater than 0.5 cm) as the samples for microarray hybridization. DNA methylation was revealed by bisulfite sequencing. Bisulfite conversion was performed using the Epitect kit (Qiagen). PCR products were cloned using the pGEM-T vector (Promega). For each sample, 8–12 clones were sequenced. The proportion of methylated Cytosine in CG, CNG, and CNN contexts was calculated for the whole sequenced regions (AT3G0837, −423 to +67; AT4G06658, −1,400 to −600). Both of sequenced regions covered CG islands. Primers are listed in supplementary table 2 (Supplementary Material online).

Results and Discussions
Genome-Wide ASE of TEs
The analysis of allele-specific gene expression in F1 hybrids has provided a sensitive way to detect cis-regulatory differences within and between species (de Meaux et al. 2006; Wittkopp et al. 2008; Zhang and Borevitz 2009). Indeed, in such individuals, both parental genomes experience the same trans-regulatory environment. Differential expression of one of the parental copies demonstrates the existence of functional differences encoded genetically or epigenetically at the locus of the expressed gene. Such functional cis-regulatory difference can be caused by nucleotide change or local chromatin modification and will be hereafter designated as ASE. We used a custom-designed array, AtSNPtile 1, designed to interrogate 250,000 SNPs in A. thaliana with four probes each (Zhang and Borevitz 2009; Atwell et al. 2010; Zhang et al. 2011). This tool has previously enabled the genome-wide characterization of cis-regulatory variation segregating within A. thaliana (Zhang and Borevitz 2009; Zhang et al. 2011).

We generated F1 interspecific hybrids by crossing the diploid paternal A. lyrata ssp. petraea genotype LF-10 and the maternal A. thaliana genotype Be-0. We then monitored genome-wide cis-regulatory divergence between these genotypes by assessing the relative expression level of each parental copy. For this, we proceeded in two steps. First, we identified informative probe sets for which each parental allele hybridizes preferentially to distinct probes at a given SNP position (see Materials and Methods). A total of 63,395 informative SNPs (~25% of 250,000) fulfilled this criterion. Of these, 3,994 SNPs were located in the exonic regions of 1,535 TE genes, allowing the assessment of ASE patterns in about 39% of the 3,908 TE genes annotated in the genome of A. thaliana (TAIR9). TEs with reliable SNP probe sets were longer on average than those without reliable SNP probe sets (2,905 vs. 2,411 bp, respectively.

Downloaded from https://academic.oup.com/mbe/article-abstract/29/3/1081/1009031 by guest on 02 February 2019
Kolmogorov–Smirnov test, \( P = 2.767 \times 10^{-12} \). In addition, two families of TEs were significantly enriched. Approximately 55.4% of \( \textit{copia} \)-like retrotransposons and 72.2% non-LTR retrotransposons harbored at least one reliable SNP, a proportion significantly higher than 39% (Fisher’s exact test \( P = 4.95 \times 10^{-10} \) and \( 7.49 \times 10^{-29} \), respectively, supplementary table 3, Supplementary Material online). TEs with SNPs also displayed significantly lower GC content (40.7% vs. 42.0%, Kolmogorov–Smirnov test, \( P < 2.2 \times 10^{-16} \)). Subsequent analyses were all performed on the subset of TEs that could be interrogated with the array.

In a second step, we identified loci with ASE differences. Relative expression levels were measured as the log value of the allelic SNP probe ratio (SPR). Among 1,535 annotated TEs that could be reliably monitored, 463, 604, and 449 were ASE-TEs in young flower buds, open flower buds, and leaves, respectively (fig. 1, FDR 1%). In total, 722 TEs (46.8% of those that could be interrogated) were differentially expressed in at least one tissue or developmental stage (supplementary table 4, Supplementary Material online). Among these, 311 TEs displayed ASE in all three tissue samples, suggesting that differential expression is often constitutive for these genes (fig. 1). Using data from a similar analysis conducted on protein-coding genes (He F, Zhang X, Hu J, Turck F, Dong X, Goebel U, Borevitz J, and de Meaux J, forthcoming), we observed that TE genes are dramatically enriched among loci displaying tissue-specific \( \textit{cis} \)-regulatory divergence (Fisher’s exact test, \( P < 10 \times 10^{-108} \), table 1).

