Distribution, Diversity, and Long-Term Retention of Grass Short Interspersed Nuclear Elements (SINEs)

Hongliang Mao[1] and Hao Wang[1,2,*]

[1]Department of Physics, T-Life Research Center, Fudan University, Shanghai, P.R. China

[2]Department of Genetics, University of Georgia

*Corresponding author: E-mail: wangh8@fudan.edu.cn.

Accepted: July 29, 2017

Abstract

Instances of highly conserved plant short interspersed nuclear element (SINE) families and their enrichment near genes have been well documented, but little is known about the general patterns of such conservation and enrichment and underlying mechanisms. Here, we perform a comprehensive investigation of the structure, distribution, and evolution of SINEs in the grass family by analyzing 14 grass and 5 other flowering plant genomes using comparative genomics methods. We identify 61 SINE families composed of 29,572 copies, in which 46 families are first described. We find that comparing with other grass TEs, grass SINEs show much higher level of conservation in terms of genomic retention: The origin of at least 26% families can be traced to early grass diversification and these families are among most abundant SINE families in 86% species. We find that these families show much higher level of enrichment near protein coding genes than families of relatively recent origin (51%:28%), and that 40% of all grass SINEs are near gene and the percentage is higher than other types of grass TEs. The pattern of enrichment suggests that differential removal of SINE copies in gene-poor regions plays an important role in shaping the genomic distribution of these elements. We also identify a sequence motif located at 3′ SINE end which is shared in 17 families. In short, this study provides insights into structure and evolution of SINEs in the grass family.

Key words: SINE, transposable elements, genome evolution, comparative genomics.

Introduction

Short interspersed nuclear elements (SINEs) are Class I transposable elements (TEs) that moves in a copy-and-paste mode via an intermediate RNA (Wicker et al. 2007). SINEs are nonautonomous elements because they do not encode enzymes which are necessary for transposition, and thus cannot transpose by themselves. It has been proposed that SINE uses the enzymatic machinery of autonomous long interspersed nuclear elements (LINEs) to transpose (Boeke 1997; Kajikawa and Okada 2002; Dewannieux et al. 2003). SINEs are 80–1,000 bp in length and their genomic copy number (CPN) varies greatly between families and host species. Canonical structural characteristics of SINE include 1) short in size (<800 bp), 2) a 5′ end related to one of small cellular RNAs synthesized by RNA polymerase III (tRNA, 7SL RNA, or 5S rRNA), 3) a 3′ end composed of simple repeats like poly-A and poly-T, or tandem array of 2–3 bp unit, and in many elements 4) target site duplications (TSD) of variable sizes (see summary in, e.g., Vassetzky and Kramerov 2013).

The first SINE, human *Alu* element, was discovered over 40 years ago (Schmid and Deininger 1975; Houck et al. 1979) and following *Alu*, the last three decades have observed great advance in studying SINEs in animals (e.g., see reviews in Kramerov and Vassetzky 2011). In plants, the first SINE, rice p-SINE, was reported in 1991 (Umeda et al. 1991). To date, plant SINEs have been studied in many families (see summary in Deragon and Zhang 2006; Wenke et al. 2011). Early small-scale experimental studies of single SINE families, observed that some families, for example, AU (Yasui et al. 2001), TS (Yoshioka et al. 1993), and S1 (Deragon et al. 1994), showed high level of conservation in evolution. For example, AU was proposed originating at least in the common ancestor of Angiosperms (Fawcett and Innan 2016). Another observation was that some SINE families were enriched in euchromatic regions (Yasui et al. 2001). Later genome-wide surveys found that the majority of plant SINEs were enriched in distal chromosomal regions (Baucom et al. 2009; Wenke et al. 2011; Schwichtenberg et al. 2016; Seibt et al. 2016). However, counter examples existed. The *Brassica* family S1 was found in favor of pericentromeric regions (Goubely et al. 1999).

To date, the observations of the above two characteristics of plant SINEs have been well documented. However, their extent, mechanisms and evolutionary impact have not been systematically evaluated in general. Specifically, the overall levels of conservation and enrichment have not been well quantified yet. Addressing the question requires comprehensively identifying SINEs and comparing them to other genomic components like genes and TEs in well-sampled taxonomy groups. In this report, we annotate SINE in publically accessible 14 grass and 5 outgroup angiosperm genomes using SINE_Scan (Mao and Wang 2017) and perform comparisons between SINE families, between SINE and conserved gene families, and between SINE and other TEs. Our results made a series of new findings on structure, genomic distribution, and amplification dynamics of grass SINEs. Specifically, we show that SINEs are far more conservative than LTR retrotransposons and DNA transposons in terms of retention of early originated families, Moreover, our data show that in general, SINEs have higher level of enrichment in gene-rich regions than other TEs and families of ancient origin are more enriched than "young" families, which suggest that biased removal of "old" elements in nongene-rich regions play a key role in shaping currently observed pattern of enrichment if suppose "old" and "young" SINE families have on average similar insertional preference.

Materials and Methods

Genomic Data and Known Grass TE Data

Plant genomes and known grass LTR elements and short TIR elements used in this study were downloaded from multiple public databases or generated in our lab. Information of these sequence data was listed in supplementary tables S10 and S11, Supplementary Material online.

