Transposon-Mediated Expansion and Diversification of a Family of *ULP*-like Genes

Douglas R. Hoen,[1] *Kyong Cheul Park,*[1] *Nabil Elrouby,*[2] *Zhihui Yu, Nadia Mohabir, Rebecca K. Cowan, and Thomas E. Bureau*

Department of Biology, McGill University, Montreal, Québec, Canada

Transposons comprise a major component of eukaryotic genomes, yet it remains controversial whether they are merely genetic parasites or instead significant contributors to organismal function and evolution. In plants, thousands of DNA transposons were recently shown to contain duplicated cellular gene fragments, a process termed transduplication. Although transduplication is a potentially rich source of novel coding sequences, virtually all appear to be pseudogenes in rice. Here we report the results of a genome-wide survey of transduplication in *Mutator*-like elements (MULEs) in *Arabidopsis thaliana*, which shows that the phenomenon is generally similar to rice transduplication, with one important exception: *KAONASHI* (*KI*). A family of more than 97 potentially functional genes and apparent pseudogenes, evidently derived at least 15 MYA from a cellular small ubiquitin-like modifier–specific protease gene, *KI* is predominantly located in potentially autonomous non–terminal inverted repeat MULEs and has evolved under purifying selection to maintain a conserved peptidase domain. Similar to the associated transposase gene but unlike cellular genes, *KI* is targeted by small RNAs and silenced in most tissues but has elevated expression in pollen. In an *Arabidopsis* double mutant deficient in histone and DNA methylation with elevated *KI* expression compared to wild type, at least one *KI*-MULE is mobile. The existence of *KI* demonstrates that transduplicated genes can retain protein-coding capacity and evolve novel functions. However, in this case, our evidence suggests that the function of *KI* may be selfish rather than cellular.

Introduction

Transposons are abundant constituents of all eukaryotic genomes. Although they are considered "selfish DNA" because they survive not by phenotypic selection but through self-replication, mounting evidence indicates that transposons contribute by a variety of mechanisms to the function and evolution of host genomes (Makalowski 2003; Kazazian 2004). For instance, eukaryotic transposons are able to duplicate and mobilize cellular gene sequences, potentially contributing to creative mutagenic processes like exon shuffling and gene duplication. Cellular genes flanking the 3′ termini of retrotransposons can be duplicated by readthrough transcription initiated from the element, reverse transcription of mature mRNA, and insertion into genomic DNA (Moran, DeBerardinis, and Kazazian 1999; Elrouby and Bureau 2001). By a different but unknown mechanism, DNA transposons can incorporate unspliced fragments of unlinked cellular genes between the transposon termini in a process called transduplication (Juretic et al. 2005).

Mutator elements were first discovered in maize, and *Mutator*-like elements (MULEs) constitute a diverse superfamily of DNA transposons in plants, fungi, and prokaryotes (Le et al. 2000; Yu, Wright, and Bureau 2000; Turcotte, Srinivasan, and Bureau 2001; Lisch 2002; Chalvet et al. 2003; Neuveglise et al. 2005). Autonomous MULEs contain a *mudrA* gene which encodes a transposase required for mobility. In most organisms, MULEs have long, high-identity terminal inverted repeats (TIRs), but roughly one-third of *Arabidopsis* MULEs do not. It was not previously known whether non-TIR MULEs are capable of transposition (Le et al. 2000; Yu, Wright, and Bureau 2000).

Maize *Mutator* elements were first observed to contain insertions of non-*Mutator* DNA (Chandler, Rivin, and Walbot 1986), and transduplication has subsequently been documented in *Arabidopsis* MULEs (Le et al. 2000; Yu, Wright, and Bureau 2000), CACTA elements in Japanese morning glory (Kawasaki and Nitasaka 2004) and soybean (Zabala and Vodkin 2005), as well as maize rolling circle elements (Morgante et al. 2005). Two genome-wide studies recently identified over 1,300 duplicated gene fragments formed by transduplication (i.e., transduplicates) in rice MULEs (Jiang et al. 2004; Juretic et al. 2005), but despite their abundance all existing transduplicates in rice appear to be pseudogenes (Juretic et al. 2005), and there is, to our knowledge, no documented case of a transduplicated gene encoding a functional protein. It has also yet to be determined whether transduplicates have other functions, such as the provision of sequence reservoirs for gene conversion or the generation of small RNAs (sRNAs) which might participate in RNA-mediated silencing of paralogous cellular genes.

Here we report the results of a genome-wide survey of MULE-mediated transduplication in *Arabidopsis* and the discovery of one highly unusual case. We find that the *Arabidopsis* genome contains at least 97 sequences containing strong similarity to peptidase C48, a conserved domain (CD) found exclusively in ubiquitin-like protein–specific protease (*ULP*)–like genes. Most of these sequences are located in predicted genes, but our analysis indicates that only eight are cellular *ULP*-like (*AtULP*) genes and the remainder form a unique family of transduplicates located in non-TIR MULEs which we name *KAONASHI* (*KI*; named after the mysterious character in Hayao Miyazaki's animated film "Spirited Away" who is split between conflicting worlds and identities). We contrast the characteristics of *KI* with other transduplicates; examine *KI* phylogeny, conservation, and age; compare *KI* expression patterns to those

[1] Present address: Max Planck Institute for Plant Breeding Research, Plant Developmental Biology, Köln, Germany.

[2] The first two authors contributed equally to this work.

Key words: genome evolution, gene duplication, transposable element, *Mutator*, *Arabidopsis thaliana*, SUMO.

E-mail: thomas.bureau@mcgill.ca.

Mol. Biol. Evol. 23(6):1254–1268. 2006
doi:10.1093/molbev/msk015
Advance Access publication April 3, 2006

of *mudrA* and cellular *AtULP* genes; and investigate *KI*-MULE mobility and transcriptional silencing. We conclude by arguing that *KI* is a functional and possibly selfish gene family and discuss its significance in understanding the evolutionary forces underlying the phenomenon of transduplication and transposon evolution in general.

Materials and Methods

Sequences

The Institute for Genomic Research (TIGR) *Arabidopsis thaliana* (Columbia-0) (hereafter referred to simply as *Arabidopsis*) genome sequences and annotations version 5.0 (TIGR5) were accessed both locally and online at http://www.tigr.org (AGI 2000; Haas et al. 2005). Pseudochromosome sequences were downloaded from http://www.tigr.org on June 2004.

Identification of MULEs, Transduplicates, and Peptidase C48

MULEs are characterized by long terminal sequences (greater than 100 bp) which are conserved within families, flanked by short direct repeats (9–11 bp) called target site duplications (TSDs) which are generated on insertion. MULE internal sequences are variable even between closely related elements due to deletions, insertions (including transduplications), and substitutions and also because only some MULEs contain a *mudrA* gene. We modified a previously described automated in silico procedure (Juretic et al. 2005) to identify MULEs in the TIGR5 *Arabidopsis* genome sequence, augmented with extensive manual curation, briefly described here. We compiled a library of terminal sequences (100 bp in length) from a set of previously characterized MULEs (Le et al. 2000; Yu, Wright, and Bureau 2000) and additional MULEs identified through manual curation. Using these as queries, we identified a complete set of putative MULE termini through similarity searches with the genome sequence using WU-Blast (version 2.0 release March 27, 2004; http://blast.wustl.edu) with MASKERAID (Bedell, Korf, and Gish 2000) as a search engine for REPEATMASKER (version 2004/03/06; options: nolow, nocut, no_is, s; http://www.repeatmasker.org). Pairs of termini from the same family, with correct orientation, separated by less than 30 kbp were matched, and the immediate flanking region was searched for 9-bp TSDs. Matched pairs of termini flanked by TSDs with at most two mismatched base pairs were considered to be intact MULEs.

Each MULE subsequently found to contain a transduplication or peptidase C48 sequence (see below) was verified by manual inspection of its termini, repetitiveness, and TSDs. Greater than two TSD mismatches were permitted, subject to manual inspection, if the MULE contained a peptidase C48 sequence or a transduplicate. MULEs were categorized as TIR or non-TIR according to the previously described library, in which TIR MULEs were defined as having at least 60% nucleotide identity in an alignment between the 5′ terminal 100 bp and the reverse complement of the 3′ terminal 100 bp (Yu, Wright, and Bureau 2000). We then identified additional MULEs surrounding *mudrA* sequences not contained in MULEs in this data set and iterating the above procedure. The distribution of MULEs on *Arabidopsis* chromosomes was visualized using the Nottingham *Arabidopsis* Stock Centre *Arabidopsis* Ensemble KaryoView tool (http://genome.arabidopsis.info).

To locate candidate transduplicates, we searched for cellular CDs within the MULEs by identifying all CDs and then filtering out CDs which are found in MULEs or other transposons. Because many transduplicates will probably not contain a CD, this method is expected to have a high false negative rate; however, it has the advantage of a low false positive rate (see *Discussion*). Putative transduplicated CDs were identified by querying the National Center for Biotechnology Information (NCBI) CD-Search (default settings; accessed July 2004; Marchler-Bauer and Bryant 2004) with six-frame conceptual translations of the MULE sequences. Transposon-related CDs were filtered out and ambiguous CDs, which may have been derived either from a transposon or a cellular gene, were also filtered out unless adjacent to a nonambiguous cellular CD. Excluding peptidase C48 (pfam02902) which was analyzed separately (see below), a putative cellular gene paralog for each candidate transduplicate was identified as the locus corresponding to the second best hit (best non–self-hit) in an NCBI BlastN (Altschul et al. 1990) search of the genome sequence.

A large number of MULEs were found to contain the peptidase C48 domain. To exhaustively locate all *Arabidopsis* sequences similar to peptidase C48, we employed two complementary methods: (1) CD-Search of TIGR5 annotated open reading frames (ORFs) and (2) PSI-TBlastN (version 2.2.8; Schaffer et al. 2001) search of the TIGR5 genome sequence, using as query a consensus of eight representative *Arabidopsis* peptidase C48 sequences computed by the NCBI Conserved Domain Database (gi|3377828, gi|5731755, gi|4309748, gi|3377837, gi|4678213, gi|3859612, gi|3080361, and gi|4733978). We ignored annotated exon and gene boundaries and extracted maximum length peptidase C48 sequences from genomic sequences at the positions identified in these searches and, because many sequences were apparent pseudogenes, we disregarded frameshifts and stop codons. Frameshifts were defined as two consecutive ORFs (putative exons) in different frames separated by a gap (putative intron) of less than 10 bp, a conservative threshold given that 99.9% of true *Arabidopsis* introns are longer than 50 bp (Yu et al. 2002). Frameshifts and premature stop codons in the peptidase C48 domain were counted. Cases where peptidase C48 sequences were not within a MULE were reexamined in an attempt to identify evidence of MULE-like sequences in the flanking genomic regions. Neighboring *mudrA* genes were identified based on TIGR5 annotations. The positions and distribution of peptidase C48 sequences were visualized using The *Arabidopsis* Information Resource (TAIR) Chromosome Map Tool (http://www.arabidopsis.org; Garcia-Hernandez et al. 2002). In four cases, where there was no TIGR5 locus at the position of the peptidase C48 sequence, we assigned ad hoc identifiers based on genomic positions (e.g., At4k0284 was used to identify a peptidase C48 sequence located at chromosome 4 in the region 28.4–28.5 Mbp).