Our array-based method assumes that the \( \textit{A. lyrata} \) and \( \textit{A. thaliana} \) cDNA or DNA fragments hybridizing with alternative allelic probes of a probe set are alleles. Yet, the array was designed based on the \( \textit{A. thaliana} \) genome and does not consider rearrangements that may have occurred in each genome. It remains possible that the alleles observed as hybridizing to the same probe set of the array are in fact independent insertions of homologous transposons. Our study therefore does not exclude the possibility that \( \textit{cis} \)-regulatory divergence could result from interspecific differences in the genomic locations of the TE copies. However, F1 DNA signals were used as reference to determine expression divergence: This controls for copy number differences so that significant SPR differences are differences in expression. In addition, in F1 hybrids, all TE copies are expressed in the same cells. Therefore, even if they potentially correspond to copy-specific expression differences, the ASE differences described above still result from functional differences in \( \textit{cis} \)-regulation.

Preferential Expression of the \( \textit{A. Lyrata} \) Allele of TEs

Variation in TE \( \textit{cis} \)-regulation was not randomly distributed. Almost all ASE-TEs (458 of 463 in young flower buds, 594 of 604 in open flowers, 447 of 449 in leaves) preferentially expressed the \( \textit{A. lyrata} \) allele (fig. 2). This clearly indicates that preferential hybridization of \( \textit{A. thaliana} \) alleles on the array has little influence on the ASE patterns detected. Intriguingly, differential expression did not hit all TE families identically. Indeed, 2 of 12 TE families were significantly overrepresented. More than 55% of the 273 \( \textit{copia} \)-like retrotransposons (Fisher exact test, FDR = 0.011) and 60.5% of the 195 non-LTR retrotransposons (long interspersed nuclear elements; FDR = 0.001) displayed interspecific \( \textit{cis} \)-regulatory divergence (supplementary table 5, Supplementary Material online). Most of these differentially regulated elements were distributed in pericentromeric regions where TEs normally accumulate (Wright et al. 2003). To validate the accuracy of our predictions, we randomly selected eight ASE-TE genes in both flower buds and open flowers (supplementary table 1, Supplementary Material online) and confirmed their preferential expression of the \( \textit{A. lyrata} \) allele. In a parallel study, genes coding for non-TE proteins were analyzed, and the microarray predictions could also be validated by pyrosequencing for 19 of 20 loci (He F, Zhang X, Hu J, Turck F, Dong X, Goebel U, Borevitz J, and de Meaux J, forthcoming), which confirms the accuracy of the microarray for monitoring ASE.

Table 1. TEs Are Significantly Overrepresented among Genes Showing ASE, That Is, \( \textit{cis} \)-Regulatory Differences.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>ASE-TE No.</th>
<th>Total TE No.</th>
<th>ASE-Gene No.</th>
<th>Total Gene No.</th>
<th>Fisher exact P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flower bud</td>
<td>463 (30.2%)</td>
<td>1,535</td>
<td>1236 (7.6%)</td>
<td>16,175</td>
<td>3.99 \times 10^{-176}</td>
</tr>
<tr>
<td>Open flower</td>
<td>604 (29.3%)</td>
<td>1,535</td>
<td>1995 (12.3%)</td>
<td>16,175</td>
<td>5.66 \times 10^{-183}</td>
</tr>
<tr>
<td>Leaf</td>
<td>449 (39.3%)</td>
<td>1,535</td>
<td>1662 (10.3%)</td>
<td>16,175</td>
<td>1.16 \times 10^{-108}</td>
</tr>
</tbody>
</table>

Note.—The proportion of the category number to the whole genome is indicated in brackets.
Variation of TE Expression Across Species