Identification of Grass SINEs

We searched SINEs in 14 published grass genome assemblies using SINE_Scan v1.1 (Mao and Wang 2017) and identified 59 SINE families under default parameter settings. All of the families were confirmed by manual inspection. Previously reported SINEs were deposited in Repbase (www.girinst.org/repbase/), PGSB Repeat Element Database (previously MIPS repeat database, http://pgsb.helmholtz-muenchen.de/plant/recat/), and SINEBase (http://sines.eimb.ru/). We compared the 61 families with valid grass SINEs deposited in these databases to identify new families. Here, valid means that these SINE have at least five qualified hits (*e*-value $\leq$ 1e-10 and the best high-scoring segment pairs cover at least 80% of the query sequence) in host genome using BLASTN search. Some known families were not found by SINE_Scan because they had noncanonical A-box and/or B-box sequences or distance between A-box and B-box was out of normal range. These families were also included in our analysis.

Genome-Wide Search of Qualified Copies of SINE Families

The representative sequences of the 61 SINE families were used as repeat database to search against the 19 plant genomes using RepeatMasker v4.0.5 (http://www.repeatmasker.org). We only kept qualified hits for further analysis (see above paragraph for meaning of qualified).

Domain Identification

To search conserved domains (sequence motifs) in these SINE families, we first excluded simple repeats (poly-A/T) at the 3′ end of SINEs and built consensus sequences for all families. To obtain the consensus at each column in the multiple sequences alignment of a family, the percentage of each kind of base at the column was calculated and the base with the highest percentage went into the consensus sequence if the percentage is higher than 80%; otherwise, there was no consensus at that column. Each consensus sequence was divided into two parts: the first 80 bp was taken as tRNA head, and the rest as body. We next applied local pairwise alignment (using *matcher* program in EMBOSS package v6.6.0 [Rice et al. 2000]) to tRNA heads to discover highly similar ($\geq$80% identity) heads. We identified conserved domains in the body region by 1) using local pairwise alignment to find highly similar strings ($\geq$80% identity) occurred in at least five families; and 2) a set of highly similar strings are considered a domain if their location are collinear in all families.

Average Intergenomic Conservation of SINE or Gene Families

If a SINE or gene family occurred in two genomes, the intergenomic conservation of this family was defined as the proportion of identical bases between the reciprocal best hits in the global alignment (the *stretcher* program in EMBOSS package version 6.6.0) (Rice et al. 2000). If a family occurred in more than two host genomes, we first calculated conservation for all two-genome combinations and took the average of these conservation values as average intergenomic conservation of this family.

Results

Structure

We identified a total of 61 SINE families (29,572 qualified copies) in the 14 grass genomes (fig. 1). Among them, 46 families were first described and the other 15 matched previously reported elements deposited in SINEBase (http://sines.eimb.ru/) or Repbase (www.girinst.org/repbase/). Detailed information of these families can be found in supplementary table S1, Supplementary Material online. The 61 families were

