Supporting Information

Hollister et al. 10.1073/pnas.1018222108

SI Materials and Methods

TE Characterization. We assembled parallel datasets of TE insertions from the *A. thaliana* and *A. lyrata* genomes, using the *A. thaliana* (TAIR 8) genome and the *A. lyrata* final 8× assembly (http://genome.jgi-psf.org/Araly1). We used RepeatModeler (Version Beta 1.0.3) to identify de novo repeats. To reduce false positives, unclassified repeats were compared with annotated genes and eliminated if they exhibited ≥80% identity to annotated genes over ≥80 bp. The remaining predictions were grouped if ≥80% identical over ≥80% of the aligned sequence of at least 80 bp in length (1). The identified repeats were appended to RepBase (Arabidopsis library: RM Database Version 20080611), resulting in a final library with 1,152 repeat units. Then the final libraries were used to annotate TEs using RepeatMasker (Version 3.2.5). If two TEs of the same family were located within 100 bp of one another in the raw annotations, they were concatenated into a single TE annotation to prevent single TE insertions from being spuriously annotated as two or more separate insertions.

To assess the quality of our TE predictions, we compared the full *A. thaliana* TE dataset with a smaller, independently derived dataset of 5,982 TE insertions (the "GR dataset") from 12 high-copy-number TE families (2). We found 98% overlap between the TEs defined in the current study and the GR dataset, which was generated by RepeatMasker with an additional BLAST-based culling procedure and substantial hand curation. Moreover, the individual TE annotations overlapped over 96% of their length, on average, between datasets.

Intact LTR retrotransposons were identified with LTR_STRUC (3), using default parameters, and LTRs were aligned using MUSCLE (4). The distance k between the two LTRs or a retrotransposon was calculated with the Kimura two-parameter model (5). The insertion time of an element was calculated as $k/2m$, where m is the rate of nucleotide substitution based on the observed mutation rate of 7×10^{-9} per site per generation (assumed to equal 1 year; ref. 6).

siRNA Data and Analysis. We used two siRNA datasets in our analyses. The *A. lyrata* data have been published (7). The *A. thaliana* (Col-0) data were generated from the same tissues (stage 1–14 floral tissues) grown under identical conditions (16 °C with a 16-h light period). Libraries were constructed as per ref. 8, except small RNA were isolated by PAGE, and RNA amplicons were reverse-transcribed using the Fermentas Revertaid kit (Fermentas Life Sciences) and then amplified by PCR using the Phusion DNA polymerase (Finnzymes). Two biological replicates were sequenced per species, by using the Illumina sequencing platform.

Sequence reads were sorted by size. The 24-nt siRNA were mapped to the *A. thaliana* and *A. lyrata* reference genomes by using the SHORE pipeline (9). We only analyzed 24-nt siRNAs with exact matches to target sequences. The density of siRNAs was measured as the number of siRNA sequences in 100-kbp bins; multiply mapping RNAs were weighted by the reciprocal of their number of total mapping locations.

Gene and TE Expression Data. We used two sources of data for gene expression. For *A. thaliana*, RNA was extracted from whole inflorescences, up to stage 14 flowers, and applied to a tiling array. Triplicate biological samples were generated, with a single technical replicate per sample. The methods used for generating and processing the tiling array data have been published (10, 11). In total, the expression levels of 24,646 genes were measured.

At the time of this study, no expression array had been developed for *A. lyrata*, so we generated a strand-specific genome-wide dataset of *A. lyrata* mRNA sequences (mRNAseq). RNA was extracted from floral tissue (stages 1–14) by using the TRIzol (Invitrogen) RNA extraction protocol. rRNAs were depleted from 10 μg of RNA by using the Ribominus kit (Invitrogen). Tobacco acid pyrophosphatase (Epicentre Biotechnologies) was used to remove the 5′ CAP, followed by metal ion fragmentation for 10 min at 70 °C (RNA fragmentation kit; Ambion). The RNA was dephosphorylated using CIP, then size selected for fragments of 100–150 nt by PAGE. RNA was extracted from the gel by semidry electrotransfer to DE81 chromatography paper (Whatman), followed by salt elution. The Illumina 3′ and 5′ sRNA adapters were sequentially ligated with size selection and rephosphorylation by PNK between ligations. After a further cycle of PAGE size selection, the RNA was reverse-transcribed with Fermentas Revertaid and PCR-amplified for 15 cycles with Phusion DNA polymerase (Finnzymes) before sequencing (Illumina sequencing-by-synthesis).