In interspecific hybrids, the upregulation of TEs has been commonly reported, possibly participating in hybrid dysgenesis (Landry et al. 2007; Michalak 2009). In *A. thaliana*, siRNA expression in the vegetative nuclei of the maternal gametophytes prior to fertilization can mediate DNA methylation to active repeats (Mosher et al. 2009; Slotkin et al. 2009; Mosher and Melnyk 2010). TE content in *A. lyrata*, the paternal genotype, differs much from TE content in *A. thaliana*, the maternal genotype (Hollister et al. 2011; Hu et al. 2011), and therefore, fail to silence paternal TEs that are most divergent between species. To verify whether the allelic imbalance observed in TEs is a direct consequence of mixing two alien genomes, we randomly picked 18 TE genes. We measured their expression in various genotypes of *A. thaliana* and *A. lyrata* and compared them with expression in interspecific hybrids generated between various genotypes by RT-PCR (supplementary table 6, Supplementary Material online). These TE genes were generally expressed in *A. lyrata* but not in *A. thaliana*, although for many of them, expression is polymorphic in *A. lyrata* (fig. 3 and supplementary table 6, Supplementary Material online). None of the interrogated loci was expressed only in the hybrids. This suggests that for approximately 95% of the differentially expressed loci, failure in *A. lyrata* TE silencing is not specific to the hybrid. This confirms the recent report that at least 8% of TEs are expressed in *A. lyrata* although *A. thaliana* expresses only 3% (Hollister et al. 2011). We also quantified copy number variation of three randomly selected ASE-TE loci using Southern blotting but could not detect that differentially expressed TEs are reactivated in the F1 hybrids, as previously reported (Beaulieu et al. 2009).

**Cis-Regulatory Divergence and Histone Methylation**

Effective TE silencing in *A. thaliana* is contingent either on siRNA expression or on preexisting DNA methylation marks and can be reinforced by histone methylation (Mathieu et al. 2007). We asked whether differentially regulated TEs tended to display a different epigenetic profile. Based on a study of chromosome 4, most TEs in *A. thaliana* are locked in the heterochromatin, where histone H3K9me2 marks are present (Turck et al. 2007). We find that 112 of the 275 TE genes encoded on chromosome 4 preferentially expressed the *A. lyrata* allele. Of those, 89 harbor the heterochromatic mark H3K9me2, a proportion compatible with the expectation for TEs interrogated in this study. We only found a significant depletion in H3K27me3 marks among cis-regulatory divergent genes,

---

**Fig. 2.** Volcano plot contrasting the significance (−log10Q on the ordinate) and the magnitude of the difference in allelic ratio between *Arabidopsis thaliana* and *Arabidopsis lyrata* (data on log2 scale). The bottom horizontal black line corresponds to the FDR acceptance level of 0.01. Purple square loci represent TEs expressing preferentially the *A. lyrata* allele. Blue diamond loci represent TEs expressing preferentially the *A. thaliana* allele. (A), (B), and (C) show allelic expression of hybrid in leaf, flower bud, and open flower, respectively.
but this mark is rarely present on TE genes in *A. thaliana* (supplementary table 7, Supplementary Material online).

In order to confirm whether epigenetic marks H3K9me2 and H3K27me3 influence interspecific differences in TE expression, we randomly selected 11 loci expressing preferentially the *A. lyrata* allele in the hybrid and determined their histone modification status by ChIP using antibodies specific against H3K9me2 and H3K27me3 marks. Eight loci were H3K9me2 methylated in *A. thaliana* and hybrid but not in *A. lyrata* (fig. 4A). Two loci were reported to carry H3K27me3 marks in *A. thaliana* and we confirmed an absence of methylation in the *A. lyrata* parent (fig. 4B). Only one locus had no positive amplification for either of these two marks, suggesting they do not always determine allelic expression differences. Therefore, interspecific differences in histone methylation are in agreement with the expression differences observed at ASE-TEs.

**Cis-Regulatory Divergence and DNA Methylation**

The silencing of TEs also depends on DNA methylation. The MET1 gene encodes for a DNA methyl-transferase that is required to maintain cytosine methylation in CG dinucleotides during DNA replication (Fujimoto et al. 2008; Tsukahara et al. 2009). In met1 mutants of *A. thaliana*, 281 TE genes are upregulated (Lister et al. 2008). In our study, these genes were significantly enriched in ASE-TEs. As much as 64 of the 281 met1 upregulated TEs displayed ASE (hypergeometric distribution test, $P = 0.0056$). Most of them expressed preferentially the *A. lyrata* allele (62 of 64 TEs). This suggests that methylation mediated by DNA methyltransferase MET1 could differ between species.

