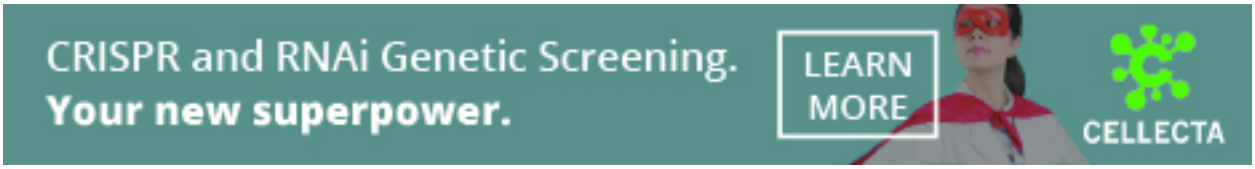
Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome

Regina S. Baucom, James C. Estill, Jim Leebens-Mack, et al.

Genome Res. 2009 19: 243-254 originally published online November 24, 2008
Access the most recent version at doi:10.1101/gr.083360.108



Letter

Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome

Regina S. Baucom,[1,3] James C. Estill,[2] Jim Leebens-Mack,[2] and Jeffrey L. Bennetzen[1]

[1]*Department of Genetics, University of Georgia, Athens, Georgia 30602-7223, USA;* [2]*Plant Biology Department, University of Georgia, Athens, Georgia 30602-7223, USA*

Although the proliferation of LTR retrotransposons can cause major genomic modification and reorganization, the evolutionary dynamics that affect their frequency in host genomes are poorly understood. We analyzed patterns of genetic variation among LTR retrotransposons from *Oryza sativa* to investigate the type of selective forces that potentially limit their amplification and subsequent population of a nuclear genome. We performed both intra- and interfamily analyses of patterns of molecular sequence variation across multiple LTR retrotransposon genes. This analysis involved more than 1000 LTR retrotransposon sequences from 14 separate families that varied in both their insertion dates and full-length copy numbers. We uncovered evidence of strong purifying selection across all gene regions, but also indications that rare episodes of positive selection and adaptation to the host genome occur. Furthermore, our results indicate that LTR retrotransposons exhibit different but predictable patterns of sequence variation depending on their date of transposition, suggesting that LTR retrotransposons, regardless of superfamily and family classifications, show similar "life-histories."

Orgel and Crick (1980) and Doolittle and Sapienza (1980) promoted the concept that the repetitive DNA found within organismal genomes was "parasitic," and provided no adaptive benefit to the host genome. Under this hypothesis, the ability to transpose alone should ensure transposable element (TE) survival, rendering an adaptive explanation, in terms of their benefit to the host, unnecessary to explain their presence (Doolittle and Sapienza 1980; Orgel and Crick 1980). That TEs are purely parasitic on host genomes has long been a topic of debate (Kidwell and Lisch 2000, 2001), especially in light of reports that show TE sequences have been co-opted for host processes (Hudson et al. 2003; Cordaux et al. 2006; Lin et al. 2007; Mason et al. 2008). However, the finding that hosts experience a fitness detriment in response to increased TE densities (Pasyukova et al. 2004) and activity (Mackay et al. 1992; Charlesworth et al. 1994) substantiates the view that TEs are arbiters of negative selection on their hosts, which in turn suggests that the type of selection on TEs should also be negative.

The authors of the parasitic DNA hypothesis suggested that a "nonphenotypic" selection regime should act on TEs, meaning that various forces within the host promote "genome-level" selection (Doolittle and Sapienza 1980; Orgel and Crick 1980). They further suggested that the pattern of sequence variation seen among TEs can be explained by selection, but they did not predict whether purifying, positive, or balancing selection should predominate. There is reason to believe that the type and intensity of selection could be variable among the extant, full-length population of transposable elements within host genomes. First, the different superfamilies of LTR-retrotransposons, *copia* and *gypsy*, are often found in different densities or copy numbers (Gao et al. 2004), as are the evolutionarily distinct subfamilies (Bennetzen 2000; Vitte and Panaud 2005). This variation in the copy number of different LTR retrotransposons suggests either a differential transposition or removal rate among families, or both. Second, transposable elements are often located away from genic regions or areas of high recombination (Bartolomé et al. 2002), suggesting that the elements with the most successful life-history strategies are the ones that do not disrupt biological processes within host genomes by inducing chromosomal rearrangements via ectopic exchange (Charlesworth and Langley 1989) cause deleterious insertions (Finnegan 1992) or induce host fitness costs due to transposition activity (Nuzhdin 1999). Finally, in the case of LTR retrotransposons, there is a high degree of conservation of the reverse transcriptase genic region among elements from widely separated taxonomic hosts (Xiong and Eickbush 1990), suggesting that it is likely to exhibit signs of historical selective regimes.

There are a number of other proteins, in addition to reverse transcriptase, that are required for the replication of LTR retrotransposons. Any of these could potentially show patterns of variation that are indicative of selection. *gag* encodes proteins involved in the maturation and packaging of retrotransposon RNA into a DNA copy that is later integrated back into the host genome. The *pol* gene region encodes both reverse transcriptase (*RT*) and RNase H (*RH*), which mediate the replication process, whereas the protein integrase (*INT*) inserts the new DNA copy of the retrotransposon into nuclear DNA (Kumar and Bennetzen 1999). Because a newly integrated LTR retrotransposon is a copy of its parent molecule, LTR retrotransposons increase in number following the replication cycle, a strategy that is different from the type II DNA elements. Thus, the "copy and paste" mode of replication of LTR retrotransposons has led to significant effects on the host genome, from increasing genome size to causing lethal mutations (SanMiguel et al. 1996; Kumar and Bennetzen 1999; Devos et al. 2002; Ma and Bennetzen 2004; Hawkins et al. 2006; Piegu et al. 2006; Ammiraju et al. 2007).

While it is clear that transposition can be deleterious to the host, what is less understood is whether host genomes impose selection on LTR retrotransposons to reduce such proliferation, and if so, what mechanisms would be involved. Given the long-term evolutionary relationship between a host and its TEs, they are

[3]Corresponding author.
E-mail gbaucom@uga.edu; fax (706) 542-3910.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.083360.108.

often compared with organismal parasites, and coevolutionary arms race models have been hypothesized to underlie the process of their transposition and subsequent removal or silencing from organismal genomes (Orgel and Crick 1980; McDonald 1998). Under this model, TEs should adapt to host defenses, and selective sweeps are expected as elements with adaptive mutations populate the genome—this model predicts low gene diversity and evidence of positive selection. However, an explicit investigation of TE sequence variation in this context has yet to be considered.

The availability of sequenced genomes, such as that of *Oryza sativa*, or rice, allows for such empirical investigations. While the transpositional dynamics and rate of removal of the LTR retrotransposons from the *O. sativa* genome have been studied (Vitte and Panaud 2003; Ma and Bennetzen 2004; Ma et al. 2004; Vitte et al. 2007), the potential for natural selection in shaping the sequence variation of extant, full-length elements remains unknown. The study herein reports an investigation of the prevalence of historical selection as well as patterns of molecular sequence variation on LTR retrotransposon families using both a population genetics approach and a comparative analysis. The data show evidence of intense purifying selection across all families, but also demonstrate that rare episodes of positive selection and adaptation may occur. Furthermore, identified differences in sequence variation among recently transposed versus older families suggests that there is a progressive post-insertion life cycle common to all LTR retrotransposons.