Estimates of the number of peptidase C48 domains in various species were obtained from the Protein Families database (Pfam; version 19.0; Bateman et al. 2004). We also searched for *KI*-like sequences in preliminary *Brassica oleracea* genome sequence contigs (less than 0.5× coverage) using a TBlastN search on the TIGR Web site (http://www.tigr.org; accessed 10/2005) with a representative *KI* peptidase C48 domain (from At2g12100) as the query. We verified that the resulting *B. oleracea* sequences contained peptidase C48 using NCBI CD-Search.

Alignment, Phylogeny, and Conservation

3DCOFFEE (standalone T-COFFEE version 2.50; option: special-mode = 3dcoffee; O'Sullivan et al. 2004), a highly accurate multiple sequence alignment tool, was used to align all *Arabidopsis* peptidase C48 sequences with sequences from *Drosophila melanogaster* (Ulp1 gi|18860521:1318–1508), *Saccharomyces cerevisiae* (Ulp1 gi|6325237:433–617; Ulp2 gi|6322158:444–673), *Schizosaccharomyces pombe* (gi|2894265:377–564), *Homo sapiens* (Senp2 gi|54607091:397–586), and human adenovirus type 2 protein (HavULP; gi|34810217:50–144), as well as three-dimensional structure data from *S. cerevisiae* Ulp1 (Protein Database [PDB] ID 1EUV; Mossessova and Lima 2000) and *H. sapiens* Senp2 (PDB ID 1TH0; Reverter and Lima 2004). Alignments were manually edited to correct the alignment of the invariant glutamine in HavULP and KI sequences as in Mossessova and Lima (2000). We used the Phylogeny Inference Package (version 3.6; http://evolution.genetics.washington.edu/phylip.html) programs SEQBOOT, PROTDIST, NEIGHBOR, and CONSENSE to construct an extended majority rule consensus tree from 1,000 bootstrap replicates. We also clustered both protein and DNA sequences using NCBI BLASTCLUST (standalone version 2.2.10; http://www.ncbi.nlm.nih.gov/BLAST) to define clusters at various levels of divergence. Both the phylogenetic and clustering analyses supported a division between eight previously documented cellular *AtULP* genes (Novatchkova et al. 2004), which are not located in MULEs, and the remaining sequences, namely *KI* genes, which are located in MULEs.

To prepare the alignment for synonymous substitution rate analysis, we made the following adjustments. Non-*Arabidopsis* sequences were removed. Genomic DNA sequences were substituted for corresponding amino acid sequences, stop codons were replaced with gaps, codons containing gaps in more than 15% of sequences were removed, and clusters of greater than 88% amino acid identity were pruned to one sequence. Two alignments were created, one that included cellular *AtULP* genes and a second that excluded them. TREEVIEW (version 1.6.6; Page 1996) was used in making corresponding edits to the consensus tree. The adjusted alignments and trees were used to estimate *d*N/*d*S (the ratio of nonsynonymous to synonymous nucleotide substitutions per site) using the Phylogenetic Analysis by Maximum Likelihood package (PAML; version 3.14 release January 2004). BASEML was used to estimate initial branch lengths and CODEML was used for *d*N/*d*S calculations (default parameters except CodonFreq = 2, clock = 0, cleandata = 0, fix_blength = 1, and model, NSSites, and fix_omega as below). Overall *d*N/*d*S values for *KI* and *AtULP* peptidase C48 domains were calculated using the several-ratios branch model (model = 2, NSSites = 0; Yang 1998; Yang and Nielsen 1998) with one *d*N/*d*S value for *KI* and a second for *AtULP*s. *KI* clade-specific *d*N/*d*S values were calculated under the same model, using only *KI* sequences (no *AtULP*), two runs per clade (one run with fix_omega = 0, one with fix_omega = 1 and omega = 1), with one *d*N/*d*S value for the selected clade and a second for all remaining nodes. Likelihood ratio tests were used to determine whether clade-specific *d*N/*d*S values differed significantly from unity. In addition to the above calculations, *d*N/*d*S was calculated for various other sequence subsets and tree branches (e.g., excluding outliers, including nearly identical sequences, including or excluding *AtULP* genes) using various models and parameters with similar results (data not shown).

Expression

We used a combination of data sources to investigate the expression of *KI*, *mudrA*, and cellular *AtULP* genes. We identified expressed sequence tag (EST) and full-length cDNA sequences using the Munich Information Center for Protein Sequences *Arabidopsis thaliana* Database (http://mips.gsf.de/proj/thal/db; Schoof et al. 2002) and TAIR (http://www.arabidopsis.org; Garcia-Hernandez et al. 2002), and we also searched for evidence of expression in the whole-genome tiled oligonucleotide array data of Yamada et al. (2003). Using GENEVESTIGATOR (https://www.genevestigator.ethz.ch; Zimmermann et al. 2004), we collected microarray measurements (ATH1 22k array, Columbia-0) in different tissues from a pooled and standardized database incorporating several large public databases (Craigon et al. 2004; Barrett et al. 2005; Parkinson et al. 2005). Finally, 17-bp mRNA and sRNA massively parallel signature sequencing (MPSS) data were extracted from the *Arabidopsis* MPSS database (http://mpss.udel.edu; Nakano et al. 2006).

Mobility

Plant Material

Mutant seeds were obtained from Tetsuji Kakutani (Department of Integrated Genetics, National Institute of Genetics, Japan). Plants were raised in a growth chamber under 24 h daylight at 22°C. Genomic DNA was isolated from leaf tissue using the DNeasy Plant Mini Kit (QIAGEN, Mississauga, Ontario, Canada).

Transposon Display Analysis

Transposon display (TD) was performed as described by Wright et al. (2001) with minor modifications. Genomic DNA (100 ng) was digested with 2.5 U *Bfa*I (New England Biolabs, Beverly, Mass.) and ligated to 15 pmol adaptor cassettes (5′-TAGCAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGGACGAGAGAAGGGAGA-3′ and 5′-TCTTCCC-TTCTCGAATCGTAACCGTTCGTACGA-GAATCGCTGTCTCTCCTTGC-3′) with T4 DNA ligase (Invitrogen, Carlsbad, Calif.). The ligation reaction was diluted fourfold. A 3-μl aliquot of diluted ligation mixture

was used as template for preselective amplification with MULE-specific primer MuP-1 (5′-GGTCAGTTTT-TGGCT(G/T)AATGGCTAA-3′) and adaptor-specific primer ap-1 (5′-CGAATCGTAACCGTTCGTACGA-GAATCGCT-3′) and the following polymerase chain reaction (PCR) conditions: initial denaturing of 10 min at 94°C; 20 cycles of 1 min at 94°C, 1 min at 55°C, and 1 min at 72°C; and final extension of 10 min at 72°C. PCR products were diluted 100-fold in MilliQ distilled water. A 3-μl aliquot of diluted PCR products was used as template for selective amplification under the same reaction conditions with nested element-specific primer MuP-2 (5′-AAAGTGGGTCAA(C/T)GGCTACTG-3′) and IRD700-labeled (LiCor, Lincoln, Nebr.) nested adaptor-specific primer ap-2 (5′-GTACGAGAATCGCTGTCCTC-3′). Five microliters of loading dye was added to the final amplification products, which were separated by size and visualized on a 5.5% denaturing polyacrylamide gel using a LiCor IR4200 DNA sequencer. LiCor IRD700-labeled DNA (50–700 bp) was used as molecular weight markers. Polymorphic DNA fragments were isolated from the gel using the same primers (MuP-2 and ap-2) cloned into pCR 2.1 vectors (TA cloning kit, Invitrogen, Canada) and sequenced using a 3730xl DNA Analyzer (Applied Biosystems, Foster City, Calif.) at the McGill University and Genome Québec Innovation Centre (Montreal, Canada).

Reverse Transcriptase–PCR

Total RNA was isolated from floral tissue using the RNeasy Plant Mini Kit (QIAGEN, Germany) and treated with RNase-free DNase (QIAGEN) to eliminate contaminating DNA. cDNA was synthesized from 1 μg of total RNA using the SuperScript III First-Strand Synthesis System for reverse transcriptase–PCR (RT-PCR) following the manufacturer's instructions. PCR was performed using the following conditions: one cycle of 2 min at 94°C; 35 cycles of 30 s at 94°C, 30 s at 58°C, and 90 s at 68°C; and a final extension of 5 min at 68°C. PCR products were separated on a 1.5% agarose gel containing 0.4 μg/ml ethidium bromide. Forward (F) and reverse (R) primer sequences used for the amplification of *KI*-At2g12100 and *mudrA*-At2g12150 transcripts were as follows: UPF, 5′-GAAGAGAAATCGG-TATGTCGTT-3′; UPR, 5′-GACGTTGCAGGCATA-TAGCT-3′; UP1F, 5′-GCAACTGGTAGTCTTCCTGTC-3′; UP1R, 5′-ACAACTTCTTCTTCAGGATTT-3′; UP2F, 5′-GTTCAGAAAGGACTTGGTGGA-3′; UP2R, 5′-ATTAA-GGCATTCTCCCTTGGA-3′; MPF, 5′-GAACATGAG-GACGAGGATAAC-3′; MPR, 5′-CTGTTGTCACTAG-ACCGTCA-3′. The actin gene *ACT4* was used as control with primers Act4-U (5′-GCATAGAGTGAGAGAA-CAGC-3′) and Act4-D (5′-GACGTTGAAGACATTCA-ACC-3′).