To confirm the existence of DNA methylation differences, we selected two typical transposons for bisulfite sequencing (fig. 5). Tnp2/En/Spm (AT3G30837), which belongs to the CACTA-like transposase family, was found to have much higher DNA methylation in *A. thaliana* allele at $-423$ to $+67$ sequenced window upstream from the TSS in all three nucleotide contexts, which are known to be methylated in plants ($\chi^2$ test, CG methylation, 53.3% vs. 30.6%, $P = 1.047 \times 10^{-3}$; CHG, 50.2% vs. 17.9%, $P = 1.747 \times 10^{-4}$; CHH, 20.0% vs. 7.4%, $P = 1.099 \times 10^{-4}$; respectively; fig. 5A), especially in the $-423$ to $-200$ bp window (not shown). Accordingly, the *A. lyrata* allele was expressed preferentially in the hybrid. On the contrary,
AT4G06658, a copia-like retrotransposon, displayed a reversed pattern with the Arabidopsis thaliana allele preferentially expressed in the hybrid. The cytosines in the CHH context were more methylated in Arabidopsis lyrata within a sequenced window of $-1400$ to $-600$ bp upstream to TSS ($\chi^2$ test, 22.5% vs. 10.9%, Arabidopsis lyrata and Arabidopsis thaliana, respectively, $P = 9.9 \times 10^{-5}$; fig. 5B), especially in the $-900$ to $-600$ bp window (not shown). However, this was not significantly different for cytosine in the CG and CHG context, which were methylated more than 50%. Therefore, MET1 is able to methylate CG dinucleotides in both species, yet DNA methylation is apparently not distributed identically. Interspecific differences in methylation appear to be quantitative, local, and context dependent.

Sequence Polymorphisms and Expression Divergence of TEs

In both humans and mouse, transcription factors were shown to bind to DNA motifs found in repeats (Bourque et al. 2008), and the expression of small noncoding RNA is strongly correlated with the presence of a specific DNA motif (Kato et al. 2009). In addition, regions enriched in GC dinucleotides tend to be hypermethylated (Fujimoto et al. 2008). We thus investigated whether DNA sequence differences in the promoter can predict the expression differences we observe. GC content in the 1,000 bp upstream to TSS did not differ in ASE-TEs versus non-ASE-TEs in the Arabidopsis thaliana genome (35.3% vs. 35.1%, respectively) but was slightly higher in ASE-TEs in the Arabidopsis lyrata genome (37.1% vs. 35.9%, respectively). However, GC-rich regions were distributed differently in the Arabidopsis thaliana and Arabidopsis lyrata genomes (Two-sample Kolmogorov–Smirnov test, $P = 8.37 \times 10^{-8}$ and $5.428 \times 10^{-6}$, respectively, fig. 6A). This difference depended on the pattern of cis-regulatory divergence: The proportion of ASE-TEs with GC-rich regions in the promoter was significantly higher only for Arabidopsis lyrata TEs displaying cis-regulatory divergence (212 of 476 or 44.5%). By contrast, it was only 35.5% (200 of 562, Fisher’s exact test, $P = 0.00204$) in promoters of TEs expressing equally both alleles. The difference was particularly apparent in Arabidopsis lyrata regions located $-600$ to $-500$ bp and $-400$ to $-300$ bp upstream of the TSS (fig. 6A). No such difference could be detected in the Arabidopsis thaliana genome (Fisher’s exact test, $P = 0.2617$). As ASE-TEs are more expressed in Arabidopsis lyrata and tend to display less DNA methylation in Arabidopsis lyrata and Arabidopsis thaliana (fig. 5), it seems that the GC-rich regions might be less methylated in this species. Alternatively, CG dinucleotides present upstream of Arabidopsis lyrata ASE-TEs may have been lost by mutations in Arabidopsis thaliana (fig. 6A). Indeed, methylated cytosine residues tend to have a greater mutability than nonmethylated ones (Ossowski et al. 2010).

We further looked for specific motifs that could explain ASE. We examined 469 experimentally validated plant
cis-acting regulatory elements and scored their occurrence in the 1,000 bp window upstream to TSS of TE genes in both genomes. We found that TE genes that are differentially expressed are significantly enriched for eight motifs in the *A. lyrata* genome (FDR < 0.021, table 2) but not in the *A. thaliana* genome (FDR < 0.021, table 2). Of these motifs, two are core promoter elements, two are found in pollen-expressed genes, and three are involved in disease resistance. This may partly explain greater expression of the *A. lyrata* alleles, although we did not find that orthologous upstream regions displayed distinct contents in known cis-regulatory motifs. We further used the MEME algorithm to identify overrepresentation in unknown motifs.

### Table 2. Enrichment of DNA Motifs in the 1,000 bp Window Upstream from TSS in TE Genes with Cis-Regulatory Differences.

