Comparative Genomic Paleontology across Plant Kingdom Reveals the Dynamics of TE-Driven Genome Evolution

Moaine El Baidouri[1,2] and Olivier Panaud[1,2,*]

[1]Université de Perpignan Via Domitia, Laboratoire Génome et développement des plantes, UMR UPVD/CNRS 5096, 66860 Perpignan, France

[2]CNRS, Laboratoire Génome et Développements des plantes, UMR CNRS/UPVD 5096, 66860 Perpignan, France

*Corresponding author: E-mail: panaud@univ-perp.fr.

Accepted: February 12, 2013

Abstract

Long terminal repeat-retrotransposons (LTR-RTs) are the most abundant class of transposable elements (TEs) in plants. They strongly impact the structure, function, and evolution of their host genome, and, in particular, their role in genome size variation has been clearly established. However, the dynamics of the process through which LTR-RTs have differentially shaped plant genomes is still poorly understood because of a lack of comparative studies. Using a new robust and automated family classification procedure, we exhaustively characterized the LTR-RTs in eight plant genomes for which a high-quality sequence is available (i.e., *Arabidopsis thaliana*, *A. lyrata*, grapevine, soybean, rice, *Brachypodium dystachion*, sorghum, and maize). This allowed us to perform a comparative genome-wide study of the retrotranspositional landscape in these eight plant lineages from both monocots and dicots. We show that retrotransposition has recurrently occurred in all plant genomes investigated, regardless their size, and through bursts, rather than a continuous process. Moreover, in each genome, only one or few LTR-RT families have been active in the recent past, and the difference in genome size among the species studied could thus mostly be accounted for by the extent of the latest transpositional burst(s). Following these bursts, LTR-RTs are efficiently eliminated from their host genomes through recombination and deletion, but we show that the removal rate is not lineage specific. These new findings lead us to propose a new model of TE-driven genome evolution in plants.

Key words: transposable elements, LTR-retrotransposons, transpositional burst, comparative genomics, genome dynamics, deletion, solo-LTR, plants.

Introduction

Transposable elements (TEs) are endogenous mobile DNA elements that are ubiquitous in nearly all living organisms. Beside their ability to move across their host genome, some TEs can increase their copy number while active and therefore densely populate the chromosomes of many eukaryotic species. For instance, 85% of the total DNA of maize is composed of TEs (Schnable et al. 2009). TEs have thus been considered for a long time as "parasitic DNA," but numerous studies have now clearly established their strong biological impact on the structure, function, and evolution of eukaryotic genomes (Jones and Gellert 2004; Kobayashi et al. 2004; Feschotte 2008; Hollister and Gaut 2009).

TEs are divided into two classes (Wicker et al. 2007): class I elements or retrotransposons (RTs) that move via a "copy and paste mechanism" and class II elements that move via a "cut and paste mechanism." Among class I elements, long terminal repeat RTs (LTR-RTs) are the most abundant in plants (Feschotte et al. 2002). Because of their transposition mechanism, as mentioned earlier, LTR-RTs can spread rapidly throughout their host genome leading in some cases to a significant increase of its size in a short evolutionary time scale (Piegu et al. 2006). These genomic amplifications can occur very rapidly through several waves of retrotransposition that involve only one or few TE families (a process referred to as bursts of transposition). In other words, large genomes would result from large, recent bursts of transposition, whereas small ones would be devoid of any trace of such amplification. Several authors have indeed pointed out that the variation of genome size observed between species is largely dependent on their content in LTR-RTs (Vitte and Panaud 2005; Hawkins et al. 2006; Zedek et al. 2010). However, several studies have also shown that LTR-RTs are eliminated efficiently from the genome through various mechanisms involving deletions and recombinations (Ma and Bennetzen 2004). These observations gave rise to a new