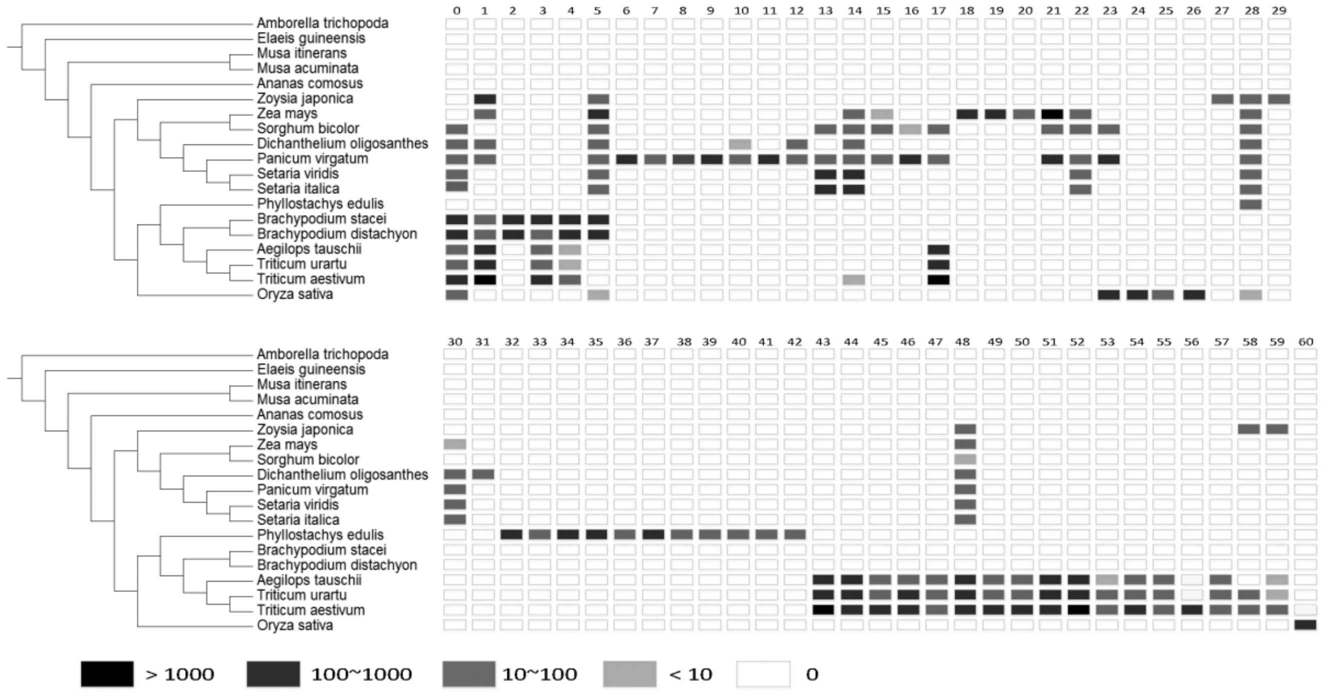

Fig. 1.—Taxonomic distribution and abundance of 61 grass SINE families. Each column shows CPN of a SINE family in the 19 genomes. Presence of families is shown by rectangles with gray gradient. Absence is represented by white rectangles. Family names are abbreviated by exclude the prefix of "Grass_." For example, Grass_0 is abbreviated as 0.

all tRNA type, and all showed canonical structure of tRNA head + body region + Poly-A/T tail. Poly-A and Poly-T tail occurred in 35 and 26 families. The length of these SINEs ranged from 102 to 399 bp, where length was calculated by excluding poly-A/T tail since this part was highly variable even between copies in a family. The distribution of size of the 61 grass SINEs exhibited bimodal pattern with one peak located at ~150 bp and the other at ~300 bp (supplementary fig. S1, Supplementary Material online).

We found the sequence motif of 5′-(CTT)$_n$AA-3′ (called CTT domain hereafter) occurred in 17 grass SINE families (supplementary table S2, Supplementary Material online). Four of these families originated in common grass ancestor and the other 13 were lineage specific. In the pattern 5′-(CTT)$_n$AA-3′, CTT usually repeated three or four times, but $n = 2$ was seen in one case (Grass_43). CTT-derived variants like CTTT and CT were also frequently observed. This CTT domain was located ~32 bp upstream of poly-A/T tail (supplementary table S2, Supplementary Material online).This sequence and location of this CTT domain was different from previously reported rice 5′-TTCTC-3′ domain (Tsuchimoto et al. 2008) in both sequence and location. The rice 5′-TTCTC-3′ domain was directly adjacent to the poly-A/T tail. The broad occurrence of conserved CTT domain in grass SINEs suggested that this domain might play critical role in SINE biology, but the exact function needed to be further identified.

Taxonomic Distribution and Dates of Origin

Forty-four percentage (27 of the 61) SINE families were lineage-specific given current taxon sampling (supplementary table S1, Supplementary Material online and fig. 1). For the other 34 families occurred in more than one genome (called conserved families hereafter), we used the most recent common ancestor (MRCA) of host genomes to represent the dates of their origin. Since one aim of this study was to evaluate the conservation of SINE families, we employed a stringent criterion on the occurrence of a SINE family (at least five copies ≥80% identity to representative sequence; see Materials and Methods) and thus the MRCA inference gave lower-limit of the origin date. For example, previous studies reported broad distribution of Grass_1 (AU) in angiosperms, including in sorghum, foxtail millet, banana, oil palm, and *Amborella* (Fawcett et al. 2006; Yagi et al. 2011; Fawcett and Innan 2016). Our criterion of qualified copies led to reporting absence of AU in the five species. However, if using weaker criterion, copies of AU could be found within all of these species. Moreover, adding more species from other grass lineages might also shift the origin to an earlier date. In short, our estimation of the dates of origin was lower-limit. According to figure 1, 18 families could be dated as originated before the divergence of grasses (ten families: Grass_0, Grass_1, Grass_5, Grass_14, Grass_17, Grass_23, Grass_28, Grass_48, Grass_58, Grass_59), before the divergence of Panicoids (six families: Grass_13, Grass_15, Grass_16, Grass_21, Grass_22, Grass_30), or Pooideae species (two families: Grass_3, Grass_4). We call them "old" families in the following texts. Total CPN of the 18 families was 14,493, ~50% of all annotated SINEs.

Abundance of SINE Families

CPN of SINE families ranged from 16 to 3,321 (median = 152). Nine families had CPN ≥ 1,000, 30 had CPN between 100 and 1,000, and the other 22 had CPN < 100 (fig. 1 and supplementary table S1, Supplementary Material online). The nine (15%) top-CPN families were composed of 62.4% (18,460/29,572) of total copies identified. Supplementary table S5, Supplementary Material online, showed CPNs of all SINE families in their host genomes. In their host genomes, SINE families also showed similar pattern: the 20% top-CPN families accounted for >50% of all SINE copies in all but three genomes. Even in the three species (sorghum and *D. oligosanthes* and *P. edulis*), the 20% top-CPN families accounted for over 40% of all SINE copies (supplementary table S3, Supplementary Material online). In 12 out of 14 (86%) grass genomes, 20% top-CPN families contained at least one "old" family, and in six genomes, all 20% top-CPN families were "old" families.

Genomic Distribution

We investigated the distribution of SINEs across chromosomes by normalization of chromosome size for seven species (supplementary fig. S2, Supplementary Material online) in which pseudo chromosomes and rough location of centromeres were accessible. Consistent to previous observation, the distribution of SINEs in the seven genomes showed that SINEs were enriched in the far ends of chromosome arms and avoided centromeric/pericentromeric regions (fig. 2*a* and supplementary fig. S2, Supplementary Material online).