Results and Discussion

Strategy for mining, identifying, and dating LTR retrotransposons in rice

We downloaded the sequences of the 12 chromosomes of the *O. sativa* genome (variety Nipponbare) from GenBank and mined for LTR retrotransposons using the program LTR_STRUC (McCarthy and McDonald 2003). LTR retrotransposons are characterized by the presence of long terminal repeats (LTRs) that flank an internal coding domain (Kumar and Bennetzen 1999). LTR_STRUC is a structural data-mining program that identifies the LTR retrotransposons based on the presence of these LTRs; the algorithm of this program requires that the two LTRs are at a distance of at least 40 bp from one another and are more than 70% percent homologous. Thus, it does not rely on sequence similarity to a known database of elements, and as such, can uncover previously uncharacterized LTR retrotransposon families. Using this program, 2226 full-length LTR retrotransposons were mined from the *O. sativa* genome and they were then placed into families using two separate strategies. First, we ran an "all-by-all" BLASTN search with the 5′ LTR of each of the mined elements at an expected value of e^{-10}. We rendered the results of this all-by-all blast into a visual image using the Perl program RepMiner (J. Estill, code available at http://repminer.sourceforge.net), which formats the output of BLAST searches for the imaging program Cytoscape (Shannon et al. 2003; v. 2.5.1). This strategy placed the 2226 LTR retrotransposons into 172 families. We uncovered extensive variation for copy number and primary DNA sequence (Fig. 1) and observed more low- than high-copy number families. Forty-eight percent of the families were represented by a single sequence, 20% of the families were made up of two sequences, and 32% of the families were represented by three or more sequences within the genome. We excluded the single-copy element families (n = 81) from further analyses, given that two or more sequences are needed for selection analyses. Second, we performed a BLASTN search (e^{-10}) of the mined LTR retrotransposons using the RetrOryza database (Chaparro et al. 2007) in order to verify our family designations (Fig. 1). The family designations of the two methods were in general agreement, although there were a few cases in which the first methodology "lumped" families that the second methodology would have determined were separate, and vice versa. For the purposes of the presented work, we used families only if they were clearly defined using both methods in concert. Current work in this lab is contrasting the RepMiner-based nomenclature of LTR retrotransposons with other methods (see Wicker et al. 2007), and is not the focus of the present work.

After mined elements were reliably assigned to families, we determined both the insertion time and gene content of each element. We measured the degree of divergence between the two LTR regions of the LTR retrotransposons to get an estimate of insertion time (SanMiguel et al. 1998; Ma and Bennetzen 2004; Ma et al. 2004; Vitte et al. 2004; Vitte and Bennetzen 2006; Wicker and

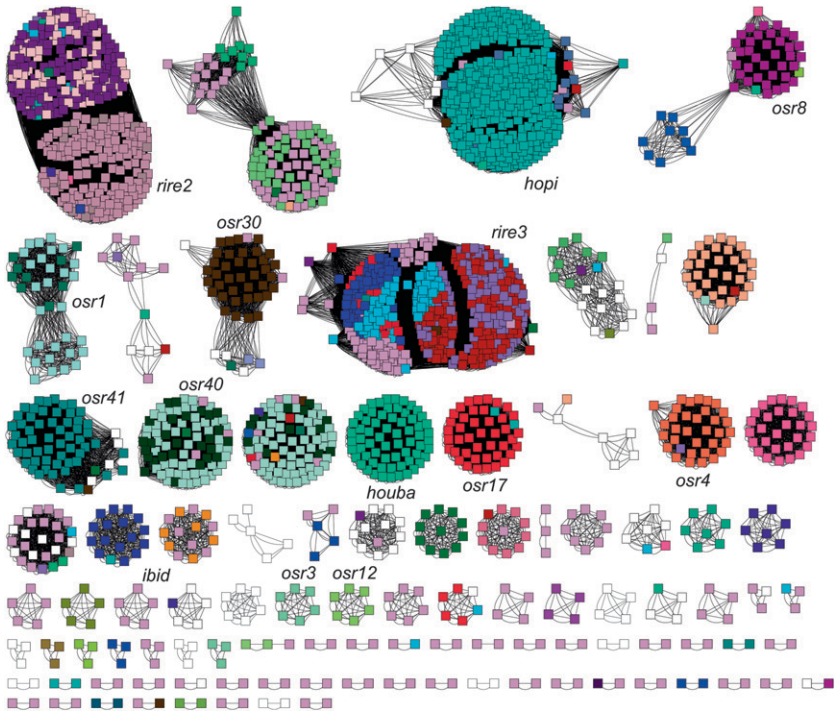


Figure 1. The LTR retrotransposon families of rice. An all-by-all BLASTN (e^{-10}) was performed using the 5′ LTR sequence of each element from our database of 2226 LTR retrotransposons. The program RepMiner (J. Estill) was then used to format the results of this BLAST for graphic display by the program CytoScape (v. 2.5.1). The 5′ LTR of each element is illustrated as a single square node, while blast hits are shown as lines connecting these nodes. Each individual cluster of connected nodes may be considered a family according to a connected components interpretation of family affinity. The colors of individual nodes represent the family name assignment from a BLASTN (e^{-10}) to the RetrOryza database (Chaparro et al. 2007). Families used in the selection analyses are labeled.

Keller 2007). The LTRs are identical at the time of insertion into the host genome, and thus the level of divergence between the two provides an estimate of the time that has elapsed since insertion (SanMiguel et al. 1998). An analysis of variance finds that these insertion dates varied significantly among families ($F_{(86,1826)} = 13.83$; $P < 0.0001$), and ranged from 0 yr to 6.59 million yr ago (Mya) among families represented by two or more sequences (Fig. 2). This analysis suggests that the majority of the families exhibit significant variation among them for insertion date, meaning that an increased rate of transposition at different times is common among families (Table 1). This observation is in line with previous reports that find significant peaks in a distribution of insertion times among LTR retrotransposon families within *O. sativa* (Vitte et al. 2007; Wicker and Keller 2007).

To find the LTR retrotransposon gene regions, we performed BLASTX searches (e^{-1}) using a database of the PFAM proteins *gag*, RNase H, integrase, and reverse transcriptase from previously characterized *O. sativa* LTR retrotransposons. We also queried our mined elements using reverse transcriptase and integrase gene models downloaded from NCBI in TBLASTX searches (e^{-1}) in order to identify genes within the *copia* families. The use of two different search strategies limits the ability to directly compare results between *copia* and *gypsy* families, yet this was necessary, as a limited number of genes were found within *copia* families using the PFAM gene models. Furthermore, we found very few gene models of *gag* and RNase H for the *copia* families, and as such, elected to consider only integrase and reverse transcriptase for this superfamily. Among the *gypsy* families in the data set represented by two or more sequences, 50% and 30% of the LTR retrotransposons exhibited homology with *gag* and RNase H, respectively. Sixty-three percent of the LTR retrotransposons exhibited homology with both integrase and reverse transcriptase among families, regardless of superfamily.

We narrowed our focus to seven each of the *gypsy* and *copia* superfamilies for a total of 14 families (Table 1). We chose families that had broad ranges of average insertion dates (45,000 yr to 1.54 Mya) and copy number (5 to 408), but also exhibited a high proportion of individuals with genes (i.e., ~80% with integrase and reverse transcriptase). Once the families were chosen for analyses, gene alignments were made at the protein level using ClustalW (Thompson et al. 1994) for each gene/family combination, and the DNA sequence was then used for further analyses.

Recently transposed families exhibit higher Ts:Tv ratios and fewer premature stop codons than older families

Gene alignments were evaluated in MEGA (v. 4.0) (Tamura et al. 2007). Using this program, we calculated the average transition: transversion (Ts:Tv) ratio as well as the number of sequences belonging to a family that exhibited at least one premature stop codon for each gene/family combination. Sequences that exhibited a premature stop codon were removed from all further analyses, as they would be expected to evolve neutrally and similarly to pseudogenes from the point in time when that stop codon arose; thus, the work presented here does not consider the evolution of defective or nonautonomous elements. We also removed elements from the analysis that exhibited a length greater than two standard deviations from the family average, as these are the elements that have previously been found to house nested elements and are putatively unable to replicate. On average, four elements from each of the high-copy families were removed from analysis; no individual elements from the low-copy families housed insertions. The average Ts:Tv ratio across all genes and all families was 3.76, suggesting that LTR retrotransposons are generally in an epigenetically silenced state (San Miguel et al. 1998). This is because host genes, including introns, exhibit an ~1:1 to ~2:1 ratio of transitions to transversions, and it has been shown that a higher Ts:Tv ratio is evidence of extensive cytosine 5-methylation, given that this epigenetic DNA modification increases the C-to-T transition rate (SanMiguel et al. 1998). Although the overall average Ts:Tv ratio is high, we found that it varied greatly among genes and families from 0.8 to 9.5 across both superfamilies (Table 2).