model for TE-driven genome evolution that posits that the genome size at a given time results from two counteracting forces that are the retrotransposition (that adds up DNA to the nuclear genome) and the elimination of TE-related sequences through deletions and recombinations. Various authors have discussed how the different mechanisms involved in TE elimination may actually lead to genome size reduction (Rabinowicz 2000; Gregory 2003). The main question that remains unanswered and is the main focus of this article is whether TE elimination rate varies from one lineage to another. If this is the case, then variation in genome size should be accounted for by both the extent of transposition bursts (higher in large genomes) and the efficiency of TE elimination (higher in small genomes). However, this last point is far from being clearly established. The two main mechanisms of TE elimination are the homologous recombination between the two LTRs of a LTR-RT (Shirasu et al. 2000) and the small deletions (Ma et al. 2004). The first leads to the elimination at once of one of the two LTRs together with the internal region of the element leaving a solo-LTR, whereas the second leads to a more gradual elimination of TE-related sequences. Solo-LTRs have been found in all organisms investigated for so far and seem ubiquitous in all eukaryotes. However, it is not clear whether the rate of solo-LTR formation varies significantly from one lineage to another and if this variation could explain the difference in genome size observed among species. For example, in maize, the solo to intact ratio (*S/I*) was previously estimated to be 0.2:1 (SanMiguel et al. 1996), whereas Ma et al. (2004) showed that in rice, the solo-LTRs outnumber the intact elements (1.5:1 *S/I*). This may suggest that in the smaller genome of rice, LTR-RTs are eliminated more efficiently than in that of maize, but we propose to test it on a larger panel of species to validate this hypothesis. Small deletions have been investigated in only few plant lineages (e.g., rice and *Arabidopsis thaliana*, Bennetzen et al. 2005), but like in the case of solo-LTRs, it is still unclear whether there exists a correlation between their rate and the genome size in a given lineage.

To date, a complete genome sequence is available for 31 plant species of both monocots and dicots. However, only few are based on physical map, which ensures the best possible assembly quality and is needed for a correct characterization of repeats in the genome. For this reason, our study focused on eight plant species for which such sequence is available, which include four monocots and four dicots and cover a wide range of genome size. The dicot species are *A. thaliana* (Arabidopsis Genome Initiative 2000), *A. lyrata* (Hu et al. 2011), grapevine (*Vitis vinifera*, Jaillon et al. 2007), and soybean (*Glycine max*, Schmutz et al. 2010), and the monocot species are rice (*Oryza sativa*, International Rice Genome sequencing Project 2005), sorghum (*Sorghum bicolor*, Paterson et al. 2009), *Brachypodium dystachion* (International Brachypodium Initiative 2010), and maize (*Zea mays*, Schnable et al. 2009). Although the genomic sequence of *A. lyrata* is purely based on whole-genome shotgun data, its close relatedness with *A. thaliana* and its small genome size enabled to produce a good quality assembly suitable for our comparative study.

The TE elimination rate can be estimated for each of these eight genomes through an in-depth characterization of LTR-RTs families, with a particular emphasis on the extent of solo-LTRs and truncated (or deleted) copies. This should allow to investigate whether there exists a correlation between elimination rate and genome size or at least a lineage-specific elimination rate. This approach is conceptually straightforward but technically challenging because it requires to use the same method for LTR-RTs characterization for all eight genomes. Few softwares are commonly used in genome sequencing projects for LTR-RTs searches (e.g., LTRharvest, LTR_STRUC, and LTR_FINDER) (McCarthy and McDonald 2003; Xu and Wang 2007; Ellinghaus et al. 2008). These usually combine searches for structural features, such as LTRs and TSDs, with functional features (the presence of *Gag-Pol* domains). Although these three softwares do not use exactly the same method for the identification of LTR-RTs, they all yielded a comprehensive list of complete LTR-RTs from the eight plant genomes concerned in this study. However, the main difficulty in completing a robust characterization of these elements lies in their correct classification into distinct families.

A considerable effort was made over the last years to propose a new TE hierarchical classification system (Wicker et al. 2007) that subdivided them into subclasses, orders, superfamilies, and families. This system has been successfully used for the TE annotation of some genome sequencing projects (Hu et al. 2011). However, the last level of classification (i.e., family) has always been considered as more error prone because of the high diversification rate and rapid evolution of TE families during and after their transposition, especially for those that have undergone several successive transpositional bursts. In such cases, a TE family consists in large "populations" of elements exhibiting highly divergent forms and distinct transpositional histories. A transposition burst can give rise to hundreds of neocopies that accumulate mutations independently. Mutations include deletions, insertions, and other type of structural variations that occur during the transposition cycle and after their integration into the genome. In particular, defective copies (that often results from large deletions) may remain transpositionally active through their transactivation by their autonomous counterpart. For these reasons, the family classification proposed by Wicker et al. (2007), which is only based on the use of homology searches using a reference sequence as a query, followed by a filtering of the search results (i.e., 80% identity over at least 80% of the reference sequence and a minimum homology length of 80 nt), may lack robustness because it could lead to an overestimation of the family number and correlatively to an underestimation of the copy number of each of these families. We, therefore, designed a new automatic and robust family classification procedure of LTR-RTs based on a clustering strategy that aims to identify all