Results

Detection of MULEs and Transduplicates

The MULE-mediated acquisition of cellular gene fragments in *Arabidopsis* was previously documented in a survey of approximately 15% of the genome (Yu, Wright, and Bureau 2000). To better characterize MULE-mediated transduplication in *Arabidopsis*, we modified a procedure previously shown to be accurate (Juretic et al. 2005) and identified 722 MULEs (table ST1, Supplementary Material online). Consistent with earlier reports (AGI 2000; Le et al. 2000; Yu, Wright, and Bureau 2000; Singer, Yordan, and Martienssen 2001), MULEs were found to be distributed across all chromosomes and concentrated at higher densities in pericentromeric regions and heterochromatic knobs (table ST1 and figure SF1, Supplementary Material online).

To locate transduplicate candidates, we searched for CDs within the MULEs, filtering out *mudrA*-related CDs. To differentiate between transduplication and transposon insertion, which occurs frequently, we also filtered out CDs characteristic of other transposons. Because most transduplicated sequences undergo frequent mutation, we identified CDs by scanning conceptual translations of MULE internal sequences in all six reading frames, ignoring frameshifts and stop codons. Each candidate transduplicate-containing MULE was manually verified, corresponding full-length cDNAs and ESTs were identified, and cellular genes corresponding to transduplicates were located and characterized (table 1; table ST2, Supplementary Material

Table 1
Summary of Selected *Arabidopsis* Transduplicates

MULE			Transduplicate		Cellular Gene		
Location[a]	TIR	Length[b]	Locus	CD[c]	GO Term[d]	Locus	% Identity[e]
Chr 1: 12926896–12932261	yes	4,597	At1g35240	pfam06507	Transcription factor	At1g35540	82
Chr 2: 4839811–4841640	no	553	At2g11990	pfam00319	Transcription factor	At3g04100	98
Chr 2: 6585875–6591729	no	5,854	At2g15160	smart00129	Microtubule motor	At3g49650	94
Chr 3[f]: 20780577–20799426	yes	6,180	At3g55980	pfam03157	Transcription factor	At2g40140	83
		1,628	At3g55990	COG1215	Expressed protein	At2g40150	
Chr 4: 1866831–1882462	no	11,964	At4g03930	pfam01095	Pectin-esterase	At3g27980	68
Chr 5: 14217004–14218072	yes	491	None[g]	smart00220	Kinase	At3g01085	88

[a] Chr, chromosome.
[b] Length in base pairs.
[c] Most have multiple CDs, of which only one representive is given. For more details see table ST2, Supplementary Material online.
[d] GO, Gene ontology term of cellular gene (http://www.geneontology.org).
[e] Percent nucleotide identity.
[f] This MULE contains two transduplicates.
[g] There is no locus assigned to this transduplicate.

online). We found a total of 86 CDs in 22 transduplicated sequences and 19 MULEs. However, because our methods only detect transduplicates that contain a CD, these almost certainly constitute only a subset of *Arabidopsis* transduplicates. Alignments of transduplicated sequences to putative paralogous cellular genes had nucleotide identities ranging from 58% to 98% (unweighted mean 81%). Sixteen MULEs contained a single transduplicate and three MULEs contained two transduplicates each. Nineteen of the transduplicated CDs were present in two groups of MULEs, at the following locations: chr1|16662129–16666762, chr3|11796038–11800933, chr3|14747509–14753917, chr3|20780577–20799426 and chr4|591671–592987, chr4|1866831–1882462, chr4|5200471–5204006. Eighty-one percent of transduplicated CDs were truncated by at least 30%.

Identification of *KI*

One apparently transduplicated CD, peptidase C48, had markedly unique characteristics. While most transduplicates were single copy and highly truncated, peptidase C48 was largely intact and found in a large number of MULEs (table ST3, Supplementary Material online). Peptidase C48 is a cysteine protease domain approximately 200 amino acids in length found in Ulp-like proteins, located at the C-terminus of Ulp1 and in the central region of Ulp2 (Li and Hochstrasser 1999, 2000). To fully characterize the occurrence of peptidase C48 in *Arabidopsis*, we scanned the genome sequence and located 84 peptidase C48 domains truncated by less than 5%, an additional 21 truncated by less than 20% (table ST3, Supplementary Material online), and at least an additional 35 truncated by at least 20% (table ST4, Supplementary Material online). We refer to domains truncated by 20% or less as "intact" and to the remainder as "truncated" and implicitly refer to intact domains unless otherwise noted. Eight of the intact domains were located in nonpseudogenic TIGR5 ORFs previously identified as probable *AtULP* genes (Novatchkova et al. 2004) and were not associated with MULE features such as adjacent *mudrA* ORFs or flanking MULE termini. We refer to these eight ORFs as cellular *AtULP* genes.

Sixty-nine of the remaining 97 intact domains were located in intact MULEs. Note that we use the term "intact MULEs" to mean not truncated, that is, MULEs which have matching termini and TSDs (see *Materials and Methods*). This is different from "autonomous," which would require that, in addition to being intact, a MULE contains a functional *mudrA* gene and is furthermore capable of mobilizing itself in the absence of other MULEs. Nevertheless, most of these intact MULEs also contained a *mudrA* gene, many of which (29 of 69) were in turn found to contain all three *mudrA*-related CDs (pfam03108, pfam00872, and smart00575). These MULEs are potentially autonomous (table ST3, Supplementary Material online). The remaining 28 intact peptidase C48 domains appeared to be located in truncated MULEs as they were highly similar to sequences in intact MULEs and were usually associated with one or more MULE features such as a single unpaired terminus or a *mudrA* ORF.

Ninety-three of these 97 intact peptidase C48 domains were located in TIGR5 annotated ORFs (table ST3, Supplementary Material online). These MULE-related sequences constitute a novel family of *ULP*-like genes, which we named *KI*. Seventy-seven percent of *KI*-MULEs contained a *mudrA* gene and 60% contained one or more additional ORFs, which may be transduplications or transposon insertions (table ST3, Supplementary Material online). *KI* and *mudrA* are located on opposite strands in convergent orientation. Interestingly, all *KI*-MULEs have the unusual characteristic that their termini do not form high-identity inverted repeats, that is, they are non-TIR MULEs (Yu, Wright, and Bureau 2000).

Phylogeny and Age

To investigate the phylogeny of *KI*, we performed multiple sequence alignments of the predicted amino acid sequences of all intact *Arabidopsis* peptidase C48 domains including both *KI* and cellular *AtULP* genes, representative sequences from diverse species, and three-dimensional X-ray crystallographic structures of *S. cerevisiae* Ulp1 (Mossessova and Lima 2000) and *H. sapiens* Senp2 (Reverter and Lima 2004; fig. 1; figure SF2, Supplementary Material online). We constructed a neighbor-joining tree based on protein distances (fig. 2; figure SF3, Supplementary Material online). *KI* and the eight cellular *AtULP* genes formed two highly diverged phylogenetic groups. Although the bootstrap support for each major *KI* clade (see below) is 100%, it is below 50% for the most ancient nodes due to the high degree of divergence between clades and between *KI* and *AtULP* genes. This indicates that there is not enough evidence either to determine which cellular *AtULP* is most closely related to the *KI* family or to evaluate the order in which *KI* clades diverged from one another.

Similarity-based clustering of the amino acid sequences was consistent with the phylogenetic analysis and permitted further definition of *KI* subgroups. At an arbitrary threshold of 1.0 bits/residue (approximately equivalent to 45% identity), we grouped *KI* into nine clades of 3–20 members, leaving four single-member cluster outliers. At the same threshold, cellular *AtULP* genes formed four groups (At4g00690, At4g15880, At3g06910; At1g60220, At1g10570; At4g33620, At1g09730; and At5g60190), consistent with our phylogenetic tree and with previous studies (Kurepa et al. 2003; Novatchkova et al. 2004). The non-*Arabidopsis* Ulp-like sequences each formed single-member clusters at this threshold. Further clustering at various thresholds identified additional subgroups. For instance, at 100% nucleotide identity there were 2 clusters of 2 and 4 members, at 99% nucleotide identity there were 8 clusters of 2–5 members, and at 88% amino acid identity there were 13 clusters of 2–15 members, distributed across all clades (figure SF3, Supplementary Material online).

To estimate a minimum age of *KI* formation, we searched for sequences similar to a representative *KI* peptidase C48 domain (At2g12100, Clade 9) in preliminary *B. oleracea* genomic contigs (less than 0.5× coverage) and identified 228 *KI*-like sequences ($E < 10^{-10}$). The *B. oleracea* sequences were most closely related to *Arabidopsis* Clades 7 and 8 with maximum 33% identity in a 208–amino acid BlastP alignment.