We also uncovered variation among genes and families for the proportion of individual sequences that exhibited premature stop codons—values ranged from 0% to >60% (Table 2). Previous work shows that 25% of *copia* reverse transcriptase sequences from the *O. sativa* genome exhibit premature stop codons, and this has been argued as evidence that many LTR retrotransposons are nonfunctional entities within host genomes (Navarro-Quezada and Schoen 2002). This number is in line with the range detected here, and likely represents an average across all *copia* families, but does not take into account the possibility that families may vary for the proportion of sequences within them that exhibit premature stop codons.

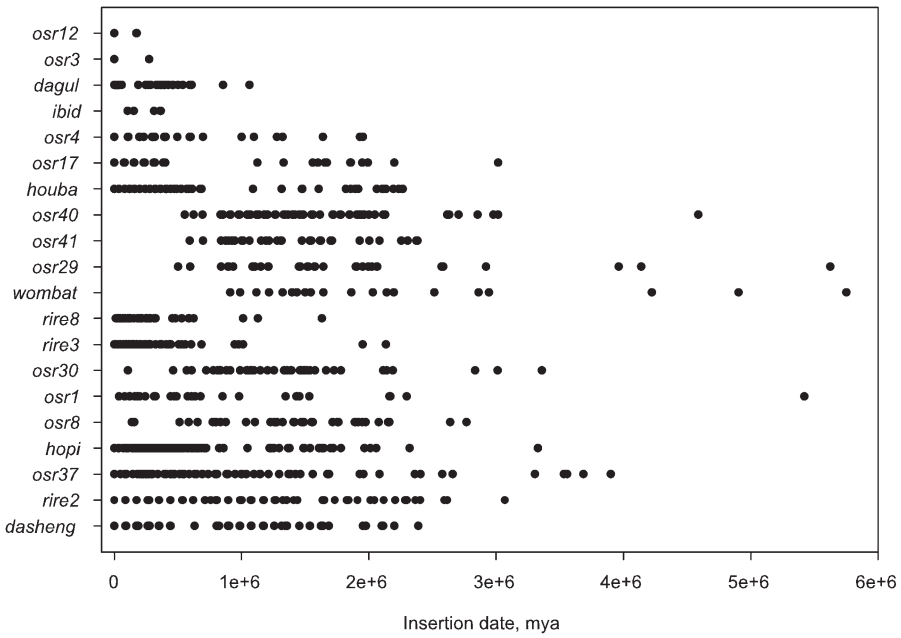


Figure 2. Insertion dates for each LTR retrotransposon from a representative sample of families from the *O. sativa* genome. The 5′ and 3′ LTRs were aligned using ClustalW (v. 1.3.8) and we determined their divergence using the BaseML module of PAML (v. 4). We then applied the formula $T = k/2r$ (k = divergence, $r = 1.3 \times 10^{-8}$) to estimate the insertion time of each element.

Table 1. Families and genes used in the analyses

			Genes[a]							
			INT		*RVT*		*GAG*		*RH*	
Superfamily[b]	Family[c]	Insertion date[d] Mya (SE),	Number[e]	Length[f]	Number[e]	Length[f]	Number[e]	Length[f]	Number[e]	Length[f]
gypsy	*rire2*	0.451 (0.034)	129	402	129	408	144	201	145	276
	hopi	0.355 (0.035)	286	354	285	408	258	255	286	306
	osr30	1.349 (0.092)	26	396	28	372	57	108	226	258
	rire3	0.282 (0.027)	58	408	64	408	64	273		
	osr41	1.426 (0.068)	26	417	30	393	27	195	35	264
	osr40	1.540 (0.066)	52	411	61	354	56	144	60	291
	ibid	0.259 (0.055)	5	387	5	408	5	219	5	234
copia	*osr8*	1.300 (0.153)	34	309	29	609				
	osr1	0.886 (0.169)	35	381	35	585				
	houba	0.671 (0.090)	58	450	59	633				
	osr17	0.638 (0.122)	18	363	31	606				
	osr4	0.522 (0.094)	31	441	30	600				
	osr3	0.045 (0.045)	6	462	6	675				
	osr12	0.145 (0.029)	4	366	6	675				

[a]GAG and RNase H were not assessed for *copia* families, and no LTR retrotransposons of *rire3* exhibited homology to RNase H.
[b]Superfamily of each LTR retrotransposon family.
[c]Families were delineated using the program RepMiner (Fig. 1) and the RetrOryza db (http://www.retroryza.org).
[d]Average insertion date of each family, using a substitution rate of 1.3×10^{-8} per site per year (Ma and Bennetzen 2004).
[e]Number of sequences used per family/gene combination.
[f]Length of nucleotide alignment, with gaps.
INT, integrase; *RVT*, reverse transcriptase; *GAG*, *gag*; *RH*, RNase H.

Furthermore, the variation among families in both the average Ts:Tv ratio and in the proportion of premature stop codons per family correlates significantly with the average insertion times of the LTR retrotransposon families (Fig. 3). Specifically, the resampled correlation coefficient between the average insertion date for each family and its Ts:Tv ratio was generally negative, meaning that younger LTR retrotransposon families in the *O. sativa* genome exhibit a higher Ts:Tv ratio than older families (Fig. 3A). After Bonferroni corrections, this relationship was significantly negative for *gag,* integrase, and reverse transcriptase of the *gypsy* families, with RNase H exhibiting a positive, but nonsignificant, correlation. Among *copia* families, the Ts:Tv ratio was found to significantly vary with average insertion time for both integrase and reverse transcriptase, providing evidence that the negative relationship between insertion date and Ts:Tv ratio is present in both superfamilies of LTR retrotransposons (Fig. 3A). These results provide indirect evidence for the epigenetic silencing of LTR retrotransposons within the *O. sativa* genome. Previous reports found that low-copy LTR retrotransposon families are more often expressed in the *Zea mays* genome than are high-copy elements (Meyers et al. 2001). While we did not observe a strong relationship between copy number and the average Ts:Tv ratio of families

Table 2. The Ts:Tv ratio and the proportion of individuals with premature stop codons for each gene/family combination

		Genes							
		INT		*RVT*		*GAG*		*RH*	
Superfamily	Family	Ts:Tv[a]	Prop. stops[b]	Ts:Tv[a]	Prop. stops[b]	Ts:Tv[a]	Prop. stops[b]	Ts:Tv[a]	Prop. stops[b]
gypsy	*rire2*	2.5	0.143	4.6	0.137	4.6	0.189	2.5	0.181
	hopi	4.4	0.091	4.2	0.053	4.5	0.058	4.3	0.115
	osr30	2.1	0.320	2.9	0.370	3.3	0.308	2.4	0.480
	rire3	3.9	0.069	4.7	0.000	7.7	0.094		
	osr41	2.8	0.348	4.5	0.172	4.1	0.296	5.7	0.294
	osr40	3.9	0.612	3.8	0.300	3.4	0.556	3.9	0.305
	ibid	3.6	0.000	4.0	0.000	1.5	0.000	1.5	0.000
copia	*osr8*	0.9	0.273	0.9	0.286				
	osr1	3.0	0.314	1.0	0.176				
	houba	4.5	0.088	2.2	0.103				
	osr17	1.9	0.176	1.6	0.600				
	osr4	9.5	0.167	7.7	0.207				
	osr3	5.8	0.000	2.7	0.333				
	osr12	2.0	0.000	9.4	0.167				

[a]Average transition:transversion ratio for each gene/family combination.
[b]Proportion of individual sequences within a family that exhibited a premature stop codon.
INT, integrase; *RVT*, reverse transcriptase; *GAG*, *gag*; *RH*, RNase H.