We further characterized the relationship between SINEs and genes in 13 grasses of which gene annotations were accessible. We found 23,831 out of the 29,572 SINEs could be linked to genes in the same scaffolds or pseudochromosomes. The other 5,741 SINEs were located in scaffolds that had no gene annotation (supplementary table S4, Supplementary Material online). 17,763 out of the 23,831 SINEs were located in intergenic regions (supplementary table S5, Supplementary Material online). We investigated the SINE content in continuous 100 bp windows for these genes and found the distance between SINEs and genes exhibited unimodal distribution and the peaks were located around 400 bp from translation start or stop sites (fig. 2*b*). Similar patterns were observed in all 13 genomes investigated (supplementary fig. S3, Supplementary Material online).

A total of 6,068 SINEs were located inside genes and genes of PACMAD species tended to harbor more SINEs than BEP species. In *Zoysia japonica* and maize, the numbers of SINEs inside genes were greater than that in intergenic regions. Most (96% = 5,824/6,068) of SINEs inside genes were located inside introns (supplementary table S6, Supplementary Material online). In total we found that 40% ([6,068 + 3,551]/23,831) SINEs were inside or within 1 kb up- or downstream of genes (supplementary tables S5 and S6, Supplementary Material online). Since the genes and their 1 kb flanking regions only accounts for a small fraction (1,843 Mb [within 1 kb]/10,954 Mb [outside 1 kb]) of these genomes, the percentage 40% is statistically higher than expectation of random insertion along the genome (Binomial test, *P* value < 2e-16). The results suggest that SINEs enriched inside or near genes and the characteristic is conserved in grasses. It is worth noting that the 18 "old" SINE families have much higher proportion (51% = 6,479/12,651) of family members located in or within 1 kb around genes than the other 43 families (28% = 3,140/11,180). Pearson's Chi-square test suggested that the difference is significant (*P* value < 2e-16).

Long-Term Retention of SINEs

We studied the amplification dynamics of SINE families by comparing their CPN to that of Long Terminal Repeat (LTR) retrotransposons and nonautonomous Terminal Inverted Repeat (TIR) elements (including CACTA, hAT, PIF/Harbinger, Mutator, and TC1/Mariner transposons). We selected these two groups of TEs because 1) they represented major Class I and Class II plant TEs; 2) they were more effectively annotated in grass genomes than other TEs like LINE or Helitron, and thus provided the basis for less biased inferences. We selected five pairs of species representing divergence events happened at different time points (table 1). For each pair of genomes, we used SINEs, LTR, and TIR families in one species to search family members in the other species and found CPNs of conserved families. Known LTR elements and TIR elements in these species were extracted from multiple resources and were classified into families using the same criteria as our SINE family classification (see Materials and Methods; supplementary table S7, Supplementary Material online). The results (table 1) showed that 1) a considerable fraction of families of all three types of TEs were conserved between *Ae. tauschii* and bread wheat, two species diverged around 5 Ma; 2) the percentage of common LTR and TIR elements drastically decreased to <5% when divergence time ≥ 10 Ma, while the percentage of common SINEs maintained a quite high number (~50%) at these time points; and 3) in all three types of TEs, the percentages of conserved families between rice and *Ae. tauschii* were low, but the value of SINEs was still much higher than LTR and TIR elements. The results showed that SINEs were far more conserved than LTR and TIR elements.

Correlation to Genome Size

SINE content (measured by total size or total CPN) was strongly correlated with the genome size (fig. 2*c* and supplementary fig. S4, Supplementary Material online). In both