members of a family even if they exhibit large variation in their DNA sequence. Using this new classification strategy on the eight species mentioned earlier, we demonstrate that only few LTR-RTs families per genome are implicated in genome size variations in plants and that there is no lineage-specific rate of TE elimination but rather a high diversity of such rate among families regardless their host genome.

Materials and Method

LTR-RT Prediction

DNA sequences of *Uwum* family were downloaded from the maize TEs project (http://maizetedb.org/~maize/, last accessed April 10, 2013) and those of soybean from SoyTEdb database (Du et al. 2010). The genomic sequences were downloaded from the Phytozome web site (http://www.phytozome.net/, last accessed April 10, 2013). De novo detection of LTR-RTs was performed using the LTRharvest software (http://www.zbh.uni-hamburg.de/?id=206, last accessed April 10, 2013). Default parameters were used except for -xdrop 37 -motif tgca -motifmis 1 -minlenltr 100 -maxlenltr 3000 -mintsd 2. A typical LTR-RT should harbor a TG..CA box, with TG at the 5′-end of 5′-LTR and CA at the 3′-end of 3′-LTR. However, some LTR-RTs families, such as Tos17 in rice, have a TG..GA motif. For this reason, we allowed one mismatch on the canonical feature TG..CA (-motif tgca -motifmis 1). Annotation of internal features of LTR-RTs was done by LTRdigest (http://www.zbh.uni-hamburg.de/?id=207, last accessed April 10, 2013).

Family Classification by Clustering

Clustering was performed using the SILIX software package (http://lbbe.univ-lyon1.fr/SiLiX, last accessed April 10, 2013). This step first requires to perform a nucleotide Basic Local Alignment Search Tool (BLAST) search of LTR sequences16 (5′-LTR or 3′-LTR) generated by LTRharvest/LTRdigest software in an all against all comparison. The following parameters were used: -r 2 (reward for a nucleotide match. Note that this is a default parameter in blast2+ version), -F F (filter query sequence = false), and -m 8 (alignment view options: tabular). The -r 2 options is more adapted for divergent sequences. Using these options will yield longer high score pairing (HSP) with less sequence identity and therefore facilitates the sequence clustering by SILIX software. Two sequences are included in the same family if the HSPs in the BLAST tabular output cover at least 70% of the sequence length with an identity of at least 60% (see http://lbbe.univ-lyon1.fr/Documentation,3012.html, last accessed April 10, 2013 for further documentation). All the family classification data of the eight plant genomes are freely available and can be accessed at the following link: http://gamay.univ-perp.fr/~moaine/, last accessed April 10, 2013.

RTs Insertion Time Estimation

The dating procedure is based on the fact that at the insertion time of an LTR-RT, the two LTRs are strictly identical. The use of these information allows the estimation of the LTR-RT insertion age (SanMiguel et al. 1998): During evolutionary time and because of the absence of selection pressure, the two LTRs randomly accumulate neutral mutations. Using a substitution rate of 1.3×10^{-8} substitutions per site per year (Ma and Bennetzen 2004), the insertion date can be computed for each LTR-RT.

Solo-LTRs and Truncated Elements Detection

A PERL script was written for solo-LTRs detection (available on request). For each LTR-RT family (for which the exact border can be determined), a consensus LTR sequence was manually chosen based on a sample of paralogs. Then a BLAST search was performed to retrieve all the LTRs of each family and verify the presence/absence of the internal region around it followed by manual inspection using DOTTER software (Sonnhammer and Durbin 1995). To evaluate the portion of genomes corresponding to fragmented and nested LTR-RTs that are not detected by structure-based methods (LTRharvest), the genomic sequence of each species was split into small fragments of 1 kb using splitter software from EMBOSS package (http://emboss.bioinformatics.nl/cgi-bin/emboss/help/splitter, last accessed April 10, 2013). Then a nucleotide BLAST search of fragmented genome was performed against the LTR-RTs database to quantify the total size of each families in the genome considering the best hits of each fragment. All the statistical analyses were done using R software (R Development Core Team 2010) (http://www.R-project.org, last accessed April 10, 2013).