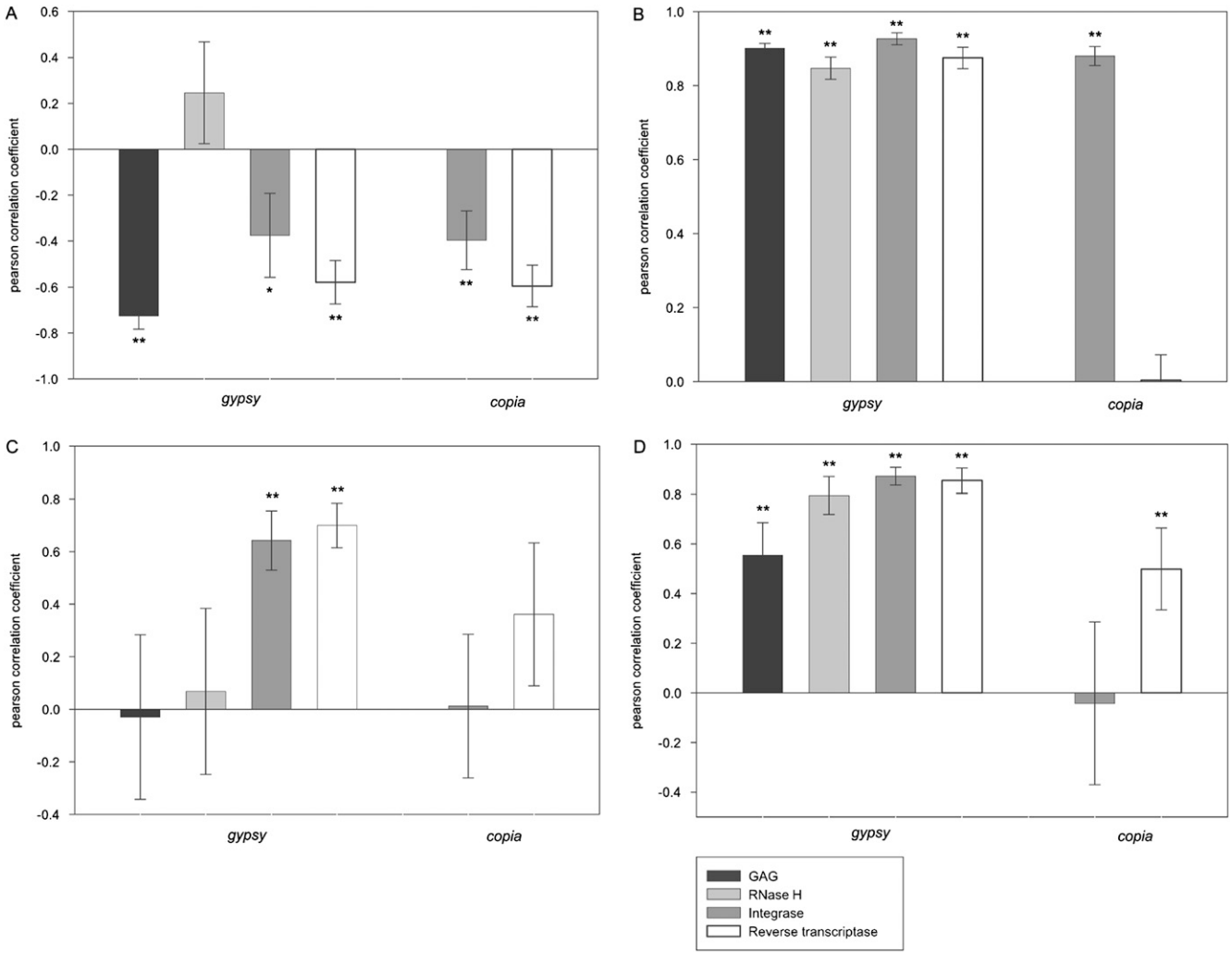


Figure 3. The strength and direction of the correlations between the average insertion date of the LTR retrotransposon families in rice and Ts:Tv (*A*), the proportion of premature stop codons within each family (*B*), π_s (*C*), and π_n/π_s (*D*), presented for each gene analyzed of each superfamily. Standard errors around the mean are represented by 95% confidence intervals, estimated from the jackknife procedure. (*) The correlation was significant at the $P < 0.05$ level; (**) significance at the $P < 0.005$ level after Bonferroni corrections. (Black bars) *gag*; (light gray bars) RNase H; (dark gray bars) integrase; (open bars) reverse transcriptase.

(data not shown), the two highest-copy *gypsy* families in rice are also the youngest.

We found the opposite relationship between insertion date and the proportion of sequences within a family that exhibited stop codons. There was a strong and significantly positive (resampled, Bonferroni-corrected) correlation between the proportion of individuals with stop codons and the average insertion date, except with the reverse transcriptase gene of the *copia* superfamily (Fig. 3B). This generally positive correlation suggests that older LTR retrotransposon families are inactive due to premature stop codons. Taken together, the relationships between the average insertion date of families and the Ts:Tv ratio and proportion of sequences with premature stop codons together suggest that full-length LTR retrotransposons are epigenetically silenced when young. As their age increases in the host genome, they are rendered nonfunctional through random mutation, perhaps enhanced by the higher rate of mutation at methylated cytosines (SanMiguel et al. 1998).

Recently transposed families exhibit lower gene diversity and lower π_n/π_s ratios than older families

We examined patterns of gene diversity among the LTR retrotransposons to determine whether the level of diversity suggested past regimes of selection. We estimated nucleotide diversity of each gene/family combination using the average number of pairwise differences per site between sequences for both nonsynonymous (π_n) and synonymous sites (π_s) (Nei 1987) using the program DnaSP (v. 4.00.5; Rozas et al. 2003). Contrary to the expectations of low gene diversity among all genes and families, the synonymous-site diversity (π_s) ranged from 0.012 to 0.591 (Table 3). Furthermore, we find that the majority of the variation resides at the synonymous nucleotide positions, suggesting that LTR retrotransposon genes are evolving under purifying selection (Table 3). More specifically, 88% of the π_n/π_s ratios are <0.40, indicating that these genes have experienced strong evolutionary constraints. There was a single exception to the generally low π_n/π_s ratios—the *gag* gene of the *gypsy* family, *osr40*, exhibited a π_n/π_s ratio of 2.73, suggesting

Table 3. Within-family estimates of sequence variation

		INT			*RVT*			*GAG*			*RH*		
Superfamily	**Family[a]**	π_n	π_s	π_n/π_s	π_n	π_s	π_n/π_s	π_n	π_s	π_n/π_s	π_n	π_s	π_n/π_s
gypsy	*rire2*	0.027	0.363	0.073	0.022	0.335	0.066	0.017	0.319	0.053	0.025	0.452	0.056
	hopi	0.015	0.197	0.074	0.016	0.172	0.093	0.011	0.239	0.046	0.019	0.238	0.082
	osr30	0.039	0.422	0.091	0.051	0.451	0.112	0.054	0.591	0.091	0.079	0.459	0.173
	rire3	0.006	0.155	0.038	0.007	0.152	0.045	0.006	0.232	0.026			
	osr41	0.048	0.317	0.153	0.035	0.339	0.103	0.026	0.196	0.132	0.054	0.279	0.195
	osr40	0.043	0.318	0.134	0.039	0.301	0.129	0.143	0.052	2.729	0.039	0.335	0.115
	ibid	0.099	0.275	0.359	0.006	0.028	0.227	0.165	0.269	0.613	0.012	0.027	0.459
copia	*osr8*	0.111	0.365	0.304	0.080	0.321	0.248						
	osr1	0.016	0.082	0.198	0.049	0.121	0.409						
	houba	0.006	0.025	0.227	0.010	0.027	0.373						
	osr17	0.006	0.027	0.210	0.014	0.031	0.447						
	osr4	0.016	0.090	0.178	0.011	0.043	0.250						
	osr3	0.020	0.216	0.091	0.016	0.250	0.064						
	osr12	0.176	0.329	0.533	0.003	0.012	0.240						

[a]Sequences with premature stop codons were removed prior to analysis.
INT, integrase; *RVT*, reverse transcriptase; *GAG*, *gag*; *RH*, RNase H; π_n, nucleotide variation at nonsynonymous positions of the codon; π_s, nucleotide variation at the synonymous codon position.

that it has evolved under a regime of positive selection. This gene/family combination also exhibited a very high proportion of sequences with premature stop codons; thus, members of this family, and specifically members that do not exhibit premature stop codons, are likely evolving under a scenario of relaxed evolutionary constraints rather than positive selection. Although the majority of the genes of the families exhibited very low π_n/π_s ratios, it is important to note that this measure ignores variation in the strength of selection through history (i.e., among branches on the gene trees [Yang 1998], across sites [Yang et al. 2000], or both [Yang and Nielsen 2002; Guindon et al. 2004]).