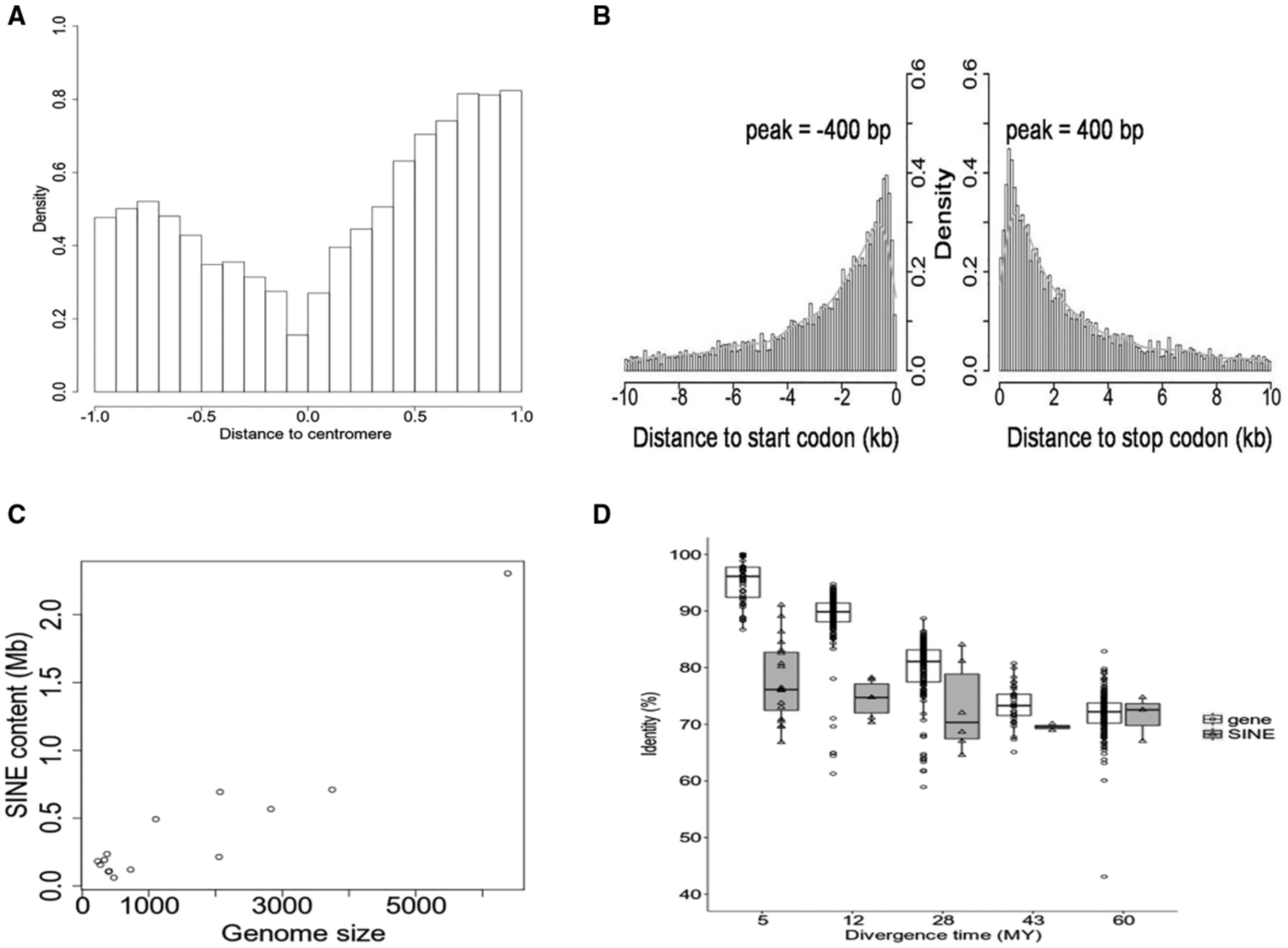

FIG. 2.—Key characteristics about genomic distribution, abundance and evolutionary speed of grass SINEs. (*a*) Chromosomal distribution of grass SINEs. Pseudo-chromosomes of seven genomes are used. Short and long arms of chromosomes are normalized separately, with the centromere is located at 0 and the short and long arm end at – 1 and 1, respectively. The locations of genes are calculated as distance from centromere divided by total length of chromosome arm, where distance is a negative number for the short arm. (*b*) Distribution of SINEs around genes. Up- and downstream 10 kb is shown. (*c*) Correlation of the SINE content and genome size. (*d*) Comparison of sequence divergence between highly conserved gene families and SINE families at five time points. At every time point, the distributions of sequence identity of genes and SINE families are showed side by side. Key features of these distributions are captured by standard box plot.

cases, Pearson's correlation coefficients were 0.92, and *P* value was 2.5e-6 for total size and 3.7e-6 for CPN.

Family Homogeneity

We performed pairwise alignment for members of each family and obtained family homogeneity by calculating average DNA similarity of these pairs. Family homogeneity of the 61 families varied from 67% to 90.5% and exhibited a unimodal distribution: around the peak at 75%, 70% families had family homogeneity value from 70% to 80% (supplementary fig. S5, Supplementary Material online).

Comparison to Gene Family

We investigated intergenomic conservation for the 34 conserved families by comparing them to previously identified highly conserved low-copy angiosperm gene families (Zhang et al. 2012). We first investigated how conservation of SINE families changed over time. Not considering horizontal gene transfer (HGT) and gene conversion, the distance of orthologous genes or SINEs should positively correlate with the divergence time of genomes. Using best reciprocal hits to represent orthologs, we calculated the sequence similarity of SINE and gene families at the five time points (see table 1) representing species divergence event happened from 5 to 60 Ma (fig. 2*d*). The intergenomic conservation of both SINEs and genes tended to decrease with the increase of the divergence time. At four time points, median values of sequence similarity of gene families were higher than SINE families and gene families had smaller interquartile range than SINE families. Unlike gene families, whose median similarity decreased over time, median similarity values of SINE families were

Table 1

Retained Conserved TE Families in Species Diverged at Different Times

Species Pair	Divergence Time (Ma)	TE Category	No. of Families in A	No. of Families in B	No. of Common Families	% of Common Families[a]
O. sativa–S. bicolor	60	LTR	27	364	0	0.00
		TIR	414	412	8	1.94
		SINE	8	12	4	40.00
Ae. tauschii–O. sativa	43	LTR	382	27	0	0.00
		TIR	373	414	3	0.76
		SINE	20	8	1	7.14
S. italica–Z. mays	28	LTR	74	177	2	1.59
		TIR	78	609	6	1.75
		SINE	8	12	6	60.00
S. bicolor–Z. mays	12	LTR	364	177	2	0.74
		TIR	412	609	26	5.09
		SINE	12	12	7	58.33
Ae. tauschii–T. aestivum	5	LTR	382	234	93	30.19
		TIR	373	229	57	18.94
		SINE	20	23	20	93.02

[a] Percentage = 100 * 2 * # of conserved families/(# of all families in the pair of species).

around 70–80% at all five time points. One reason underlying this was that we only studied SINE copies that shared ≥80% similarity with family representative sequence (see Materials and Methods).