<table>
<thead>
<tr>
<th>Motif</th>
<th>ASE-TEs</th>
<th>Non-ASE-TEs</th>
<th>FDR</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAG</td>
<td>474</td>
<td>499</td>
<td>2.06 × 10⁻¹³</td>
<td>Binding of Dof proteins</td>
</tr>
<tr>
<td>CAAT</td>
<td>468</td>
<td>502</td>
<td>2.61 × 10⁻⁰⁸</td>
<td>CAAT BOX1</td>
</tr>
<tr>
<td>GTGA</td>
<td>458</td>
<td>479</td>
<td>2.61 × 10⁻⁰⁸</td>
<td>GTGA motif, late pollen gene g10</td>
</tr>
<tr>
<td>CTCCT</td>
<td>376</td>
<td>371</td>
<td>5.63 × 10⁻⁰⁵</td>
<td>Nodulin consensus sequences</td>
</tr>
<tr>
<td>TGAC</td>
<td>442</td>
<td>471</td>
<td>9.53 × 10⁻⁰⁵</td>
<td>W-box, WRKY binding sites</td>
</tr>
<tr>
<td>GATA</td>
<td>456</td>
<td>496</td>
<td>1.05 × 10⁻⁰⁴</td>
<td>GATA box</td>
</tr>
<tr>
<td>AGAAAA</td>
<td>431</td>
<td>458</td>
<td>3.29 × 10⁻⁰⁴</td>
<td>Pollen-specific activation</td>
</tr>
<tr>
<td>TGACT</td>
<td>278</td>
<td>275</td>
<td>2.09 × 10⁻⁰²</td>
<td>Binding to W-box element</td>
</tr>
<tr>
<td>Total TEs</td>
<td>475</td>
<td>562</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 6.** Sequence polymorphisms in ASE-TEs and non-ASE-TEs. (A) Distribution of CpG islands along the 1,000 bp upstream sequence to TSS. Red: sequence of *Arabidopsis lyrata* allele showing cis-regulatory difference (ASE-lyr); blue: sequence of *A. lyrata* showing equal parental expression (Non-ASE-lyr); orange: sequence of *Arabidopsis thaliana* showing ASE (ASE-tha); green: sequence of *A. thaliana* showing non-ASE (Non-ASE-tha). (B) SNP density in 100 bp windows in the transcribed region as well as upstream and downstream of TE genes between species. Red: TEs expressing preferentially one parental copy (ASE); blue: TEs expressing equally both parental copies (non-ASE).
poly-A tract “AAAAAAAAAAAA” and the motif “TGGATACAGACC.” In addition, the A. lyrata promoters of ASE-TEs were specifically enriched in the motifs “TTTGGCCAGTTGCC” and “CCTCCCCGMMMMMC,” whereas the promoter in A. thaliana showed a specific enrichment in motifs “TTTCTTYYTTTTY” and “TTTGGCYYGGTTGCC.” Taken together, these results suggest that differential cis-regulation of TE genes, if driven by differences in epigenetic marks, could also depend on the nucleotide sequence.

We further used SNP predicted by the genomic array hybridizations to measure nucleotide divergence between the two parental lines. ASE-TEs tended to have twice as many divergent sites in transcribed regions (3.55 SNPs per 100 bp) than non-ASE-TEs (1.75 SNPs per 100 bp; fig. 6B). This result shows that fast evolving TEs tend to be less silenced in A. lyrata.