We uncovered a relationship between these measures of sequence evolution and the average insertion date of the LTR retrotransposon families (Fig. 3). The level of gene diversity, as measured by π_s, positively correlates with the average insertion date for the reverse transcriptase and integrase genes of the *gypsy* families (Fig. 3C), while the π_n/π_s ratio positively correlates to insertion date of all genes of the *gypsy* families and reverse transcriptase of the *copia* families (Fig. 3D). The positive relationship between insertion date and the level of gene diversity is potentially indicative of selective sweeps if transpositional bursts are due to positive selection and adaptation to the host genome. This is because selective sweeps would lead to lower genetic diversity of those families that had experienced the sweep—in this case, the families that most recently transposed. Unfortunately, little is currently known about the causes of transpositional bursts (i.e., are they due to element adaptation to the host genome or caused by a random process?), and thus the data are also consistent with expectations under neutrality, given that the younger element families have not had as much time to accumulate differences. Furthermore, that the π_n/π_s ratio increases with the age of the LTR retrotransposon family suggests that older families are either experiencing positive selection, at least at some codon sites, or relaxed selective constraints. Given that the majority of π_n/π_s ratios are less than one, and evidence that gene diversity also increases with the age of the LTR retrotransposon family, it is likely that the increased π_n/π_s ratios of the older families are due to lowered evolutionary constraints in comparison to the younger families.

Both superfamilies exhibit purifying selection, but allowing for rate "switches" suggests episodic positive selection

To assess the potential for selection on LTR retrotransposons, we performed an interfamily comparison by estimating the d_N/d_S (ω) ratio separately across both of the *copia* and *gypsy* superfamilies using the genes integrase and reverse transcriptase. We specifically wanted to investigate whether codon sites had experienced historical regimes of positive selection. Protein alignments were made with ClustalW using a single individual from each family of each superfamily, with two representative LTR retrotransposon genes mined from the *Z. mays* genome and used as outgroup sequences. We estimated the ω ratio in the codeml module of PAML (Yang 2007) using the M0 (one ratio), M1 (nearly neutral), M2 (positive selection), and M3 (discrete) models of sequence evolution. Comparison of twice the difference in log-likelihood scores between the M0 and M3 models assesses whether one or four ω ratios best fit the data, whereas comparisons between the M1 and M2 models determine whether the sequences are exhibiting evidence of positive selection (Yang and Nielsen 2002; PAML manual).

For both superfamily phylogenies (shown in Fig. 4) of both genes, we find no evidence for positive selection because none of the ω values are more than one under the M0 model of sequence evolution (Table 4). However, the M3 model of sequence evolution best fits the data, meaning that allowing the ω ratio to vary across the sequence provides a better fit than assuming one ω ratio (Table 5). Under the M3 model of sequence evolution, the highest ω ratio was found in the *copia* integrase gene, with 26% of the codon sites exhibiting a ω ratio of 0.72. The majority of the ω ratios estimated under this model of sequence evolution are very low, however, suggesting that the majority of sites are evolving under purifying selection. For example, 100% of the codons exhibit ω ratios $\leq$0.05 among the *gypsy* families for the integrase gene, and 74% of codons show a ω ratio of 0.07 for the reverse transcriptase gene. And, although the ω ratio of the integrase gene among the *copia* families was much higher in comparison to these values, 81% of the codon sites of this superfamily's reverse transcriptase gene were $\leq$0.04. Thus, the codon sites model of sequence evolution provides no evidence that particular codons are evolving under

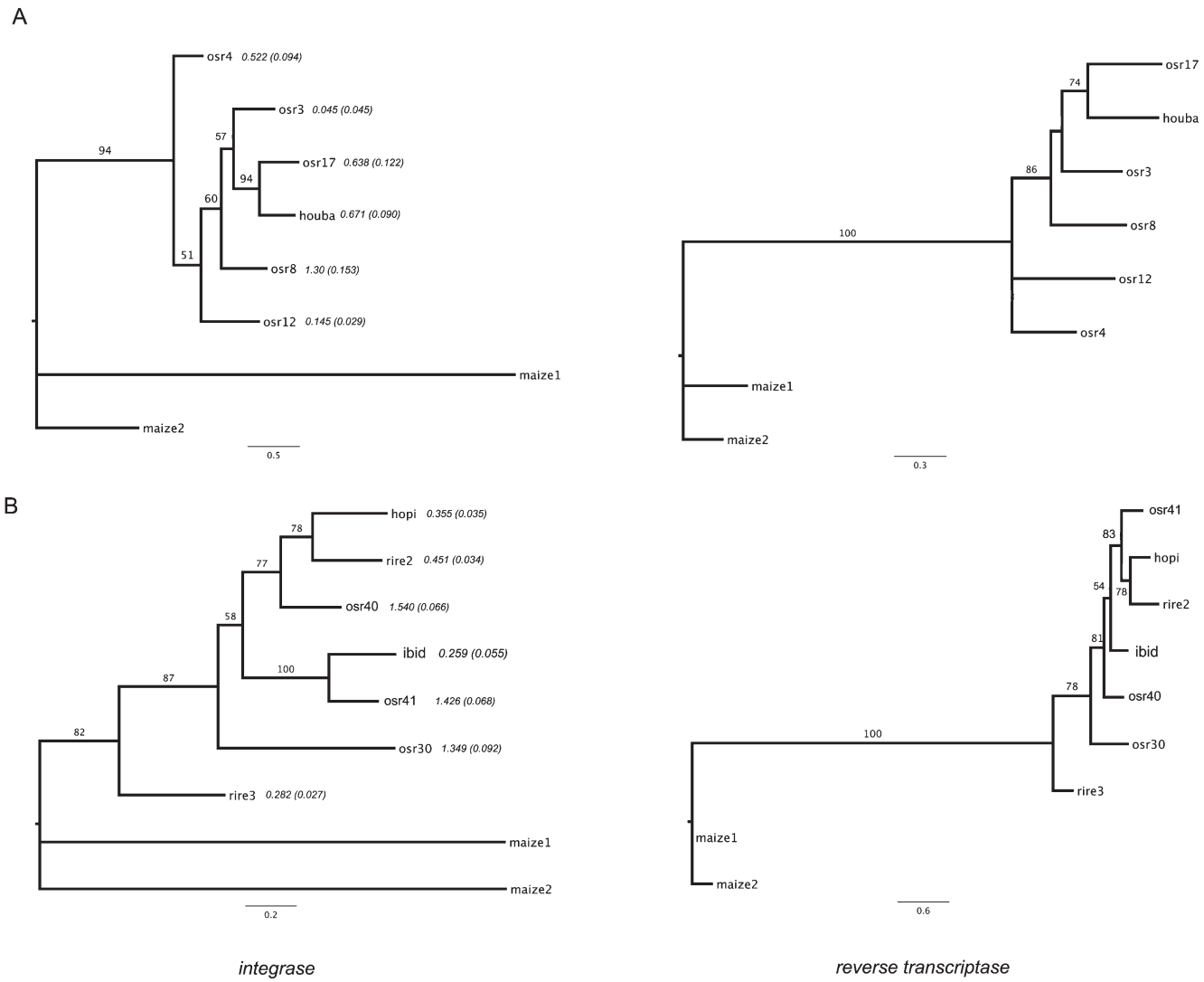


Figure 4. Phylogeny of the integrase and reverse transcriptase genes for *copia* (*A*) and *gypsy* (*B*) families. Phylogenies were made in PhyML. Bootstrapped values from 1000 pseudo-replicates of the data are presented next to the branch. Two *Z. mays* sequences were used as outgroups for each phylogeny. Average insertion dates (SE) for each family are presented on the integrase tree for each superfamily.

a regime of positive selection. Furthermore, the comparison between models M1 and M2 finds no evidence of positive selection for either gene of either superfamily (Table 5).