We further compared the conservation of SINEs to genes in all host genomes. We drew scatter plot on average intergenomic conservation (see Materials and Methods) of SINE families and gene families. Each data point was calculated using family members of the same groups of host genomes. Consistent to the above results, the results showed that average intergenomic conservation of SINE families were lower than gene families (supplementary fig. S6 and table S8, Supplementary Material online).

Structural Evolution

We observed domain shuffling between three groups of families. 1) domain shuffling between rice SINE Grass_23 (OsSN1), Grass_0 (OsSN2), and Grass_24 (OsSN3) was reported by (Tsuchimoto et al. 2008): OsSN1 and OsSN2 had highly similar body but different head, while OsSN1 and OsSN3had highly similar head but different body. Figure 1 showed that Grass_23 and Grass_0 were widely distributed in grasses, but OsSN3 were rice specific. To investigate the origin of these families, we performed homology search of these SINE families against 11 publically accessible *Oryza* genomes and a close relative species *Leersia perrieri* (supplementary table S9, Supplementary Material online). The results showed that OsSN3 only occurred in genus *Oryza* while Grass_23 and Grass_0 were found in *L. perrieri*. This indicated that the origin of OsSN3 happened after the divergence of *Oryza* from other species of the tribe Oryzeae. The *Oryza* orgin of OsSN3 is based only one outgroup species, adding more high-quality outgroups in future is needed to further test this hypothesis. 2) Grass_11and Grass_17 had highly similar head but different body. The origin of Grass_17 was traced to the common ancestor of grass, while Grass_11only occurred in *P. virgatum* (fig. 1). 3) Grass_0, Grass_17, Grass_23 and Grass_43 had highly similar body but different tRNA head. Grass_0, Grass_17, Grass_23 widely occurred in Panicoideae and BEP species, while Grass_43 only occurred in Triticeae. It is well known that domain shuffling underlies animal SINE evolution (see, e.g., Kramerov and Vassetzky 2011). These results seem to suggest that domain shuffling also frequently occur during the evolution of plant SINEs.

Besides domain shuffling, point substitution and insertion–deletion were also considered as important evolutionary forces underlying structural evolution of SINE (Kramerov and Vassetzky 2011). We found a group of eight SINE families which differentiated through a series of insertion–deletion events. The eight families were highly similar in nonindel regions and alignment gaps were located in the internal regions (supplementary fig. S7, Supplementary Material online). Three terminal duplicates could be identified at the boundary of indels. In the three duplicates, one was located within a 93 bp tandem repeats, and the other two were of 2 and 13 bp in size. The above patterns suggest that these indels might be generated by intragenomic illegitimate recombination (Ma et al. 2004).

Discussion

We have studied the distribution, diversity, and evolutionary patterns of grass SINEs. All identified grass SINEs are tRNA-derived. Like other types of TEs, in every genome investigated, the majority SINE copies come from a few families while the majority of families keep moderate copy numbers. Similar to grass LTR elements, SINE content positive correlates to genome size strongly (fig. 2*c*). Unlike other grass TEs, where the most abundant families in a genome are lineage-specific families in at least most (if not all) studied cases (Devos 2010), "old" SINE families of early origin are found among top-ranked abundant SINE families in over 85% genomes (supplementary table S3, Supplementary Material online). We also discover that 28% grass SINE families share a strong sequence motif at the 3′ end. Besides these, our comprehensive survey have revealed that a considerable fraction of SINE families have been retained in host genome for tens of millions of years and exceptionally high level of enrichment near genes.

Exceptional Long-Term Retention of SINE Families

It has been well documented that grass TE elements are highly dynamic (see review in, e.g., Devos 2010). For example, grass LTR elements are believed to amplify via a species-specific burst manner and "old" elements are removed from genome rapidly, usually within 2–4 Ma (Devos et al. 2002; Ma et al. 2004). The rapid amplification and removal lead to dramatic variance of composition of LTR elements even between closely related species. At sequence level, only protein domains of transposase (e.g., RT) were conserved, and noncoding parts were highly lineage-specific. In contrast, SINEs such as AU elements, have long been found distributed broadly within flowering plants (Yasui et al. 2001; Fawcett and Innan 2016). However, these analyses have only focused on single families and cannot quantify the level of conservation of SINEs in general.

Our analysis suggests that grass SINE families are more likely to be retained in host genomes for longer time than other types of TEs: MRCA analysis found at least 18 families originated before diversification of grass major lineages (fig. 1) and the percentage of conserved SINE families greatly outnumbers other types of TEs: most common LTR and short TIR families are only found in species diverged around 4–5 Ma (table 1). We also found that "old" family has on average much more elements than a younger family (702 = 12,651/18 vs. 260 = 11,180/43), which may be the result of ancient bursts of these families and/or recent independent bursts of these families in multiple host genomes.

On the other hand, our data show that the average evolutionary rate of SINE families are higher than highly conserved genes (fig. 2*d* and supplementary fig. S6, Supplementary Material online) and that sequence insertion–deletion alters the structure of SINEs (supplementary fig. S7, Supplementary Material online), which indicates that removal of SINE copies is an ongoing process. Unlike in primate genomes, where the total length SINEs can reach >15% of the genome, SINEs only account for 1% or less of these grass genomes. However, their high level of conservation and significantly close physical association to genes (*P* value < 2e-16; and see below) suggest that this group of TEs may play important roles in grass genome evolution, or suggest the sequence-level conservation are critical to ensuring the transposition and thus the proliferation of SINEs, or both.