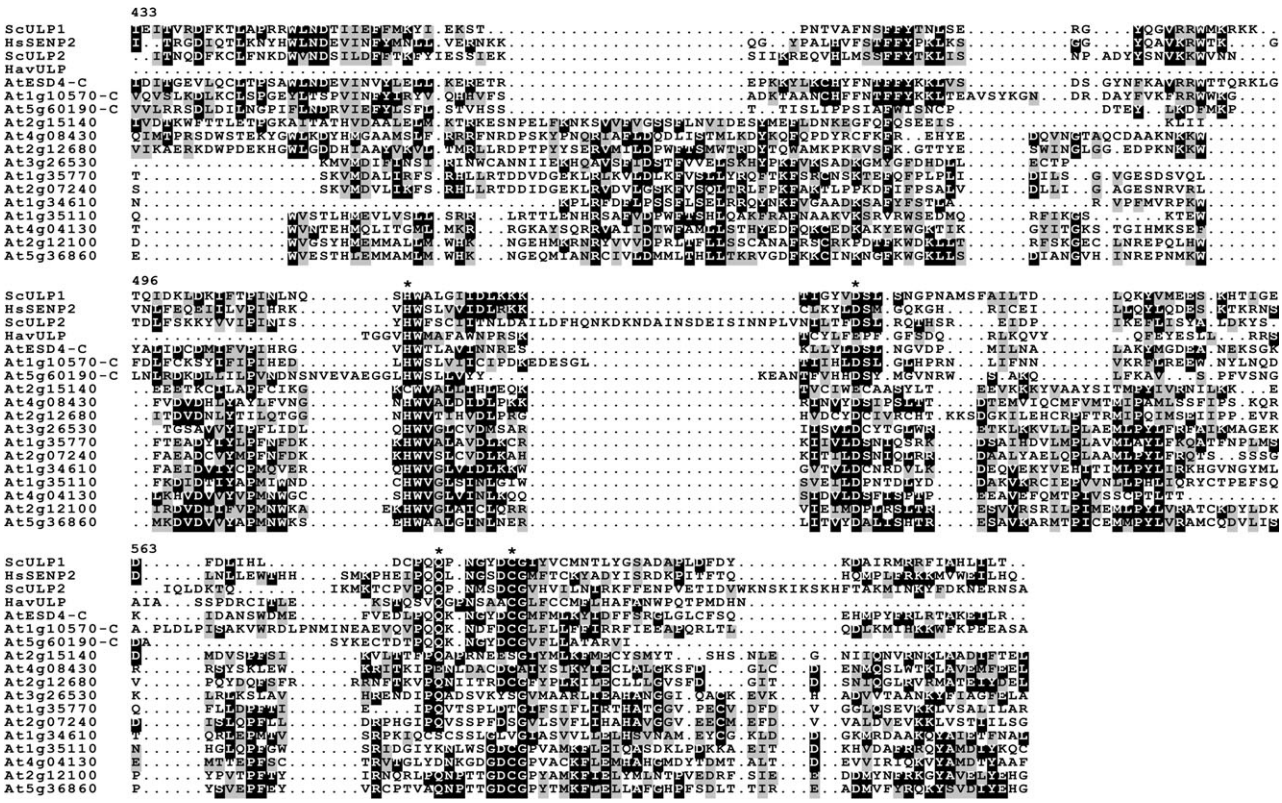

FIG. 1.—Multiple alignment of selected *Arabidopsis* peptidase C48 protein sequences with several outgroups. Probable cellular *AtULP* genes are indicated by the suffix "-C." Each KI clade is represented (Clade 1, At2g15140; Clade 3, At4g08430; Clade 2, At2g12680; Clade 5, At3g26530; Clade 4, At1g35770 and At2g07240; Clade 6, At1g34610; Clade 7, At1g35110; Clade 8, At4g04130; and Clade 9, At2g12100 and At5g36860). Outgroups are *Saccharomyces cerevisiae* ULP1 (ScULP1, gi|18860521:1318–1508) and ULP2 (ScULP2, gi|6325237:433–617), *Homo sapiens* SENP2 (HsSENP2, gi|54607091:397–586), and human adenovirus type 2 protein (HavULP, gi|34810217:50–144). Similar residues are highlighted in gray and identical residues in black. Probable catalytic residues are indicated by asterisks and are conserved only in some clades. A full alignment of KI and cellular AtULP sequences is given in figure SF3, Supplementary Material online.

Conservation

Peptidase C48 contains a putative catalytic triad of histidine, aspartate, and cysteine as well as a highly conserved glutamine residue positioned near the active site (Li and Hochstrasser 1999, 2000). We examined the *KI* peptidase C48 domains to determine whether these features were conserved and also whether the domains contained obvious disablements such as frameshifts or premature termination codons. Although many of the *KI* peptidase C48 domains were found to have one or more obvious defects, as expected for transposon-related ORFs, 53 had no obvious disablement and were positioned at the C-terminus like in Ulp1. In three of the *KI* clades (Clades 2, 3, and 9), the majority of sequences encoded all four invariant residues; in two clades (Clades 7 and 8), the majority of sequences encoded all three residues in the catalytic triad (histidine, aspartate, cysteine) but not the putative invariant glutamine; in three clades (Clades 4, 5, and 6), the majority of sequences encoded histidine, aspartate, and glutamine but not cysteine; and in the remaining clade (Clade 1), the majority of sequences encoded the invariant glutamine but none of the catalytic triad. It is possible that some of the apparently missing invariant residues are actually present but were improperly aligned; however, the regions adjacent to the invariant sites are particularly well conserved making this unlikely (fig. 1; figure SF2, Supplementary Material online).

To evaluate whether the amino acid sequences of *KI* genes had been subject to selective constraint, we estimated peptidase C48 domain dN/dS ratios for the entire *KI* subtree, for each *KI* clade, and for cellular *AtULP* genes using maximum likelihood (fig. 2; table ST5, Supplementary Material online). The overall dN/dS for *KI* and *AtULP* genes were 0.24 and 0.12, respectively. The dN/dS values of eight of the nine clades ranged from 0.11 to 0.51 and were all significantly smaller than unity (likelihood ratio test, $P < 0.001$). The only exception was Clade 2, which had dN/dS of 1.35, a value which was, however, not significantly different from unity ($P = 0.08$). In additional analyses, Clade 2 did not have an exceptionally high dN/dS ratio compared to other clades and so has probably not been subject to positive selection.

Expression

In addition to EST and full-length cDNA sequencing projects, the *Arabidopsis* transcriptome has been characterized by large-scale microarray and MPSS investigations (Brenner et al. 2000; Meyers et al. 2004; Lu et al. 2005). We identified ESTs and full-length cDNAs corresponding to *KI* and cellular *AtULP* genes in large public

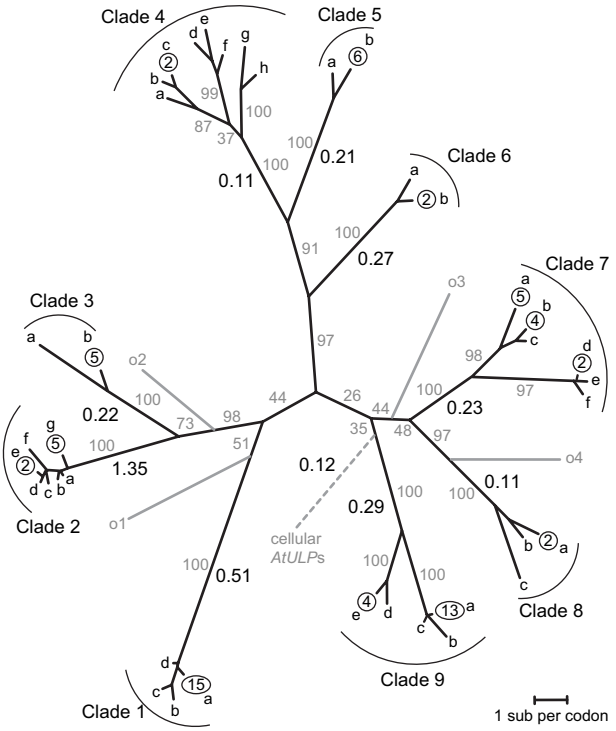

FIG. 2.—Extended majority rule phylogram of the *KI* gene family. Individual *KI* genes and 88% amino acid identity clusters are shown as labeled leaves (see figure SF2, Supplementary Material online) with circled numbers indicating the size of clusters. Bootstrap percentages are indicated in gray and *d*N/*d*S values for each clade in black. All *d*N/*d*S values are significantly smaller than unity ($P < 0.0001$), indicating that these groups have been subject to purifying selection, except Clade 2 ($P > 0.05$). Distances were estimated by PAML, and a scale of 1-nt substitution per codon is shown. Bootstrap values near the divergence between *KI* and *AtULP* are low, indicating uncertainty about the order in which *KI* clades diverged. The position of the branch containing the cellular *AtULP* genes is shown as a dashed line. A full cladogram of *KI* and *AtULP* genes is given in figure SF2, Supplementary Material online.

databases. Each of the eight cellular *AtULP* genes had between 1 and 30 corresponding ESTs, and all but two (At4g33620 and At4g00690) had at least one full-length cDNA (table ST3, Supplementary Material online). In contrast, only eight of the *KI* genes (9%) had a corresponding EST or full-length cDNA, three of which were full-length cDNAs that overlapped only a small fraction of the gene. Data from a high-density whole-genome tiled oligonucleotide array (Yamada et al. 2003) confirmed this pattern, supporting expression for only 5 of 97 *KI* genes (5%) compared to 3 of 8 cellular *AtULP* genes (38%; table ST3, Supplementary Material online).

We also compiled microarray measurements of *KI* and cellular *AtULP* gene expression levels in various *Arabidopsis* tissues using GENEVESTIGATOR, a public Web interface and standardized database of Affymetrix GeneChip data (Zimmermann et al. 2004), which consolidates several large public databases (Craigon et al. 2004; Barrett et al. 2005; Parkinson et al. 2005). Because most *KI* genes have multiple high-identity copies, many (28 of 61) probesets corresponding to *KI* genes were ambiguous; however, the ambiguity was mainly restricted to closely related *KI* genes and both ambiguous and nonambiguous probesets gave similar results (table ST6, Supplementary Material online). Whereas all eight cellular *AtULP* genes had expression levels significantly greater than background in most tissues ($P < 0.06$), maximum *KI* signal intensities were typically lower and not significantly above background in any tissue (Pina et al. 2005). The tissue-wide intensities for *KI* and *AtULP* probesets were 103 ± 17 and 597 ± 178 (normalized units; mean of inflorescence, rosette, and roots ± standard error), respectively. The majority (six of eight) of cellular *AtULP* genes had signal intensities which were 177%–1779% lower in pollen than their tissue-wide means, but two, At4g33620 and At4g15880 (*ESD4*), had signals that were, respectively, 364% and 141% higher in pollen. Conversely, most *KI* probesets had signal intensities significantly higher in pollen than their tissue-wide mean (two-tailed paired *t*-test, $P = 9.2 \times 10^{-5}$; 281 ± 64% increase). Only 23% of *KI* probesets had decreased expression in pollen. The signal intensities of ambiguous probesets appeared to increase roughly in proportion to increased ambiguity, as would be expected if multiple loci were contributing to the signal. For instance, Clade 1 was represented by five probesets with fourfold redundancy (each probeset hybridizes to the same four *KI* genes which formed a 97% nucleotide identity cluster) and six probesets with twofold redundancy which, respectively, had average pollen signal intensities of 667 and 202 and average overall signal intensities of 394 and 157. Furthermore, even if only nonambiguous *KI* probesets were considered, their signal intensities in pollen remained significantly elevated compared to their tissue-wide means (two-tailed paired *t*-test, $P = 2.1 \times 10^{-4}$). Finally, elevated expression in pollen was supported by lower but still elevated levels of expression in the stamen and inflorescence generally, as recorded by a larger number of microarrays (two for pollen, eight for inflorescence), typically with lower standard error (table ST6, Supplementary Material online).