Although these models of sequence evolution allow for variation in selection among codon sites, they assume that selection affects all branches of the phylogeny equally. The widely used branch model in PAML that allows branches to vary requires an a priori assignment of the branch that is potentially experiencing positive selection, and thus is not applicable in this analysis, given that there was no reason to expect that one particular family was experiencing a different selective regime versus another. To take potential variation among branches in the superfamily phylogenies into account, we used the program fitModeL, which implements a maximum likelihood phylogeny-based codon-substitution model that includes parameters for switching between selection processes at individual codon sites across the phylogeny (Guindon et al. 2004). This means that the site-specific selection process is allowed to vary across branches of the phylogenetic tree, and in our case, allows for variation among families in the superfamily phylogeny.

Comparison of the M3 and M3 + 1 (discrete vs. switch) models in fitModeL provides evidence of positive selection among both genes and superfamilies (Table 5) at low to moderate equilibrium frequencies. For example, the integrase gene of *gypsy* families exhibited purifying selection 99.7% of the time, and positive selection 0.3% of evolutionary time with a ω ratio >13. For reverse transcriptase, 31% of the time we find a ω ratio of 2.69. The findings among *copia* families are similar, providing evidence of positive selection (ω ratio = 3.30) 23% of the time for integrase, and 30% of the time for reverse transcriptase (ω ratio = 2.69). These results suggest that allowing site-specific variation in the nonsynonymous/synonymous rate provides a better fit to these data than does averaging this ratio over lineages. Furthermore, the strength of positive selection that has acted on these LTR retrotransposon genes is likely underestimated when the site-specific variation of selection across lineages is not taken

Table 4. Likelihood ratio statistic and model parameters for each model of sequence evolution

Superfamily	Gene[a]	M0[b]	M1[c]	M2[d]	M3[e]	M3, fitModeL[f]	M3 + 1[g]
gypsy	Integrase						
	ln*L*	−3527.58	−3525.82	−3525.82	−3507.38	−3504.69	−3494.34
	w1 w2 w3	0.02	0.02 1.0	0.02 1.0 1.0	0.01 0.05 0.14 5.01	0.007 0.03 0.05	0.01 13.49 13.92
	p1 p2 p3	1.0	0.97 0.03	0.96 0.02 0.02	0.33 0.67 0 0	0.30 0.65 0.05	0.997 0.002 0.001
	Reverse transcriptase						
	ln*L*	−2936.57	−2919.04	−2919.04	−2910.30	−2907.39	−2843.85
	w1 w2 w3	0.12	0.10 1.0	0.10 1.0 1.0	0.07 0.33 5.05 22.46	0.03 0.11 0.42	0.001 0.05 2.69
	p1 p2 p3	1.0	0.86 0.14	0.94 0.06 0	0.74 0.26 0 0	0.27 0.56 0.17	0.36 0.33 0.31
copia	Integrase						
	ln*L*	−2919.23	−2881.98	−2881.98	−2874.57	−2872.48	−2860.61
	w1 w2 w3	0.04922	0.07 1.0	0.07 1.0 28.62	0.02 0.11 0.72 0.78	0.01 0.10 0.83	0.01 0.079 3.30
	p1 p2 p3	1.0	0.65 0.35	0.65 0.35 0	0.23 0.51 0.26 0	0.20 0.54 0.26	0.28 0.48 0.23
	Reverse transcriptase						
	ln*L*	−5452.60	−5406.24	−5406.24	−5391.83	−5390.10	−5358.46
	w1 w2 w3	0.022	0.03 1.0	0.02 1.0 5.75	0.005 0.02 0.04 0.15	0.007 0.02 0.08	0.004 0.12 2.69
	p1 p2 p3	1.0	0.82 0.18	0.82 0.18 0	0.05 0.38 0.38 0.19	0.25 0.55 0.20	0.23 0.47 0.30

[a]Sequence evolution further analyzed for reverse transcriptase and integrase among *gypsy* and *copia* phylogenies.
[b]M0 model, one rate ratio across all codon sites and all branches of phylogeny.
[c]Nearly neutral model, constrains w ratio to nearly 0 and 1.
[d]Positive selection model.
[e]Discrete model of sequence evolution with four categories (*k*).
[f]Discrete model of sequence evolution estimated by fitModeL for comparison to M3 + 1.
[g]Discrete model of sequence evolution, which allows "switches" of rates along phylogeny.

Table 5. Likelihood ratio tests of selection from PAML and fitModeL

Superfamily	Gene	M0:M3	M1:M2	M3:M3 + 1
gypsy	Integrase	df = 4	df = 2	df = 2
	2LL	20.20	0.00	10.35
	P-value	0.0005		0.0057
	Reverse transcriptase			
	2LL	26.27	0.00	63.54
	P-value	<0.0001		<0.0001
copia	Integrase			
	2LL	44.66	0.00	11.87
	P-value	<0.0001		0.0026
	Reverse transcriptase			
	2LL	60.77	0.00	31.64
	P-value	<0.0001		<0.0001

into account. There is a potential caveat to this analysis: Biased gene conversion could homogenize copies of LTR retrotransposons and thus affect our estimates of selection. We have attempted to minimize this impact, however, by only studying elements with matching target-site duplications (TSDs), thus removing elements with clear evidence of gene conversion tracts that overlap element ends.

Concluding remarks

There have been few investigations of molecular selection on the LTR retrotransposons that inhabit plant genomes. Of the available research, two reports find high levels of purifying selection on the reverse transcriptase gene region among LTR retrotransposons belonging to the *copia* superfamily from various dicot and monocot species (Navarro-Quezada and Schoen 2002), as well as across grass genomes (Matsuoka and Tsunewaki 1999). The results reported herein, in which we determined the prevalence of selection across 14 LTR retrotransposon families from two evolutionarily distinct superfamilies (Xiong and Eickbush 1990), demonstrate generally similar results—the predominance of high levels of purifying selection on the genes involved in the life cycle of these genomic parasites. Yet, allowing for site variability in the nonsynonymous/synonymous ratios leads to estimates that are greater than one, suggesting that positive selection potentially underlies LTR retrotransposon sequence evolution.

How can these disparate results be reconciled? A model of sequence evolution in which purifying selection across codons and evolutionary time predominates, but rare episodes of positive selection occur, allowing for population expansions and transpositional bursts, could resolve all of the apparent conflicts in these results. In this scenario, a regime of severe purifying selection is the norm but, rarely, a mutant LTR retrotransposon arises that is not recognized by the host genome nor subsequently silenced, at least when the LTR retrotransposon population remains at low frequency. It should thus transpose until it reaches a level that allows for host recognition, at which time it is silenced. That LTR retrotransposon proliferation is often described as due to transpositional "bursts" (Vitte and Panaud 2003) or "waves of genome invasion" (Wicker and Keller 2007) supports the idea that episodic regimes of positive selection and genome adaptation occur, followed by TE amplification and subsequent periods of inactivity.