We sequenced two technical replicates each of two biological replicates, yielding 44,008,365 sequences of 42 bp (1.8 Gb total). We mapped mRNA-seq data to the *A. lyrata* genome using SHORE; 21% (9,507,594) reads mapped uniquely to the *A. lyrata* genome. To quantify expression, we weighted multiple mapping mRNA reads (which generally mapped to less than four locations) by the reciprocal of their number of mapping locations. To make explicit comparisons between the two gene expression datasets, which were based on different methods, we standardized the distribution of expression values. For the RNAseq data, the mean-per-bp coverage across genes was log normalized, and the resulting normal distribution of gene expression levels was transformed into standard (Z) scores, i.e., units of SD from the mean. The tiling array data were also transformed into standard scores for direct comparison with the RNAseq data.

To measure TE expression, we relied on our mRNAseq data in *A. lyrata*, and published mRNAseq data from *A. thaliana*. For *A. thaliana*, we downloaded 16,391,875 5-bp reads from the NCBI Short Read Archive (12). These reads were mapped to the *A. thaliana* TAIR 8 reference genome by using SHORE, in the same manner as with the *A. lyrata* RNAseq data above.

1. Wicker T, et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8:973–982.
2. 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* 19:1419–1428.
3. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19:362–367.
4. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113–132.
5. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120.
6. Ossowski S, et al. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94.
7. Fahlgren N, et al. (2010) MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* 22:1074–1089.
8. Mosher RA, et al. (2009) Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. *Nature* 460:283–286.
9. Ossowski S, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024–2033.
10. Naouar N, et al. (2009) Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes. *Plant J* 57:184–194.
11. Laubinger S, et al. (2008) At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in *Arabidopsis thaliana*. *Genome Biol* 9:R112.
12. Lister R, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133:523–536.