Estimation of Internal Deletion

The estimation of removal rate of TEs through deletions required to build for each genome a data set of TE-related sequences with both their insertion date and the amount of deleted sequences compared with their original master copy. This is not straightforward for the following reasons: First, the size of the original master copy that a paralog should be compared with was not trivial to establish for most families (or subfamilies). Figure 6 shows the distribution of the size of all paralogs of the family Gmr19 in soybean. This distribution is unimodal, the mode corresponding to the size of the most frequent paralog found in the genome for the family. One could hypothesize that this mode corresponds to the original master copy. Under this assumption, paralogs with smaller size would correspond to deleted forms, whereas paralogs with larger size should correspond to sequences with nested insertions of other TEs. Deletions were therefore computed with the former, whereas the latter were removed from the data set. Moreover, several families showed a multimodal distribution of their size, indicating the presence of more than

one master copy with distinct size. These families were discarded from the data set. Finally, the major limitation of the analysis lies in the estimation of the insertion date for each paralogous sequence. The older the element, the more deletions it will accumulate. This often leads to truncated elements lacking part or totality of one LTR. In such case, the estimation of insertion date is not possible through the comparison of the divergence between both LTRs. To circumvent this problem, one could first build a phenetic tree of all paralogous sequences and compute the average pairwise distance among all paralogs that belong to the same cluster. This method is, however, far too time consuming because it requires a lot of manual checking during the alignment step. For all these reasons, we chose to build our data set solely on elements harboring both their LTRs. One could argue that this lead to a bias in our analysis, because only young insertions would thus be analyzed. However, the results shown in figures 5 and 6 clearly show that our data set allows to cover a time range of at least 2 Myr, which is enough to measure the deletion rate in all six genomes.

Phylogeny Reconstruction

LTR sequences of the two families *Uwum* and *RLC_Gmr6/18* were aligned using MAFFT software (Katoh et al. 2002). Highly variable regions were deleted, and alignments were modified by hand using Seaview (Galtier et al. 1996; http://pbil.univ-lyon1.fr/software/seaview.html, last accessed April 10, 2013). For each family, a Neighbor-Joining phenetic tree was drawn performing 1,000 bootstrap replicates (ClustalX software; Thompson et al. 1997). Finally, a circular tree was drawn using the Treedyn package (Chevenet et al. 2006; http://www.treedyn.org/, last accessed April 10, 2013).