The evolutionary dynamics of TEs are successfully modeled under "transposition/excision" balance, which predicts increased selection against TEs as they become more numerous in the genome (Langley et al. 1988; Charlesworth and Langley 1989). These hypotheses include the ectopic exchange model (Charlesworth and Langley 1989), the deleterious insertion model (Finnegan 1992), and the host fitness cost model (Nuzhdin 1999). These models assume that TE population sizes are managed through their effects on the host, i.e., their presence causes host lethality or reduced host fitness, which in turn selects on the extant population of TEs. There is substantial support for the role of ectopic exchange imposing selection on both DNA transposons (Montgomery et al. 1991; Bartolomé et al. 2002; Dolgin and Charlesworth 2008) and non-LTR retrotransposons (Petrov et al. 2003). However, for the interactions of TEs and the host to be a true co-evolutionary dynamic, there should also be the involvement of host defenses that reduce the ability of TEs to proliferate. This is not a new idea; rather, it was suggested as part of the parasitic DNA hypothesis (Orgel and Crick 1980) and has since been discussed by other authors (Bennetzen 1998; Nuzhdin 1999; Kidwell and Lisch 2001; McDonald et al. 2005).

Epigenetic silencing mechanisms, such as methylation, play an important role in mediating the effects of LTR retrotransposons on host genomes, and especially plant genomes (Okamoto and Hirochika 2001). It has even been hypothesized that the physical association between epigenetically silenced TEs and host genes in rice has led to host gene regulation (Zhang 2008). The prevalent evidence for the expansion of TEs following escapes from host silencing mechanisms (Weil and Martienssen 2008) suggests that consideration of another model is needed that could account for the control of transposition, one of "selective silencing" (in the sense of Zilberman and Henikoff 2005). This model does not require deviation from the parasitic DNA hypothesis, in that positive selection and adaptation to the host genome among the extant population of full-length elements can occur, so long as host defenses cannot recognize a variant allowing the population of TEs to reproduce. The data presented here do not distinguish among the different models suggested to control LTR retrotransposon copy number—the levels of selection that we uncovered could stem from both the effects of elements on the survival of their hosts and mechanisms by which hosts maintain their amplification. However, the increased Ts:Tv ratios of younger element families relative to older families in our data suggests that host methylation plays an important role in LTR retrotransposon population dynamics.

The significant relationships between sequence characteristics and the age of LTR retrotransposon families suggest that there is a "life-history" common to all LTR retrotransposons, such that selection against their proliferation through host defenses is of greater importance when families are young and are comprised of mostly full-length copies. In support of this hypothesis, Ma et al. (2004) found evidence that older LTR retrotransposon families from the *O. sativa* genome are highly truncated and deleted with a half-life of <6 million yr (Myr); this number was later revised to <3 Myr by a more realistic molecular clock analysis (Bennetzen et al. 2005). A recent analysis of *copia* elements from the *O. sativa* genome finds a substantially shorter half-life of 790,000 yr (Wicker and Keller 2007). Taken together, these data suggest that the life-history stages of LTR retrotransposons are comprised of birth through transposition, followed by "selective silencing" and other forms of host control (or lack thereof), with death by random mutation and eventual deletion from the genome being the final stages.

Methods

Sequence mining, naming, and dating

The DNA sequences of the twelve rice chromosomes (Release 4, IRGSP) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov), and LTR_STRUC was run on all chromosomes using default settings (McCarthy and McDonald 2003). We discovered a total of 2226 LTR retrotransposons and placed them into families using both the Perl program RepMiner (J. Estill; http://repminer.sourceforge.net) and a BLASTN search with an expected value of e^{-10} to the RetrOryza database (Chaparro et al. 2007). Results of the RepMiner process were visualized in Cytoscape (v. 2.5.1; http://www.cytoscape.org).

Perl programs were used to automate the LTR-retrotransposon dating process; the two LTRs of each mined LTR retrotransposon were first aligned using ClustalW (v. 1.8.3) (Thompson et al. 1994), and genetic divergence between the two was estimated using the baseml module of PAML (v. 4) (Yang 2007). The time since insertion of each LTR retrotransposon was estimated using the substitution rate of 1.3×10^{-8} per site per year (Ma and Bennetzen 2004). The GLIMMIX procedure of SAS (v. 9.1) was used to determine whether LTR retrotransposon families varied for their average insertion date. This procedure is particularly useful for fitting data that exhibit non-normality and/or nonconstant variance, which were both properties of the insertion date data (Kolmogorov-Smirnov D statistic = 0.15; $P < 0.01$), even after log-transformation (D-statistic = 0.25; $P < 0.01$). We modeled the data in GLIMMIX using a log-link function and a gamma distribution.

Isolation and alignment of genes from LTR retrotransposon families

All *gypsy* gene models were downloaded from the PFAM database (*gag*, PF00607; RNase H, PF00075; integrase. PF00665; reverse transcriptase, PF00078), and *copia* gene models were downloaded from NCBI. We performed separate BLASTX queries of each gene model to all of the LTR retrotransposed that we mined from the rice genome using an expect value of e^{-1}. Genes in each of the LTR retrotransposons were parsed from the full-length sequence and visualized in MEGA (v. 4.0) (Tamura et al. 2007) to ensure that all copies in these families began with a start codon and that they mirrored the sequence of the gene model (i.e., there were no parsing errors). Translated genes were aligned using ClustalW (v. 1.3.8) in MEGA (v. 4.0) (Tamura et al. 2007) using standard alignment values. We then determined the number of individuals within each gene/family combination that exhibited a premature stop codon, and returned to the nucleotide level for analysis.

Sequence analyses

The average Ts:Tv value was estimated for each gene/family combination in MEGA (v. 4.0). The program DnaSP (v. 4.00.5) (Rozas et al. 2003) was used to estimate levels of nucleotide polymorphism at both nonsynonymous and synonymous codon sites (π_n, π_s).

Selection analyses

Using a representative individual from each family within either *copia* or *gypsy* superfamilies, phylogenetic trees of each superfamily were made using integrase and reverse transcriptase in PhyML (Guindon and Gascuel 2003), which uses maximum likelihood under a general time-reversible model of nucleotide substitution. BLASTN at an expect value of e^{-10} against a database of *Z. mays* LTR retrotransposons was performed with a representative element from each rice LTR retrotransposon family to find appropriate outgroups. In addition, ~200,000 BAC-end sequences of *O. sativa* subsp. *japonica*, and *Oryza alta* were downloaded from OMAP (http://www.omap.org) and queried for each gene in BLASTN searches against our database of rice LTR retrotransposon genes. Subsequent alignments showed almost exact sequence matches between the LTR retrotransposon genes from any of these subspecies or species and the database of *O. sativa* LTR retrotransposon genes, such that they were not sufficiently diverged to be outgroup sequences to the identified rice LTR retrotransposons—for this reason, we chose *Z. mays* LTR retrotransposons as outgroups. To determine whether the genes exhibited varying levels of selection, the CodeML module of PAML (v. 4) (Yang 2007) was used to estimate the ω ratio across all branches and sites of each phylogenetic tree under the M0 (one-ratio), M1 (nearly neutral), M2 (positive selection), and M3 (discrete) models of sequence evolution (Yang et al. 2000). Twice the log-likelihood values of the M0:M3 and M1:M2 comparisons were assessed for statistical significance to determine whether a single versus four ω ratios best fit the data, and if the sequences exhibited positive selection, respectively. The log-likelihood ratio statistics are assumed to be χ^2 distributed with degrees of freedom equal to the difference in the number of parameters between models.

The potential for variable ω ratios over branches in the phylogenetic trees was assessed using the program fitModeL (Guindon et al. 2004). fitModeL is similar to PAML in that it is a codon-based model of DNA substitution that uses maximum likelihood for estimating parameters of sequence evolution, such as branch lengths, the transition/transversion ratio, equilibrium frequencies of the selection classes, and nonsynonymous/synonymous ratios (Guindon et al. 2004). However, it also estimates switching parameters, such that fitModeL allows changes between selection classes to occur through time. (i.e., switches between selection patterns in the phylogenetic tree at individual sites) (Guindon et al. 2004). We determined the statistical significance of log-likelihood ratios by taking twice the difference in log-likelihood estimates from the M3 and M3 + 1 models of sequence evolution.