Although microarrays remain the most widely used technology for performing large-scale analyses of expression patterns, the technology is hybridization based and so, as illustrated by *KI*, has an inherently limited ability to distinguish weakly expressed genes from the background. MPSS, which involves sequencing millions of short (17–20 bp) cDNA signatures, is both quantitative and sensitive enough to detect transcripts at concentrations as low as three to five transcripts per million (TPM; Brenner et al. 2000). MPSS was recently used to sequence a set of over 36 million 17-bp signatures (268,000 unique signatures) in 14 mRNA libraries from various *Arabidopsis* tissues, mutants, and treatments (Meyers et al. 2004). We compiled signatures from this set corresponding to *KI* and *mudrA* genes in *KI*-MULEs and to cellular *AtULP* genes (table ST7, Supplementary Material online). Because of the repetitiveness of *KI*-MULEs, many *KI* and *mudrA* signatures were nonunique; however, as with the microarray probes, ambiguities usually corresponded to *KI* or *mudrA* genes in closely related MULEs, and both unique and nonunique signatures yielded similar results. The cellular *AtULP* genes each had several unique sense signatures (2.4 ± 0.3 per gene) at high maximum abundance across all libraries (26 ± 7.4 TPM) and few or no antisense signatures (0.6 ± 0.3 per gene) at relatively low maximum

abundance (8.1 ± 1.8 TPM). Conversely, the combined number of sense signatures for all *KI* and *mudrA* genes were only seven (five unique) and six (one unique), respectively, with roughly the same number of antisense signatures, eight (one unique) and three (zero unique), respectively. The maximum abundance for sense and antisense signatures for *KI* were, respectively, 1.4 ± 0.3 TPM and 3.6 ± 0.7 TPM. Considering only sense signatures (including nonunique signatures), the maximum abundance of *KI* signatures was significantly lower than that for *AtULP* genes, consistent with the microarray results (two-tailed heteroscedastic *t*-test, $P = 3.6 \times 10^{-3}$).

To investigate whether *KI* may be targeted for RNA-mediated silencing, we examined a database of over 2 million sRNAs (over 75,000 nonredundant) derived from *Arabidopsis* inflorescence tissues and seedlings that were sequenced by MPSS (Lu et al. 2005) and identified sRNAs corresponding to *KI*, *mudrA*, and cellular *AtULP* genes (table ST8, Supplementary Material online). As with the microarray and mRNA MPSS results, many of the sRNA MPSS signatures for *KI* and *mudrA* were nonunique and so could have been generated by any of several closely related genes. However, in this case, the redundancy correlates with biological function because, in vivo, sRNAs presumably target all sequences with which they are able to hybridize. Whereas only two cellular *AtULP* genes had even a single sRNA, these having low abundance (two and three transcripts per quarter million [TPQ]), *KI* and *mudrA* genes had a large total number of sRNA signatures (24 ± 3 and 18 ± 2, respectively) at high abundance (57 ± 6 TPQ and 50 ± 6 TPQ, respectively; excluding the outlier *KI*-At4g08340 which had an abnormally high abundance of 1512 TPQ due to a nonunique signature which also matches a very highly expressed, unrelated chloroplast gene).

Mobility

We screened for insertions in wild-type *Arabidopsis* and various mutants (Columbia-0 background) with elevated transposition rates—*ddm1*, *cmt3*, *met1*, and *cmt3 met1*—using TD, a modification of the amplified fragment length polymorphism technique (Korswagen et al. 1996; Wright et al. 2001). We detected a single new insertion in a *cmt3 met1* plant, which we verified by sequencing its termini and the flanking genomic DNA (fig. 3; figure SF4, Supplementary Material online). The terminal sequence uniquely identified the element as the MULE-*KI*-At2g12100. The flanking genomic sequence contained a perfect 9-bp TSD and mapped to the short arm of chromosome 3, immediately downstream of a geranylgeranyl pyrophosphate synthase gene (figure SF5, Supplementary Material online). RT-PCR experiments showed that the *mudrA* and *KI* genes of *KI*-MULE-At2g12100 have elevated expression in *met1 cmt3* compared to wild type (fig. 4).

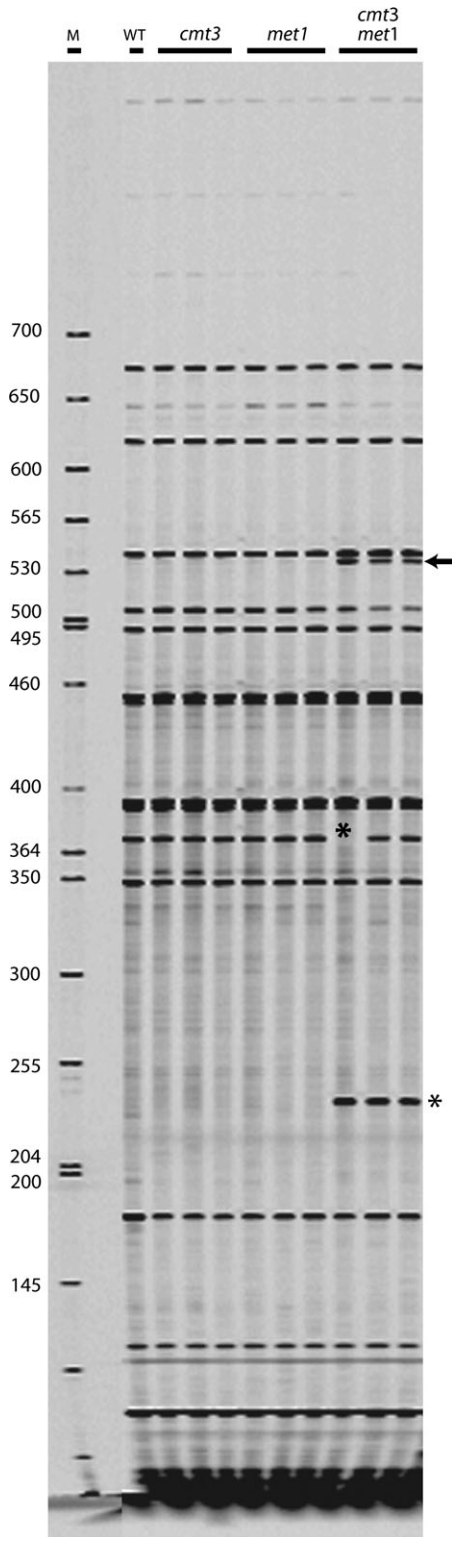

FIG. 3.—TD profile in *cmt3*, *met1*, and *cmt3 met1* mutants using MuP-1 and MuP-2 primers. Three individual plants were analyzed for each mutant line. M, molecular weight marker and WT, wild type (Columbia-0). Molecular marker sizes are indicated (base pairs). The bands at a position indicated by an arrow represent a new insertion of a MULE containing the *KI* gene At2g12100. Note that the polymorphisms marked by asterisks are not due to insertions or excisions but due to a single 153-bp deletion adjacent to a preexisting MULE.

Discussion

Transduplication

We conducted a genome-wide survey of *Arabidopsis* transduplicates, first applying an accurate procedure we previously developed to identify MULEs and then identifying

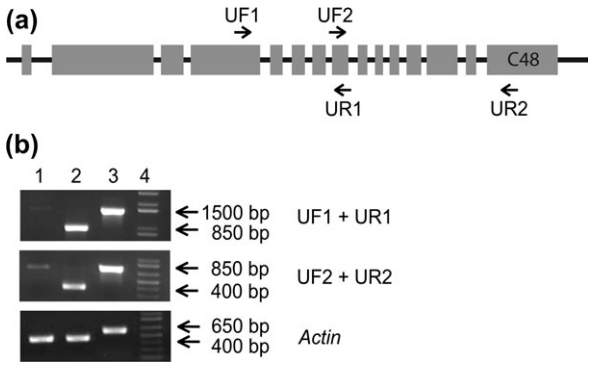

FIG. 4.—Expression of the mobile *KI* gene, At2g12100. (*a*) The structure of At2g12100. The narrow line represents genomic DNA and boxes represent exons. The exon containing the peptidase C48 domain is labeled. Arrows indicate the position of primers used for RT-PCR. (*b*) RT-PCR of At2g12100 in different backgrounds using total RNA isolated from flowers. Actin (*ACT4*) was used as a positive control. Lane 1, wild type (Columbia-0); lane 2, *cmt3 met1*; lane 3, genomic DNA; and lane 4, molecular weight marker.

cellular CDs within these elements. Because many transduplicates do not contain CDs (e.g., approximately two-thirds in rice MULEs; Juretic et al. 2005), this approach is limited in sensitivity, but it also has key advantages. (1) CDs are compact, highly conserved sequences corresponding to three-dimensional protein-folding units; therefore, sequences similar to CDs are originally derived from genuine coding sequences, not potentially misannotated genes. (2) Since most transposon-related CDs are known, CDs derived from transposon insertions can be filtered out. (3) CD truncations can be identified with high confidence and, along with frameshifts and premature stop codons, are strong indicators of gene disablement (Harrison et al. 2005). (4) Most importantly, in transduplicates that encode functional proteins maintained by selection, it is possible to detect CDs long after their nucleotide sequences and the less highly conserved regions of their amino acid sequences become too divergent from the paralogous cellular genes for any similarity to be detected.

The results of our survey indicate that the general characteristics of MULE-mediated transduplication in *Arabidopsis* (a eudicot) are similar to those previously documented in rice (a monocot), implying that its mechanism may be widely conserved in higher plants. Eighty-one percent of transduplicated CDs in *Arabidopsis* (excluding *KI*; table 1; table ST2, Supplementary Material online) and 83% of expressed transduplicated CDs in rice (Juretic et al. 2005) are truncated by more than 30%, suggesting that they may be pseudogenes (Harrison et al. 2005). Seventy-eight percent of transduplicated CDs in *Arabidopsis* (excluding *KI*; table 1; table ST2, Supplementary Material online) and 64% of transduplicates in rice (Juretic et al. 2005) are single copy, and virtually all have fewer than 10 copies in each organism, suggesting that transduplicates do not usually convey a direct selective advantage on the corresponding MULEs. Interestingly, in both *Arabidopsis* and rice, cellular genes encoding DNA-binding and transcription factors appear to have frequently been the targets of transduplication (table 1; table ST2, Supplementary Material online; Juretic et al. 2005).