Cis-Regulatory Divergence and TE Amplification

The sequence of the A. lyrata genome has revealed a recent amplification of several TE families, among which the TE LTR retrotransposon of the copia family, which is consistent with the higher expression of the A. lyrata copies observed in this study (Hu et al. 2011). We therefore examined whether cis-regulatory divergence in TEs associates with differences in TE mobility. The 1,535 annotated TEs in A. thaliana that were interrogated in this study had sufficient sequence homology in A. lyrata to bind allelic probes of at least one SNP probe set. Of these, 885 presented an extended sequence homology with a TE in A. lyrata (374 ASE-TEs and 511 non-ASE-TEs). Among those, ASE-TEs tended to have conserved neighbors in both species and were thus enriched in true orthologs (248 of 384 and 60 of 511, ASE-TEs and non-ASE-TEs, respectively, Fisher’s exact test, $P = 4.463 \times 10^{-66}$). This is expected since the most abundant TEs encode for retrotransposases and retroelements, which remain in the same location even if they are active. We therefore further tested whether cis-regulatory divergence was associated with copy numbers within genomes. In the A. lyrata genome, ASE-TEs had on average 6.34 copies. It was not significantly different from the average copy number in non-ASE-TEs (7.05 copies; $\chi^2$ test, $P = 0.30$). Furthermore, ASE-TEs and non-ASE-TEs had on average the same TE copy number in the A. thaliana genome (3.85 and 3.54 copies, respectively, $\chi^2$ test, $P = 0.39$). Yet, we found that divergence in cis-regulation was not independent from the age of TE insertions. Among LTR retrotransposons, ASE-TEs tended to be significantly younger than non-ASE-TEs in A. thaliana (3.009 ± 1.956 and 4.246 ± 3.944 My, ASE vs. non-ASE, respectively, Kolmogorov–Smirnov test, $P = 0.00998$). The age difference was marginal but not significant in A. lyrata (1.185 ± 1.091 and 1.485 ± 1.091 My, ASE vs. non-ASE, respectively, Kolmogorov–Smirnov test, $P = 0.0813$). This is presumably because LTR transposons are generally much younger in A. lyrata (Hollister et al. 2011; Hu et al. 2011). Since ASE-TEs overwhelmingly express the A. lyrata allele at a higher level, this means that A. thaliana is comparatively better than A. lyrata at silencing its youngest TEs. This pattern could result from the interspecific differences in the ratio of uniquely to multiply mapping siRNAs reported in Hollister et al. (2011). Altogether, this data show that the differences in cis-regulatory control of TEs does associate with TE mobility because it associates with younger A. thaliana TEs, but it is not associated with the extent of recent TE family expansions.

Concluding Remarks

Our study reveals that widespread interspecific differences in the control of TEs in the Arabidopsis genus are due to regulatory differences encoded in cis-. In A. thaliana, the transcriptional control of TEs depends on DNA and histone methylation marks (Lisch 2009; Law and Jacobsen 2010), which, by contrast with animals, are heritable both mitotically and meiotically (Kakutani et al. 1999; Vaughn et al. 2007). Widespread cis-regulatory variation of TE silencing was not observed in a previous study assessing cis-regulatory variation within A. thaliana (Zhang and Borevitz 2009). It is therefore possible that the cis-regulatory divergence in TE control observed in this study is a result of differences in the machinery depositing methylation marks. In yeast, there is accumulating evidence that variation in chromatin regulators explain a large part of interspecific gene expression divergence (Choi and Kim 2008; Field et al. 2009; Tirosh et al. 2009). Here, we observe that cis-regulatory variation also correlates strikingly with nucleotide content in upstream regions. This suggests that nucleotide divergence also participates in the differences in TE silencing, possibly through differential guidance of MET1 toward the regions onto which epigenetic marks are to be deposited.

Population genetics simulations have shown the crucial importance of mutations affecting the expression of the transposition machinery for the spread or maintenance of TE families (Le Rouzic et al. 2007). Recently, a comparative analysis of siRNA suggested differences in the epigenetic control of TE expression (Hollister et al. 2011). Our work now demonstrates that these expression differences are controlled in cis-, likely by differences of nucleotide and/or epigenetic nature. Although, we provide evidence that the expression of TE genes varies dramatically between these two species, whether this higher TE expression is the cause for differences in TE transposition in A. lyrata is not yet known. Recent work in A. thaliana has revealed that the transcriptional suppression of TEs can vary among individual copies (Lippman et al. 2003; Mirouze et al. 2009; Tsukahara et al. 2009). While confirming higher TE expression in A. lyrata, we also uncovered variation in individual TE expression in A. lyrata and A. thaliana (supplementary table 6, Supplementary Material online), suggesting that there are genetic differences acting on the silencing of specific TE gene copies. The elucidation of these differences promises to bring interesting insight into the evolution of TE control and possibly into the mechanistic basis of genome size differences among closely related species.
Supplementary Material

Supplementary tables 1–7 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We want to thank J. Gagnier and L. Steinmetz at the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) for advice at early stages of this study, B. Hüttel at the Max Planck Institute for Plant Breeding Research (MPIZ, Köln, Germany) for performing additional array hybridizations, and Marnik Vuylsteke at Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Gent, Belgium for beneficial discussion. We appreciate the support from Maarten Koomen at MPIZ. This research was funded by the Max Planck Society.

References