We used the CORR procedure of the SAS statistical software package (v. 9.1) to estimate the strength and direction of correlations between the average insertion date of each family within each superfamily and the following: the average Ts:Tv ratio, genetic diversity as measured by π_s, the proportion of sequences with premature stop codons, and π_n/π_s. We performed these correlations for each gene for each superfamily, and subsequently removed the *gypsy* family *ibid* from these analyses, as it proved to be an outlier for π_n/π_s as determined by the ESD test (Neter et al. 1996). We assessed the significance of these correlations by resampling the data using the jackknife procedure across the families within each superfamily. We then used a one-tailed *t* statistic to calculate a *P*-value for the confidence interval in order to determine whether the correlations between the two measures were different than zero; this *P*-value was corrected for Type I errors using the Bonferroni procedure (Neter et al. 1996).

Acknowledgments

We thank Kelly Dyer for critically reviewing this manuscript. This research was supported by NSF DBI 0607123, and J.L.M. was supported by DEB 0732818 and DBI 0638595.

References

Ammiraju, J.S.S., Zuccolo, A., Yu, Y., Song, X., Piegu, B., Chevalier, F., Walling, J.G., Ma, J., Talag, J., Brar, D.S., et al. 2007. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. *Plant J.* **52:** 342–351.

Bartolomé, C., Maside, X., and Charlesworth, B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol. Biol. Evol.* **19:** 926–937.
Bennetzen, J.L. 1998. The structure and evolution of angiosperm nuclear genomes. *Curr. Opin. Plant Biol.* **1:** 103–108.
Bennetzen, J.L. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42:** 251–269.
Bennetzen, J.L., Ma, J., and Devos, K.M. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95:** 127–132.
Chaparro, C., Guyot, R., Zuccolo, A., Piegu, B., and Panaud, O. 2007. RetrOryza: A database of the rice LTR-retrotransposons. *Nucleic Acids Res.* **35:** D66–D70.
Charlesworth, B. and Langley, C.H. 1989. The population genetics of *Drosophila* transposable elements. *Annu. Rev. Genet.* **23:** 251–287.
Charlesworth, B., Sniegowski, P., and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371:** 215–220.
Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci.* **103:** 8101–8106.
Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12:** 1075–1079.
Dolgin, E.S. and Charlesworth, B. 2008. The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* **178:** 2169–2177.
Doolittle, W.F. and Sapienza, C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601–603.
Finnegan, D.J. 1992. Transposable elements. In *The genome of* Drosophila melanogaster. Academic Press, New York.
Gao, L.Z., McCarthy, E.M., Ganko, E.W., and McDonald, J.F. 2004. Evolutionary history of *Oryza sativa* LTR retrotransposons: A preliminary survey of the rice genome sequences. *BMC Genomics* **5:** 18. doi: 10.1186/1471-2164-5-18.
Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:** 696–704.
Guindon, S., Rodrigo, A.G., Dyer, K.A., and Huelsenbeck, J.P. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl. Acad. Sci.* **101:** 12957–12962.
Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A., and Wendel, J.F. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res.* **16:** 1252–1261.
Hudson, M.E., Lisch, D.R., and Quail, P.H. 2003. The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. *Plant J.* **34:** 453–471.
Kidwell, M.G. and Lisch, D.R. 2000. Transposable elements and host genome evolution. *Trends Ecol. Evol.* **15:** 95–99.
Kidwell, M.G. and Lisch, D.R. 2001. Perspective: Transposable elements, parasitic DNA, and genome evolution. *Evolution Int. J. Org. Evolution* **55:** 1–24.
Kumar, A. and Bennetzen, J.L. 1999. Plant retrotransposons. *Annu. Rev. Genet.* **33:** 479–532.
Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* **52:** 223–235.
Lin, R.C., Ding, L., Casola, C., Ripoll, D.R., Feschotte, C., and Wang, H.Y. 2007. Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. *Science* **318:** 1302–1305.
Ma, J.X. and Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* **101:** 12404–12410.
Ma, J.X., Devos, K.M., and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14:** 860–869.
Mackay, T.F.C., Lyman, R.F., and Jackson, M.S. 1992. Effects of P-element insertions on quantitative traits in *Drosophila melanogaster*. *Genetics* **130:** 315–332.
Mason, J.M., Frydrychova, R.C., and Biessmann, H. 2008. *Drosophila* telomeres: An exception providing new insights. *Bioessays* **30:** 25–37.
Matsuoka, Y. and Tsunewaki, K. 1999. Evolutionary dynamics of Ty1-copia group retrotransposons in grass shown by reverse transcriptase domain analysis. *Mol. Biol. Evol.* **16:** 208–217.
McCarthy, E.M. and McDonald, J.F. 2003. LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* **19:** 362–367.
McDonald, J.F. 1998. Transposable elements, gene silencing and macroevolution. *Trends Ecol. Evol.* **13:** 94–95.
McDonald, J.F., Matzke, M.A., and Matzke, A.J. 2005. Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenet. Genome Res.* **110:** 242–249.
Meyers, B.C., Tingley, S.V., and Morgante, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11:** 1660–1676.
Montgomery, E.A., Huang, S.M., Langley, C.H., and Judd, B.H. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: Genome structure and evolution. *Genetics* **129:** 1085–1098.
Navarro-Quezada, A. and Schoen, D.J. 2002. Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes. *Proc. Natl. Acad. Sci.* **99:** 268–273.
Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.
Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. 1996. *Applied linear statistical models*. McGraw-Hill, Boston, MA.
Nuzhdin, S.V. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107:** 129–137.
Okamoto, H. and Hirochika, H. 2001. Silencing of transposable elements in plants. *Trends Plant Sci.* **6:** 527–534.
Orgel, L.E. and Crick, F.H.C. 1980. Selfish DNA—the ultimate parasite. *Nature* **284:** 604–607.
Pasyukova, E.G., Nuzhdin, S.V., Morozova, T.V., and Mackay, T.F.C. 2004. Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. *J. Hered.* **95:** 284–290.
Petrov, D.A., Aminetzach, Y.T., Davis, J.C., Bensasson, D., and Hirsh, A.E. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol. Biol. Evol.* **20:** 880–892.
Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., et al. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16:** 1262–1269.
Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19:** 2496–2497.
SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., MelakeBerhan, P.A., Springer, S., Edwards, K.J., Lee, M., Avramova, Z., et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.
SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.
Shannon, P., Ozier, A.M.O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:** 2498–2504.
Tamura, K.J.D., Nei, M., and Kumar, S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24:** 1596–1599.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Vitte, C. and Bennetzen, J.L. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl. Acad. Sci.* **103:** 17638–17643.
Vitte, C. and Panaud, O. 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol. Biol. Evol.* **20:** 528–540.
Vitte, C. and Panaud, O. 2005. LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenet. Genome Res.* **110:** 91–107.
Vitte, C., Ishii, T., Lamy, F., Brar, D., and Panaud, O. 2004. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). *Mol. Genet. Genomics* **272:** 504–511.
Vitte, C., Panaud, O., and Quesneville, H. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): Recent burst amplifications followed by rapid DNA loss. *BMC Genomics* **8:** 218. doi: 10.1186/1471-2164-8-218.
Weil, C. and Martienssen, R. 2008. Epigenetic interactions between transposons and genes: Lessons from plants. *Curr. Opin. Genet. Dev.* **18:** 188–192.
Wicker, T. and Keller, B. 2007. Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res.* **17:** 1072–1081.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8:** 973–982.

Xiong, Y. and Eickbush, T.H. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9:** 3353–3362.
Yang, Z.H. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15:** 568–573.
Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24:** 1586–1591.
Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908–917.
Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Zhang, X. 2008. The epigenetic landscape of plants. *Science* **320:** 489–492.
Zilberman, D. and Henikoff, S. 2005. Epigenetic inheritance in *Arabidopsis*: Selective silence. *Curr. Opin. Genet. Dev.* **15:** 557–562.

Received July 15, 2008; accepted in revised form November 18, 2008.