KI Sequence Evolution

Despite the many similarities between *Arabidopsis* and rice transduplication, there is one key difference. *KI* is a large family of *Arabidopsis* transduplicates with unique characteristics. Unlike most transduplicates, which have low copy number and contain truncated CDs, there are 97 *KI* genes with intact peptidase C48 domains as well as at least 30 with truncated domains (tables ST3 and ST4, Supplementary Material online). The *KI* family is also unusual in its diversity. Most other identified transduplicates have high nucleotide sequence similarity to their putative cellular gene paralogs (over four-fifths have greater than 70% identity), indicating that they formed recently (table 1; table ST2, Supplementary Material online). In contrast, *KI* has split into nine highly diverged clades with as much evolutionary distance between clades as between *KI* and cellular *AtULP* genes, suggesting that extant *KI* genes may have arisen from an ancient transduplication event. This is supported by the presence of hundreds of *KI*-like sequences in *B. oleracea* (despite 0.5× coverage; data not shown), consistent with a *KI* origin prior to the divergence of *Arabidopsis* and *B. oleracea*, 15–20 MYA (Yang et al. 1999; AGI 2000). Curiously, rice, which diverged from *Arabidopsis* approximately 200 MYA (Yang et al. 1999; AGI 2000), contains at least 161 genes with peptidase C48 domains (Bateman et al. 2004); however, preliminary analysis indicates that they are predominantly associated with other DNA transposons (data not shown).

Two alternative hypotheses might explain the large number of intact peptidase C48 domains in *KI*-MULEs: either *KI*-MULEs have coincidentally undergone a large expansion recently enough for the domains to have escaped mutation or *KI* has evolved under selective constraint. While the aforementioned evidence supporting an ancient origin of the *KI* gene family provides only circumstantial evidence that the second of these possibilities is correct, nucleotide substitution patterns convincingly demonstrate that *KI* has been subject to selection. Ratios of nonsynonymous (*d*N) to synonymous (*d*S) nucleotide substitutions per site that are much smaller, somewhat smaller, or not significantly different from unity indicate, respectively, that many, some, or none of a set of putative coding sequences have been subject to purifying selection (Li, Gojobori, and Nei 1981). Similarly, values significantly larger than unity indicate positive selection. For the *KI* gene family, both overall and clade-specific *d*N/*d*S values strongly support the claim that *KI* has evolved under purifying selection. The overall *d*N/*d*S for intact *KI* peptidase C48 sequences (0.24) is within the range typically observed for functional *Arabidopsis* genes (Zhang, Vision, and Gaut 2002) and comparable to that of cellular *AtULP* peptidase C48 sequences (0.12). Also, eight of nine *KI* clades have individual *d*N/*d*S values significantly smaller than unity (fig. 2; table ST5, Supplementary Material online). Furthermore, the *KI* gene family is typical of transposon-related genes in that, in addition to putatively functional members, it contains many apparent pseudogenes which have obvious disablements such as truncations, premature stop codons, and frameshifts (table ST3, Supplementary Material online). Because the codon sequences of these pseudogenes have

presumably been drifting neutrally since becoming disabled, we would expect that the strength of selection on the functional fraction of *KI* genes has been underestimated by these *d*N/*d*S calculations. It is important to keep in mind that because *KI* is located in transposons and therefore presumably subject to different selective constraints than cellular *AtULP* genes (see below), direct comparisons between the two should be interpreted with caution. Nevertheless, the observed pattern of peptidase C48 domain sequence evolution indicates that *KI* has maintained a protein-coding function which utilizes the peptidase C48 domain.

The conclusion that *KI* encodes a peptidase is further supported by the conservation of invariant residues in widely diverged *KI* sequences. Peptidase C48 contains four highly conserved residues: a putative catalytic triad of histidine, aspartate, and cysteine and a glutamine residue predicted to help form the oxyanion hole (Li and Hochstrasser 1999, 2000). There are some reported exceptions, including human adenovirus type 2 and African swine fever virus, which contain peptidase C48 domains with a glutamate and an asparagine, respectively, at the aspartate site (fig. 1; Li and Hochstrasser 1999). The majority of sequences in three *KI* clades (Clades 2, 3, and 9) encode all four invariant residues, and the active site sequences appear to be more highly conserved than other parts of the domain, consistent with the preservation of catalytic activity in these gene products (fig. 1; figure SF2, Supplementary Material online). To varying degrees, most other KI clades also appear to have maintained some invariant residues.

Expression

The *Arabidopsis* transcriptome has been exceptionally well characterized by EST and full-length cDNA sequencing, whole-genome microarrays (Edgar, Domrachev, and Lash 2002; Brazma et al. 2003; Craigon et al. 2004), tiled oligonucleotide arrays (Yamada et al. 2003), and MPSS (Meyers et al. 2004; Lu et al. 2005). We used these resources to compile a detailed picture of the expression pattern and sRNA matches of *KI* and, for comparison, associated *mudrA* genes and cellular *AtULP*s. The repetitiveness of *KI*-MULEs presents an inherent complication in the interpretation of microarray and mRNA (but not sRNA) MPSS results because most of the microarray probesets and virtually all MPSS signatures match multiple genomic locations. However, in many cases, most or all of the ambiguous locations are within closely related *KI* (or *mudrA*) genes. Therefore, although it is impossible to determine which *KI* (or *mudrA*) gene among a set of matches contributes to signal amplitudes, we can be reasonably confident that the observed expression patterns are due to some *KI* (or *mudrA*) gene in the set. The low level of *KI* and *mudrA* expression further complicates the interpretation of the microarray (but not the MPSS) data because many individual signals are not significantly above background.

Despite these complications, our results across all data sets and most loci consistently show that *KI* expression and sRNA patterns are similar to *mudrA* and different from cellular *AtULP* genes. EST, cDNA, tiling array, microarray, and MPSS data all indicate that whereas cellular *AtULP* genes are generally expressed at significant levels, *KI* and *mudrA* genes are expressed at low or undetectable levels (tables ST3, ST6, ST7, and ST8, Supplementary Material online). Microarray data suggest that this pattern is reversed in pollen, where the expression of six of eight cellular *AtULP* genes is roughly 2- to 17-fold lower than in other tissues while the expression of *KI* genes is increased on an average of threefold (table ST6, Supplementary Material online). Interestingly, the two cellular genes with increased expression in pollen (*ESD4* and At4g33620) represent widely diverged branches of the cellular *AtULP* phylogeny (figure SF3, Supplementary Material online), which would be consistent with a separation of functional roles like that of *S. cerevisiae ULP1* and *ULP2*. These two genes may play specific specialized roles in pollen.

The MPSS results confirm these trends and provide a higher resolution picture, showing that *KI* and *mudrA* genes generate few sense transcripts and a disproportionately large number of antisense transcripts (table ST7, Supplementary Material online). The 97 *KI* genes with intact peptidase C48 domains have only 14 mRNA signatures in total, roughly the same number of which are antisense and sense. The mean abundance of the antisense transcripts is almost threefold higher than that of the sense transcripts. The opposite pattern is true of the eight cellular *AtULP* genes, which have a total of 26 mRNA signatures. Only one-quarter of these signatures are antisense, and these have mean abundance more than threefold lower than the sense transcripts. Only one *KI* gene (At1g40078) has a sense signature with abundance (16 TPM) greater than that of the least abundant primary signature of any cellular *AtULP* (At4g33620, 14 TPM).

sRNAs, which are generated by the cleavage of double-stranded RNAs including long RNA hybrids (e.g., a sense and an antisense transcript) and short RNA "hairpins," target complementary genomic DNA sequences for transcriptional silencing by heterochromatin formation, and target complementary mRNA sequences for posttranscriptional cleavage (Baulcombe 2005). The sRNA MPSS data show that, while there are virtually no sRNAs for cellular *AtULP* genes, *KI* and *mudrA* match numerous, highly abundant sRNAs (table ST8, Supplementary Material online). Only a single *KI* gene (At5g33259) has no sRNA, and each *KI* has an average of approximately 24 sRNAs with abundance 57 TPQ. The large quantity and abundance antisense transcripts and sRNAs are consistent with sRNA-mediated gene silencing of *KI* and *mudrA*.

Mobility

MULEs and other DNA transposons have low transposition rates in wild-type *Arabidopsis* and elevated rates in DNA and histone methylation mutants (Singer, Yordan, and Martienssen 2001; Lippman et al. 2003). There is to our knowledge no documented case of a non-TIR MULE (e.g., *KI*-MULEs) being mobilized. Several of our results provide circumstantial evidence of recent *KI*-MULEs transposition. (1) *KI* is predominantly located in *mudrA*-containing MULEs (i.e., potentially autonomous MULEs) which must presumably have been recently mobile in order to have maintained these *mudrA* ORFs (see also the discussion of nonphenotypic selection, below). (2) Many

KI-MULEs have perfect target site duplications, indicating that they are recent insertions. (3) Upon replicating, duplicated transposons are identical, so divergence between paralogous transposon sequences is a measure of time since duplication. The nucleotide sequences of several *KI* genes are nearly identical, for example, two clusters (six sequences in total) have 100% identity and eight clusters (25 sequences from six clades) have 99% identity (figure SF3, Supplementary Material online). While a few of these copies may have arisen through segmental duplication or other cell-mediated mechanisms (e.g., *KI*-At1k1483 and *KI*-At1g40078 form one of the 100% identity clusters but appear to have been duplicated in an inverted segmental duplication), differences in TSDs and flanking sequences show that the majority resulted from replicative transposition. This strongly suggests that a diverse group of *KI*-MULEs were recently mobile; however, it does not exclude the possibility that transposition has since been silenced.

Transposons may be epigenetically silenced by sRNA-directed heterochromatin formation, which involves DNA and histone methylation, primarily via DDM1 (decrease in DNA methylation 1)–dependent histone H3 lysine 9 methylation (H3mK9) as well as MET1 (METHYLTRANSFERASE1)-dependent cytosine methylation at CG sites and, in plants, CMT3 (CHROMOMETHYLASE3)-dependent cytosine methylation at non-CG sites (Bartee, Malagnac, and Bender 2001; Miura et al. 2001; Gendrel et al. 2002; Kato et al. 2003; Lippman et al. 2004). Like most *Arabidopsis* DNA transposons, *KI*-MULEs are concentrated in heterochromatin at the pericentromeres and at two isolated islands on chromosomes 4 and 5 (table ST1 and figure SF1, Supplementary Material online; AGI 2000; Lippman et al. 2004). Transposons in these regions, as well as those in cytologically defined euchromatin, have been shown to be subject to H3mK9 and cytosine methylation (Lippman et al. 2003; Zilberman, Cao, and Jacobsen 2003; Lippman et al. 2004). This is consistent with our results, which indicate that both *KI* and *mudrA* genes in most *KI*-MULEs are associated with antisense transcripts and sRNAs and are only weakly expressed.