Besides evolutionary conservation, HGT can also result in highly similar sequences in distantly related species. Although cases of HGT of DNA transposon and LTR elements have been reported (Diao et al. 2006; El Baidouri et al. 2014), no well-defined case of HGT of plant SINEs has been reported yet. Our data do not strongly support to HGT having great contribution to observed conservation pattern of SINEs. According to figure 1, some families (e.g., Grass_58 and Grass_59) show patchy distribution in the phylogenetic tree. However, we could not found a case in which sequence similarity between family members in distantly related taxa is high enough to make HGT the most convincing explanation.

SINEs Show Higher Level of Enrichment in Gene-Rich Regions than Other TEs

In 6,068 SINEs that are inside genes, 244 are located in or overlapping with exons (supplementary table S6, Supplementary Material online). The ratio 244:5,824 is much lower than 330*5:327*4 (the average size of exon and intron of grass genes are 330 and 327 bp; a typical grass gene has on average five exons and four introns [Wang et al. 2014]). The difference is statistically significant (Binomial test, *P* value < 2e-16). If assuming the insertion of SINEs is random, the above result was consistent to the observations that most protein-coding regions were under strong purifying selection (see, e.g., Kimura 1986). We note that the reported 244 exon-related SINEs were solely based on gene annotations of generated by sequencing projects of these genomes. Due to rather low accuracy of capturing correct exon–intron structure in all current computational gene prediction methods, further investigations are needed to draw solid conclusions about the contribution of SINEs to grass protein-coding genes.

Our data show that 40% identified SINEs are located inside or within 1 kb up- or downstream of genes (fig. 2*b* and supplementary fig. S3 and tables S5 and S6, Supplementary Material online). The number 40% is similar to Tc1/Mariner DNA transposon (39% including *Stowaway* elements; [Han et al. 2013]; results based on five species and the same definition of "near gene"), and much higher than any other types of DNA transposons (CACTA: 10%; hAT: 25%; Mutator: 32%; PIF/Harbinger [including *Tourist* elements]: 32%; Han et al. 2013). In addition, previous studies have showed that grass Helitrons and LTR elements are in general have lower level of enrichment near genes (see, e.g., Baucom et al. 2009; Yang and Bennetzen 2009). Overall, the above results suggest that, comparing to other types of TEs, grass

SINEs have the highest level of enrichment in gene region. Moreover, this number is also higher than Solanaceae SINEs (30% [Seibt et al. 2016]; results based on five species). Though previous direct evidence of interaction of plant SINEs with cellular gene has been rare, a considerable fraction of grass SINEs enriched in or near genes indicates that it is very likely they do have impact on gene and genomic functions.

Mechanisms Underlying the Enrichment of SINEs in Gene-Rich Regions

Besides high level of enrichment near genes, the other key feature related to enrichment was that SINE families of relatively recent origin have far lower level of gene region enrichment than "old" families (28%:51%). The pattern of SINE enrichment near genes should be the result of balance between insertional specificity and differential removal of elements in different genomic regions (Bennetzen and Wang 2014). As it is difficult to estimate the level of insertional specificity of grass SINEs in general, it is not known if there is difference in average insertional preference between "old" and "young" families. Because this information is not available currently, it is reasonable to use the simplest model, that is, suppose such difference is small and can be neglected. Given similar insertional behavior between "old" and "young" SINE families, one can further assume the insertional bias to genes to be positive, negative or week. In any of the three cases, that the removal favors elements in gene-poor regions is required in order to get distribution patterns that are consistent to the above two key features. Reversely, if differential removal does not bias to gene-poor region (that is, no bias or bias to gene-rich region), no matter what level the insertional specificity is, the distribution of SINEs is unlikely satisfying both key features unless introducing in much more complicated models of SINE amplification and removal. In short, our data suggest that differential removal of elements in gene-poor regions play important role in shaping the genomic distribution of SINEs if "old" and "young" SINE families on average have similar insertional preference.

However, if there have been strong difference of insertional preference between "old" and "young" families, the current observed pattern of enrichment could not be explained by the above simple model. If so, for example, preferential insertion of "old" SINEs in gene-rich-region could be one possible explanation. Further investigations on a phylogenetic tree with denser and more balance taxa sampling would help to address this question. With the rapid advances in genome sequencing techniques, such research will become feasible in very near future.

Grass SINEs show interesting features in the patterns of evolution comparing to the well-studied animal *Alu* family. Previous studies have reported that 1) *Alu* elements are preferentially enriched in regions that are generally gene rich (Lander et al. 2001; Batzer and Deininger 2002); and 2) amplification of *Alu* varies in a lineage-specifc manner (Liu et al. 2009). Our study finds that grass SINEs also show similar patterns (fig. 2*b* and supplementary table S1, Supplementary Material online). As to the pattern of genome retention, it has been proposed that *Alu* evolution is dominated by the accumulation of new insertions; and new *Alu* copies accumulate sequence variation over time and are rarely removed by nonspecifc deletion processes (Shen et al. 1991; Deininger et al. 1992; Deininger 2011). As a results of accumulation, *Alu* accounts for a considerable fraction of their host genomes. In grass, however, SINEs only account for very small proportion of host genomes and confirmed "old" elements have rarely been reported. It is known that plant genomes evolve in a more dynamic manner than vertebrate genomes, in terms of genome-wide duplications, genome size variations, genome structure, TE amplification pattern, etc. (Murat et al. 2012). For example, studies in many grass species have found that the majority of plant dominant TE, LTR elements, are inserted in their host genomes recently and old copies have been removed (see, e.g., Ma et al. 2004; Piegu et al. 2006; Vitte and Bennetzen 2006; Murat et al. 2012). Therefore it is possible that, like LTR elements, old SINE copies have been removed from the host genomes. The forces undying fast change of plant genomes may also be responsible for, at least partly, the removal process.