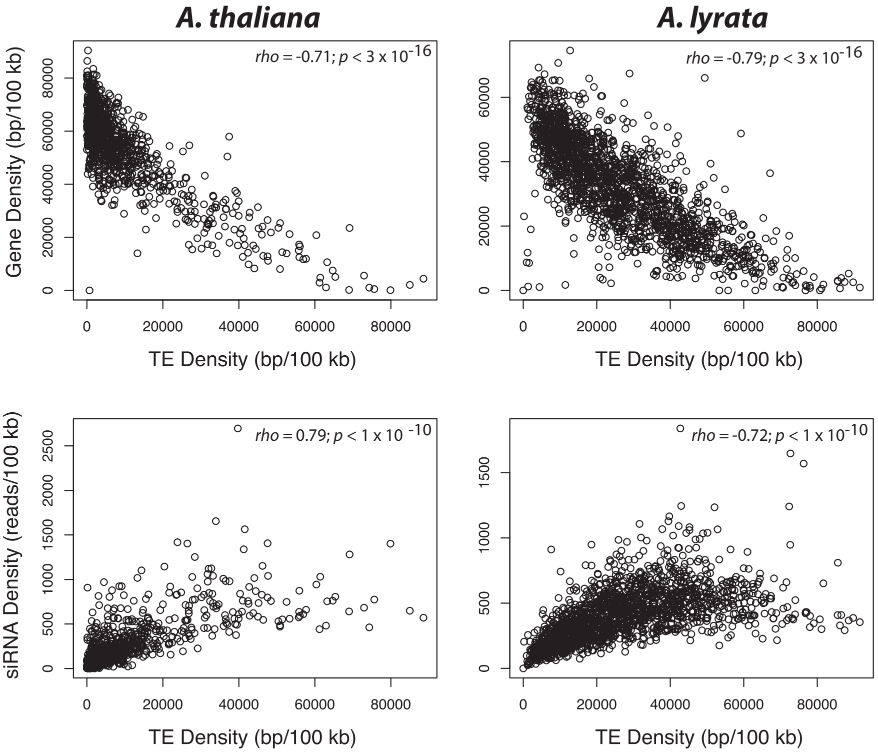

Fig. S1. Correlations between TE density and either gene density or siRNA density. Circles represent 100-kbp fixed windows.

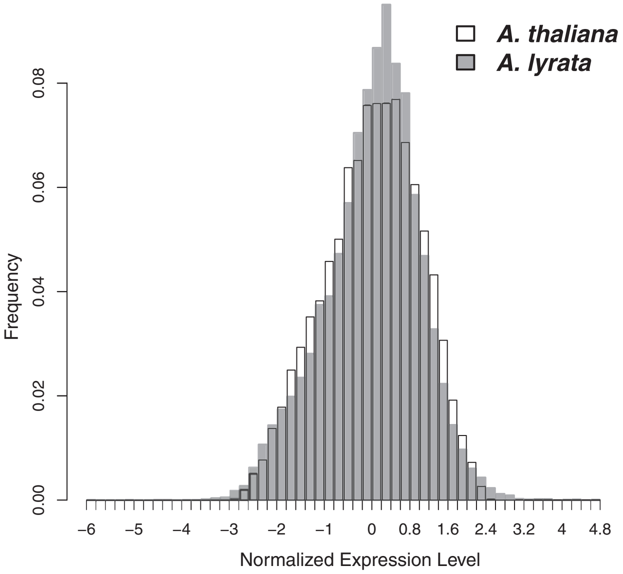

Fig. S2. Normalized gene expression data for *A. thaliana* tiling array and *A. lyrata* mRNAseq data.

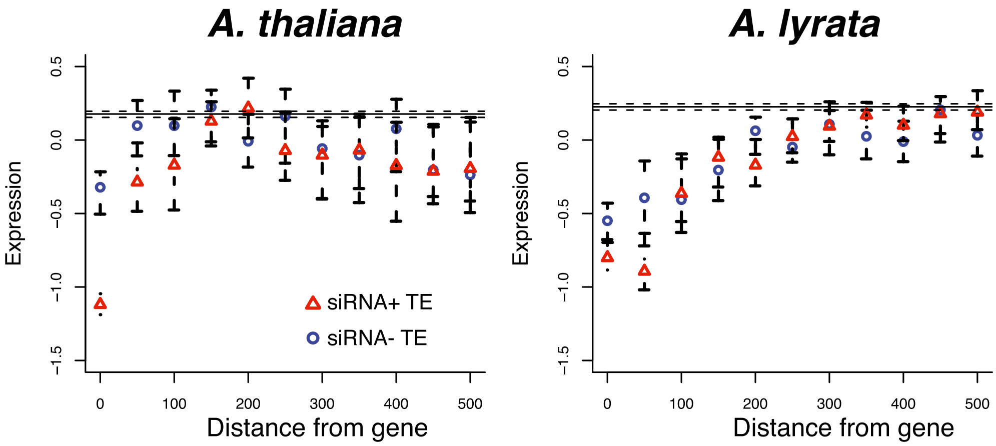

Fig. S3. Gene expression analyses performed as in Fig. 1, but using 100-bp windows instead of 500 bp. Red triangles represent genes with the closest TE siRNA+; blue circles represent genes with the closest TE siRNA−. Whiskers are 95% confidence intervals. For *A. thaliana*, although there is an apparent peak at 200 bp, genes with nearby TEs remain below the median level of gene expression for ~2.5 kbp (Fig. 1).