To test whether *KI*-MULEs remain transpositionally competent, we screened for insertions in wild-type *Arabidopsis* and various mutants (Columbia-0 background) known to have elevated transposition rates—*ddm1*, *cmt3*, *met1*, and *cmt3 met1*—using TD (Korswagen et al. 1996; Wright et al. 2001). Although previous studies found increased transposition of a MULE and CACTA elements in *ddm1* mutants (Miura et al. 2001; Kato et al. 2003; Lippman et al. 2003), we did not detect *KI*-MULE mobility in these mutants (figure SF4, Supplementary Material online). However, we did detect a single new insertion of the MULE containing the *KI* gene At2g12100 in a *cmt3 met1* background. This confirms that at least one *KI*-MULE remains capable of mobility and provides the first experimental evidence of non-TIR MULE mobility (fig. 3). The mobility of this MULE may be related to the expression levels of its *mudrA* and *KI* genes, which are elevated in the *met1 cmt3* background compared to wild type (fig. 4).

Interestingly, At2g12100 belongs to a cluster of four closely related *KI* genes: it has 99%, 97%, and 96% nucleotide sequence identity with the peptidase C48 domains of At1g45090, At2g16180, and At2g05450, respectively. This suggests that the corresponding MULEs were recently mobile in wild type. Although we found no evidence in public databases (i.e., ESTs, full-length cDNAs, tiled oligonucleotide arrays, microarrays, and MPSS mRNAs) that At2g12100 is expressed in wild type, it has no obvious disablement, contains 99.5% of the peptidase C48 domain including all four invariant residues, and the corresponding *mudrA* gene (At2g12150) contains 100% of two *mudrA*-related CDs (table ST3, Supplementary Material Online).

Is *KI* Selfish?

Selfish genetic elements have a unique mode of survival in the genome. Cellular (i.e., nonselfish) elements are selected through their contribution to beneficial phenotypes which increase reproductive success (i.e., phenotypic selection). Selfish elements are also subject to phenotypic selection, but the phenotypes associated with selfish elements are probably deleterious or neutral in most cases because new insertions are likely to jeopardize the function of nearby sequences (Kidwell and Lisch 2001). Thus, phenotypic selection likely works to remove selfish elements rather than conserve them. But selfish elements do not require phenotypic selection to survive, and can even escape mild negative phenotypic selection, because of their ability to self-replicate. By duplicating with sufficient frequency, at least one copy of a selfish element (in an interbreeding population of host organisms) can continually escape disablement and remain able to self-replicate. This process, which has been termed "nonphenotypic" selection, inherently selects for self-replication. (Doolittle and Sapienza 1980; Orgel and Crick 1980; Hurst and Werren 2001; Brookfield 2005).

Unlike other selfish elements, the evolution of eukaryotic DNA transposons is also shaped by the constraint that, to catalyze self-replication, transposase-encoded proteins must presumably act *in trans* after being imported into the nucleus. Thus, transposons that encode functional transposases (i.e., autonomous elements) may be no more likely to be replicated by their products than related nonautonomous elements. This leads to the replication and accumulation of nonautonomous elements, which often significantly outnumber autonomous elements, and can result in decreased rates of autonomous transposition and eventually the complete silencing of DNA transposon families (Brookfield 2005).

Thus, MULEs evolve under a tension of opposing evolutionary forces. Whereas nonphenotypic selection favors elevated replication rates, new transposon copies can insert into functional cellular sequences causing negative phenotypic selection. As a result, not only do hosts maintain transposition silencing systems such as sRNA-directed transcriptional and posttranscriptional gene silencing, but transposons also evolve self-regulatory mechanisms such as tissue-specific promoters to maximize their reproductive success while minimizing deleterious effects (Kidwell and Lisch 2001). For instance, promoters in autonomous maize MULEs contain nested sets of pollen-specific motifs, and reporter gene expression is increased more than 20-fold in pollen compared to leaves (Walbot and Rudenko

2002). Preferential accumulation in heterochromatic regions with low gene density may also limit the deleterious consequences of transposition (Kidwell and Lisch 2001).

KI has the expected characteristics of a family of selfish genes. Like *mudrA* and unlike cellular *AtULP* genes, *KI* genes are targeted for silencing by numerous abundant sRNAs, are located primarily in heterochromatic regions, are not expressed at high levels, and are preferentially transcribed in pollen. Because *KI* genes have replicated to high copy number within potentially autonomous MULEs, nonphenotypic selection is sufficient to explain the observed conservation and *d*N/*d*S values. Although transposon-related genes are sometimes co-opted to perform cellular functions, these sequences subsequently lose their mobility, which is no longer required for conservation and may be a liability because mobilization could lead to deletion or inactivation of a beneficial cellular function (Cowan et al. 2005). However, *KI*-MULEs generally have intact termini, contain *mudrA* genes and so are potentially autonomous, and at least one has retained the ability to transpose. These observations may suggest that *KI* evolution has been predominantly shaped by selfish selective forces; nevertheless, the possibility remains that a contribution has been made by selection for advantageous phenotypes that *KI* may contribute to its host.

Function

All known proteins containing the peptidase C48 domain are small ubiquitin-like modifier (SUMO)–specific proteases (Li and Hochstrasser 1999). SUMO is a peptide tag that modulates the function of target proteins in diverse processes, including nucleocytoplasmic transport, signal transduction, cell-cycle progression, stress response, and transcriptional regulation (Novatchkova et al. 2004; Hay 2005). *Arabidopsis* contains nine SUMO genes, at least one of which is a pseudogene (Novatchkova et al. 2004). Among the four which show significant levels of expression, two appear to be involved in stress response signal transduction pathways (Kurepa et al. 2003; Lois, Lima, and Chua 2003). The SUMOylation of transcription factors usually correlates with transcriptional repression (Gill 2005). SUMO-specific proteases (i.e., Ulps) function as endopeptidases to activate SUMO from its inactive precursor and isopeptidases to deconjugate it from target proteins. Yeast encodes two Ulps which differ in their primary function and localization: Ulp1 is an endopeptidase and localizes to the nuclear periphery and Ulp2 is an isopeptidase and is distributed throughout the nucleus (Li and Hochstrasser 1999, 2000). Interestingly, although complete *ULP1* deletion is lethal, nonlethal mutations have been reported that result in the proliferation of yeast plasmids, a type of selfish element (Dobson et al. 2005). Furthermore, SUMO plays a role in plant defense responses, and some viruses and pathogenic bacteria encode Ulps which disrupt host defense mechanisms (Xia 2004).

Consistent with our results, previous reports have noted the relatively large number of *ULP*-like genes and pseudogenes in *Arabidopsis* (Kurepa et al. 2003; Murtas et al. 2003; Novatchkova et al. 2004). Although Kurepa et al. (2003) identified only four of the eight cellular *AtULP* genes found in this study (which they classified as *ULP1*-like, as well as eight *KI* genes, which they classified as *ULP2*-like), Novatchkova et al. (2004) suggest as candidate functional genes exactly the same set of eight cellular *AtULP* genes as we identified. Our results provide strong evidence that this set of eight genes, which form a separate phylogenetic group from *KI*, are the only cellular *Arabidopsis ULP*s (figure SF3, Supplementary Material online). They each contain an intact peptidase C48 domain, are annotated as nonpseudogenic ORFs (TIGR5 annotations), are located in gene-rich euchromatic regions, are not associated with MULE features such as nearby *mudrA* ORFs or flanking MULE termini, and are expressed at levels similar to other cellular genes.

Murtas et al. (2003) carried out the only functional analysis to date of a plant *ULP*-like gene, showing that one of the cellular *AtULP* genes (*ESD4*, *early in short days 4*, At4g15880) encodes a SUMO isopeptidase (like *ULP1*) which localizes to the nuclear periphery (like *ULP2*). If, as we suggest above, *KI* has been subject to nonphenotypic selection, then it must function to increase the replicative success of *KI*-MULEs. Although we have yet to determine the mechanism by which this may occur, we may speculate based on the known functions of SUMO and eukaryotic and pathogen-encoded Ulps that the KI protein might act as a deSUMOylase to disrupt host transposon–silencing mechanisms. For instance, KI could deSUMOylate regulators of MULE transcription, especially in pollen, enabling increased mobility. On the other hand, the substitution of invariant residues in several *KI* clades may indicate that these proteins no longer act as proteases but may, for instance, instead function as competitive inhibitors of deSUMOylation or might be nonfunctional. Another possibility is that some *KI* clades have undergone subfunctionalization and, for instance, may act on different SUMO isoforms or even alternative substrates. Finally, it is possible that, rather than directly increasing MULE mobility, KI might decrease phenotypic selection against *KI*-MULE transposition by playing a self-regulatory role, such as fine-tuning transpositional timing or frequency.

Broader Evolutionary Implications

The phenomenon of transduplication has recently generated interest largely because of its potential but unproven capacity to drive cellular protein-coding gene diversification, especially given its ability to create chimeric genes by joining duplicated sequences from distant loci (Jiang et al. 2004; Lisch 2005). However, both our current analysis of *Arabidopsis* transduplicates and our previous analysis in rice (Juretic et al. 2005) suggest that virtually all transduplicates are pseudogenes. Transduplication also has the potential to contribute to cellular function by other mechanisms, such as regulating paralogous gene expression through the generation of complementary sRNAs, or providing sequence reservoirs for gene conversion. *KI* is the first clear example to show that transduplicated genes sometimes (but rarely) do retain a protein-coding function. Yet instead of directly contributing to cellular function and evolution, we suggest that transduplication may perhaps be more important as a mechanism for transposon survival,

generating selfish gene diversity that enables transposons to overcome host repression mechanisms. Furthermore, it is now well documented that transposon-related genes can adopt cellular functions (Brookfield 2005), suggesting that a cycle of transduplication followed by "domestication" is a possible albeit circuitous route to the expansion of the cellular protein-coding gene repertoire.

Supplementary Material

Supplementary figures SF1–SF5 and tables ST1–ST8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Acknowledgments

We would like to thank Tetsuji Kakutani for supplying the mutant lines. This study was supported by grants to T.E.B. from the National Science and Engineering Research Council of Canada and the McGill University William Dawson Scholar Chair. K.C.P. was supported by the Korea Research Foundation grant (Ministry of Education & Human Resources Development, KRF-2003-214-C00229; Korea). Sequencing of *B. oleracea* was funded by the National Science Foundation, and preliminary sequence data were obtained from TIGR (http://www.tigr.org).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

[AGI] Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408**:796–815.

Barrett, T., T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. 2005. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res. **33**: D562–D566.

Bartee, L., F. Malagnac, and J. Bender. 2001. Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. Genes Dev. **15**:1753–1758.