Supplementary Material

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

Acknowledgment

The authors thank the National Basic Research Program of China (973 Project Grant No. 2013CB34100) for supports.

Literature Cited

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. Nat Rev Genet. 3(5):370–379.

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. 2009. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. 19(2):243–254.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 65:505–530.

Boeke JD. 1997. LINEs and Alus: the polyA connection. Nat Genet. 16:6–7.

Deininger P. 2011. Alu elements: know the SINEs. Genome Biol. 12(12):236.

Deininger PL, Batzer MA, Hutchison CA 3rd, Edgell MH. 1992. Master genes in mammalian repetitive DNA amplification. Trends Genet. 8(9):307–311.

Deragon JM, et al. 1994. An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. J Mol Evol. 39(4):378–386.

Deragon JM, Zhang X. 2006. Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. Syst Biol. 55(6):949–956.

Devos KM. 2010. Grass genome organization and evolution. Curr Opin Plant Biol. 13(2):139–145.

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 12(7):1075–1079.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 35(1):41–48.

Diao X, Freeling M, Lisch D. 2006. Horizontal transfer of a plant transposon. PLoS Biol. 4(1):e5.

El Baidouri M, et al. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. 24(5):831–838.

Fawcett JA, Innan H. 2016. High similarity between distantly related species of a plant SINE family is consistent with a scenario of vertical transmission without horizontal transfers. Mol Biol Evol. 33(10):2593–2604.

Fawcett JA, Kawahara T, Watanabe H, Yasui Y. 2006. A SINE family widely distributed in the plant kingdom and its evolutionary history. Plant Mol Biol. 61(3):505–514.

Goubely C, Arnaud P, Tatout C, Heslop-Harrison JS, Deragon JM. 1999. S1 SINE retroposons are methylated at symmetrical and non-symmetrical positions in *Brassica napus*: identification of a preferred target site for asymmetrical methylation. Plant Mol Biol. 39(2):243–255.

Han Y, Qin S, Wessler SR. 2013. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics 14:71.

Houck CM, Rinehart FP, Schmid CW. 1979. A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol. 132(3):289–306.

Kajikawa M, Okada N. 2002. LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell 111(3):433–444.

Kimura M. 1986. DNA and the neutral theory. Philos Trans R Soc Lond B Biol Sci. 312(1154):343–354.

Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity (Edinb) 107(6):487–495.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.

Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE. 2009. Comparative analysis of Alu repeats in primate genomes. Genome Res. 19(5):876–885.

Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14(5):860–869.

Mao H, Wang H. 2017. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics 33(5):743–745.

Murat F, Van de Peer Y, Salse J. 2012. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. Genome Biol Evol. 4(9):917–928.

Piegu B, 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(10):1262–1269.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16(6):276–277.

Schmid CW, Deininger PL. 1975. Sequence organization of the human genome. Cell 6(3):345–358.

Schwichtenberg K, et al. 2016. Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species. Plant J. 85(2):229–244.

Seibt KM, Wenke T, Muders K, Truberg B, Schmidt T. 2016. Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. Plant J. 86(3):268–285.

Shen MR, Batzer MA, Deininger PL. 1991. Evolution of the master Alu gene(s). J Mol Evol. 33(4):311–320.

Tsuchimoto S, Hirao Y, Ohtsubo E, Ohtsubo H. 2008. New SINE families from rice, OsSN, with poly(A) at the 3′ ends. Genes Genet Syst. 83(3):227–236.

Umeda M, Ohtsubo H, Ohtsubo E. 1991. Diversification of the rice Waxy gene by insertion of mobile DNA elements into introns. Jpn J Genet. 66(5):569–586.

Vassetzky NS, Kramerov DA. 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. 41(Database issue):D83–D89.

Vitte C, Bennetzen JL. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci U S A. 103(47):17638–17643.

Wang H, Devos KM, Bennetzen JL. 2014. Recurrent loss of specific introns during angiosperm evolution. PLoS Genet. 10(12):e1004843.

Wenke T, et al. 2011. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. Plant Cell 23(9):3117–3128.

Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973–982.

Yagi E, Akita T, Kawahara T. 2011. A novel Au SINE sequence found in a gymnosperm. Genes Genet Syst. 86(1):19–25.

Yang L, Bennetzen JL. 2009. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A. 106(47):19922–19927.

Yasui Y, Nasuda S, Matsuoka Y, Kawahara T. 2001. The Au family, a novel short interspersed element (SINE) from *Aegilops umbellulata*. Theor Appl Genet. 102(4):463–470.

Yoshioka Y, et al. 1993. Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. Proc Natl Acad Sci U S A. 90(14):6562–6566.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195(4):923–937.

Associate editor: Mar Alba