Bateman, A., L. Coin, R. Durbin et al. (13 co-authors). 2004. The Pfam protein families database. Nucleic Acids Res. **32**: D138–D141.

Baulcombe, D. 2005. RNA silencing. Trends Biochem. Sci. **30**:290–293.

Bedell, J. A., I. Korf, and W. Gish. 2000. MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics **16**:1040–1041.

Brazma, A., H. Parkinson, U. Sarkans et al. (13 co-authors). 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. **31**:68–71.

Brenner, S., M. Johnson, J. Bridgham et al. (24 co-authors). 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. **18**:630–634.

Brookfield, J. F. 2005. The ecology of the genome—mobile DNA elements and their hosts. Nat. Rev. Genet. **6**:128–136.

Chalvet, F., C. Grimaldi, F. Kaper, T. Langin, and M. J. Daboussi. 2003. Hop, an active Mutator-like element in the genome of the fungus Fusarium oxysporum. Mol. Biol. Evol. **20**: 1362–1375.

Chandler, V., C. Rivin, and V. Walbot. 1986. Stable non-mutator stocks of maize have sequences homologous to the Mu1 transposable element. Genetics **114**:1007–1021.

Cowan, R. K., D. R. Hoen, D. J. Schoen, and T. E. Bureau. 2005. MUSTANG is a novel family of domesticated transposase genes found in diverse angiosperms. Mol. Biol. Evol. **22**:2084–2089.

Craigon, D. J., N. James, J. Okyere, J. Higgins, J. Jotham, and S. May. 2004. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res. **32**:D575–D577.

Dobson, M. J., A. J. Pickett, S. Velmurugan, J. B. Pinder, L. A. Barrett, M. Jayaram, and J. S. Chew. 2005. The 2 microm plasmid causes cell death in Saccharomyces cerevisiae with a mutation in Ulp1 protease. Mol. Cell. Biol. **25**:4299–4310.

Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601–603.

Edgar, R., M. Domrachev, and A. E. Lash. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**:207–210.

Elrouby, N., and T. E. Bureau. 2001. A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. J. Biol. Chem. **276**: 41963–41968.

Garcia-Hernandez, M., T. Z. Berardini, G. Chen et al. (21 co-authors). 2002. TAIR: a resource for integrated Arabidopsis data. Funct. Integr. Genomics **2**:239–253.

Gendrel, A. V., Z. Lippman, C. Yordan, V. Colot, and R. A. Martienssen. 2002. Dependence of heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1. Science **297**:1871–1873.

Gill, G. 2005. Something about SUMO inhibits transcription. Curr. Opin. Genet. Dev. **15**:536–541.

Haas, B. J., J. R. Wortman, C. M. Ronning et al. (12 co-authors). 2005. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biol. **3**:7.

Harrison, P. M., D. Zheng, Z. Zhang, N. Carriero, and M. Gerstein. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. **33**:2374–2383.

Hay, R. T. 2005. SUMO: a history of modification. Mol. Cell. **18**:1–12.

Hurst, G. D., and J. H. Werren. 2001. The role of selfish genetic elements in eukaryotic evolution. Nat. Rev. Genet. **2**:597–606.

Jiang, N., Z. Bao, X. Zhang, S. R. Eddy, and S. R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature **431**:569–573.

Juretic, N., D. R. Hoen, M. L. Huynh, P. M. Harrison, and T. E. Bureau. 2005. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res. **15**: 1292–1297.

Kato, M., A. Miura, J. Bender, S. E. Jacobsen, and T. Kakutani. 2003. Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. Curr. Biol. **13**:421–426.

Kawasaki, S., and E. Nitasaka. 2004. Characterization of Tpn1 family in the Japanese morning glory: En/Spm-related transposable elements capturing host genes. Plant Cell Physiol. **45**:933–944.

Kazazian, H. H. Jr. 2004. Mobile elements: drivers of genome evolution. Science **303**:1626–1632.

Kidwell, M. G., and D. R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution **55**:1–24.

Korswagen, H. C., R. M. Durbin, M. T. Smits, and R. H. Plasterk. 1996. Transposon Tc1-derived, sequence-tagged sites in Caenorhabditis elegans as markers for gene mapping. Proc. Natl. Acad. Sci. USA **93**:14680–14685.

Kurepa, J., J. M. Walker, J. Smalle, M. M. Gosink, S. J. Davis, T. L. Durham, D. Y. Sung, and R. D. Vierstra. 2003. The small ubiquitin-like modifier (SUMO) protein modification system in Arabidopsis. Accumulation of SUMO1 and -2 conjugates is increased by stress. J. Biol. Chem. **278**: 6862–6872.

Le, Q. H., S. Wright, Z. Yu, and T. Bureau. 2000. Transposon diversity in Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA **97**:7376–7381.

Li, S. J., and M. Hochstrasser. 1999. A new protease required for cell-cycle progression in yeast. Nature **398**: 246–251.

———. 2000. The yeast ULP2 (SMT4) gene encodes a novel protease specific for the ubiquitin-like Smt3 protein. Mol. Cell. Biol. **20**:2367–2377.

Li, W. H., T. Gojobori, and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237–239.

Lippman, Z., A. V. Gendrel, M. Black et al. (14 co-authors). 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature **430**:471–476.

Lippman, Z., B. May, C. Yordan, T. Singer, and R. Martienssen. 2003. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. PLoS Biol. **1**:E67.

Lisch, D. 2002. Mutator transposons. Trends Plant Sci. **7**: 498–504.

———. 2005. Pack-MULEs: theft on a massive scale. Bioessays **27**:353–355.

Lois, L. M., C. D. Lima, and N. H. Chua. 2003. Small ubiquitin-like modifier modulates abscisic acid signaling in Arabidopsis. Plant Cell **15**:1347–1359.

Lu, C., S. S. Tej, S. Luo, C. D. Haudenschild, B. C. Meyers, and P. J. Green. 2005. Elucidation of the small RNA component of the transcriptome. Science **309**:1567–1569.

Makalowski, W. 2003. Genomics. Not junk after all. Science **300**:1246–1247.

Marchler-Bauer, A., and S. H. Bryant. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. **32**: W327–W331.

Meyers, B. C., S. S. Tej, T. H. Vu, C. D. Haudenschild, V. Agrawal, S. B. Edberg, H. Ghazal, and S. Decola. 2004. The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res. **14**:1641–1653.

Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature **411**:212–214.

Moran, J. V., R. J. DeBerardinis, and H. H. Kazazian Jr. 1999. Exon shuffling by L1 retrotransposition. Science **283**: 1530–1534.

Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo, and A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. **37**:997–1002.

Mossessova, E., and C. D. Lima. 2000. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. Mol. Cell **5**:865–876.

Murtas, G., P. H. Reeves, Y. F. Fu, I. Bancroft, C. Dean, and G. Coupland. 2003. A nuclear protease required for flowering-time regulation in Arabidopsis reduces the abundance of small ubiquitin-related modifier conjugates. Plant Cell **15**: 2308–2319.

Nakano, M., K. Nobuta, K. Vemaraju, S. S. Tej, J. W. Skogen, and B. C. Meyers. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res. **34**:D731–D735.

Neuveglise, C., F. Chalvet, P. Wincker, C. Gaillardin, and S. Casaregola. 2005. Mutator-like element in the yeast Yarrowia lipolytica displays multiple alternative splicings. Eukaryot. Cell **4**:615–624.

Novatchkova, M., R. Budhiraja, G. Coupland, F. Eisenhaber, and A. Bachmair. 2004. SUMO conjugation in plants. Planta **220**:1–8.

Orgel, L. E., and F. H. Crick. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

O'Sullivan, O., K. Suhre, C. Abergel, D. G. Higgins, and C. Notredame. 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol. **340**:385–395.

Page, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**:357–358.

Parkinson, H., U. Sarkans, M. Shojatalab et al. (18 co-authors). 2005. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. **33**: D553–D555.

Pina, C., F. Pinto, J. A. Feijo, and J. D. Becker. 2005. Gene family analysis of the Arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. Plant Physiol. **138**: 744–756.

Reverter, D., and C. D. Lima. 2004. A basis for SUMO protease specificity provided by analysis of human Senp2 and a Senp2-SUMO complex. Structure **12**:1519–1531.

Schaffer, A. A., L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. **29**:2994–3005.

Schoof, H., P. Zaccaria, H. Gundlach, K. Lemcke, S. Rudd, G. Kolesov, R. Arnold, H. W. Mewes, and K. F. Mayer. 2002. MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res. **30**: 91–93.

Singer, T., C. Yordan, and R. A. Martienssen. 2001. Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene decrease in DNA methylation (DDM1). Genes Dev. **15**:591–602.

Turcotte, K., S. Srinivasan, and T. Bureau. 2001. Survey of transposable elements from rice genomic sequences. Plant J. **25**:169–179.

Walbot, V., and G. N. Rudenko. 2002. *MuDR/Mu* transposable elements of maize. Pp. 533–564 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. ASM Press, Washington, D.C.

Wright, S. I., Q. H. Le, D. J. Schoen, and T. E. Bureau. 2001. Population dynamics of an Ac-like transposable element in self- and cross-pollinating Arabidopsis. Genetics **158**: 1279–1288.

Xia, Y. 2004. Proteases in pathogenesis and plant defence. Cell Microbiol. **6**:905–913.

Yamada, K., J. Lim, J. M. Dale et al. (70 co-authors). 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302**:842–846.

Yang, Y. W., K. N. Lai, P. Y. Tai, and W. H. Li. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J. Mol. Evol. **48**:597–604.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

Yang, Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**:409–418.

Yu, J., Z. Yang, M. Kibukawa, M. Paddock, D. A. Passey, and G. K. Wong. 2002. Minimal introns are not "junk". Genome Res. **12**:1185–1189.

Yu, Z., S. I. Wright, and T. E. Bureau. 2000. Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. Genetics **156**:2019–2031.

Zabala, G., and L. O. Vodkin. 2005. The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell **17**:2619–2632.

Zhang, L., T. J. Vision, and B. S. Gaut. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in Arabidopsis thaliana. Mol. Biol. Evol. **19**: 1464–1473.

Zilberman, D., X. Cao, and S. E. Jacobsen. 2003. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science **299**:716–719.

Zimmermann, P., M. Hirsch-Hoffmann, L. Hennig, and W. Gruissem. 2004. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol. **136**: 2621–2632.

Dan Graur, Associate Editor

Accepted March 27, 2006