Results and Discussion

A New Definition of LTR-RTs Family

As mentioned in the Introduction section, the definition of a LTR-RT family is not straightforward because of the complex evolutionary dynamics of such sequences. To establish a robust automated method for the complete family classification of LTR-RTs, we first selected two highly repeated LTR-RT families from maize and soybean that we characterized manually to establish a list of relevant parameters for subsequent analyses of other families in all the eight plant genomes retained in this study. These two families are *Uwum* from maize and *RLC_Gmr6/18* from soybean. Figure 1 shows a schematic representation and comparison of the structure of the different members of these two families together with a phylogenetic tree based on multiple alignment of the LTRs sequences. *Uwum* family is composed of three different populations or subfamilies (fig. 1*A*1 and *A*2). The first subfamily (subF1) (195 copies) (referenced in the maize TEs project http://maizetedb.org/~maize/ (last accessed April 10, 2013) as *RLG_uwum_AC190887-2701*) harbors the structural features of LTR-RTs and contains the *Gag-Pol* polyprotein gene that encodes the enzymes that are necessary for the retrotransposition cycle. The other two subfamilies (subF2 and subF3) (*RLG_uwum_AC213069-12092* and *RLG_uwum_AC177933-415*, respectively) have LTRs that are homologous to that of subF1, but their internal region does not present any homology with the *Gag-Pol* domains. However, both these subfamilies appear to have transposed recently because they are present in 282 and 400 copies in the maize genome for subF2 and subF3, respectively. SubF2 and subF3 do not encode any ORFs and belong to a group of nonautonomous RTs that are probably mobilized in trans using the subF1 retrotranspositional machinery. In this regard, subF2 and subF3 can be considered as LArge Retrotransposons Derivatives (i.e., LARDs; Kalendar et al. 2004) of subF1 and should be classified into one family, which would not be possible using the 80/80/80 rule proposed by Wicker et al. (2007). The soybean *RLC_Gmr6* and *RLC_Gmr18* elements exhibit homology in their LTRs. However, like in the case of the *Uwum* family described earlier, their internal region does not show any homology except for three distinct regions (fig. 1*B*). *RLC_Gmr18* does not encode any ORFs in its internal region, in contrast with *RLC_Gmr6* that harbors both the integrase and the reverse transcriptase domains. Like in the case of *Uwum*, a good classification procedure should place both *RLC_Gmr18* and *RLC_Gmr6* elements into the same family despite the lack of homology of their internal region. These two examples illustrate the difficulty to correctly classify LTR-RTs families. Even if the LTRs evolve faster than the coding internal region because of lower functional constraints, they "offer the most specificity in defining families" (Wicker et al. 2007). Thus, similar to these authors, we believe that it is more accurate to define LTRs-RTs family first based on the LTR similarity, then eventually define different subfamilies taking in account the full-length sequence. This second step (i.e., the classification into subfamilies) is essential to pertain the information regarding the evolutionary relationships between related groups. Such information, which is shown later, is needed to unravel the mechanisms through which LTR-RTs shape plant genomes. The thorough analysis of both the maize *Uwum* and the soybean *RLC_Gmr6/18* families enabled us to define threshold parameter values for the definition of both families and subfamilies: Two LTR-RTs belong to the same family if they share at least 60% of identity over 70% of their LTR length (nondefault Blastn parameters under relaxed settings; see Materials and Methods), whereas they belong to the same subfamily if they share at least 60% of identity over 70% of the full-length sequence. Using these new criteria, we defined a new automatic procedure for the classification of LTR-RTs. The workflow of our method is as follows:

1. The first step consists in the mining out of all LTR-RTs in a given genome using LTRharvest (structure-based methods;

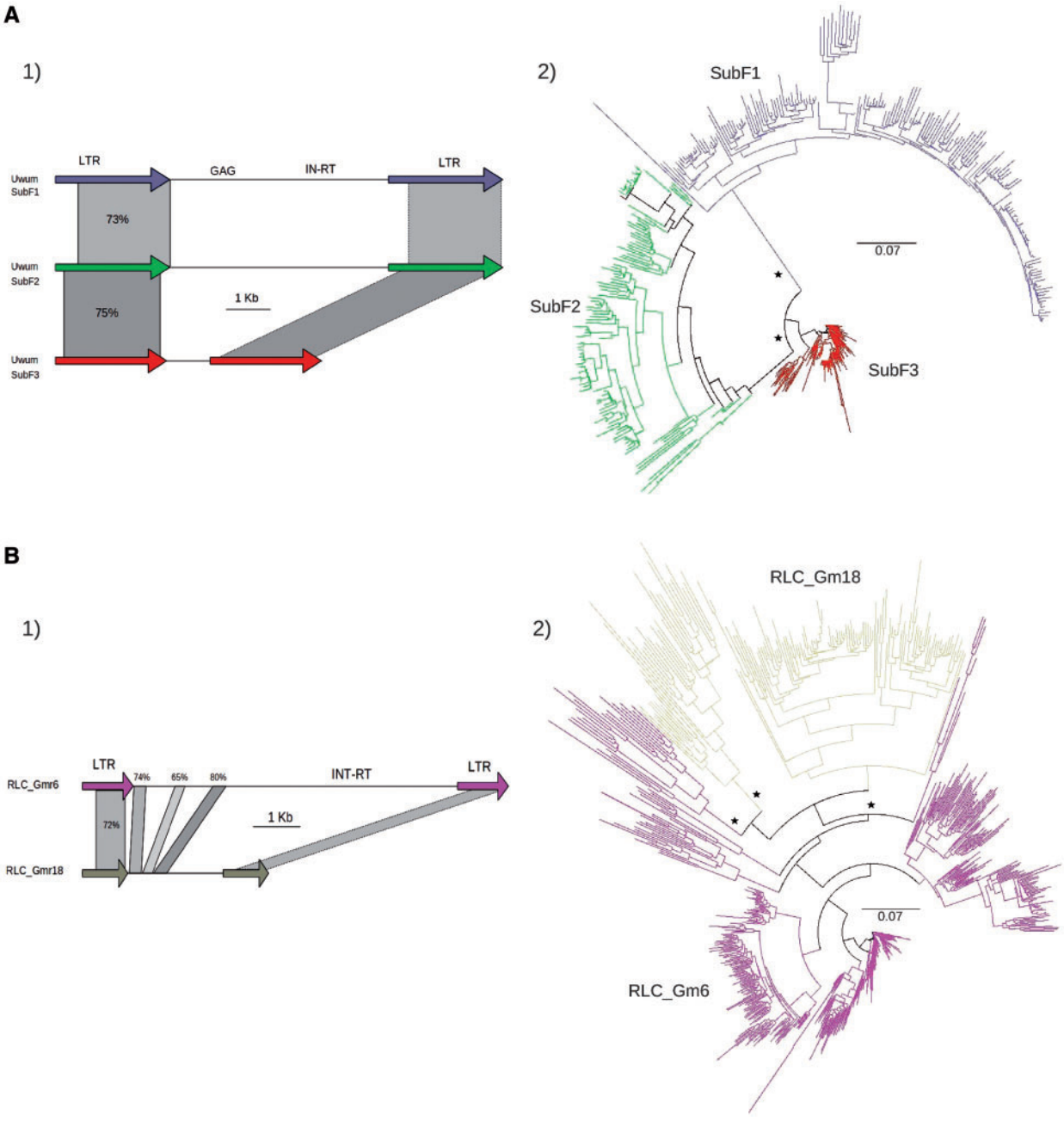

Fig. 1.—(*A*2 and *B*2) Phylogenetic tree and (*A*1 and *B*1) schematic representation and comparison of different subfamilies of LTR-RTs *Uwum* from maize (*A*) and *RLC_Gm6* and *RLC_Gm18* from soybean (*B*). The neighbor-joining tree was constructed based on the alignment of the most conserved part of the LTRs sequences. The asterisks indicate the branch with a bootstrap value higher than 90. Color coding: different color indicates different subfamilies of the same LTR-RT family. Gray areas represent conserved region between subfamilies. Scale bar indicates nucleotide sequence divergence. SubF, subfamily; GAG, group-specific antigens; IN, intergrase; RT, reverse transcriptase; kb, kilobase.

Ellinghaus et al. 2008) and annotate each element using LTR digest (Steinbiss et al. 2009).

2. The second step is the clustering of the whole-LTR sequences based on their similarity (see Materials and Methods) using SILIX software (Miele et al. 2011) to define families as described earlier. SILIX software can cluster two or more DNA sequences taking into account the alignment coverage constraints and their sequence identity (see Materials and Methods). In each genome, the LTRs that belong to different families are sufficiently divergent to be considered as distinct "clusters."
3. Finally, a clustering of the different LTR-RTs within the same family using the full-length sequence to define subfamilies. This step can lead to different cases: 1) only

one subfamily is detected, which is often the case when a LTR-RT family underwent a single recent transposition burst; 2) two or more subfamilies are detected, each one of them containing several copies; and 3) several subfamilies are defined but only one has multiple copies, whereas the remaining are single copy. These "singleton clusters" often result from an error during the mining process using LTRharvest, for instance, the detection of two closely inserted solo-LTRs. These "singleton clusters" are eliminated from the data set. In addition, all families containing less than four members are eliminated if their internal region does not harbor at least one functional domain of the *Gag-Pol* polyprotein gene.

Using this new classification method, we investigated the evolutionary dynamics of LTR-RTs in eight plant species using the same criteria for family definition. We performed a complete characterization and classification of all "intacts copies" (elements with two LTRs). A total of 111,628 elements belonging to 7,412 different families were identified (table 1). The total copy number LTR-RTs in these eight genomes is as follows: 68,462 for maize, 13,038 for soybean, 17,022 for sorghum, 4,672 for grapevine, 3,663 for rice, 2,162 for *Brachypodium*, 2,134 for *A. lyrata,* and 475 for *A. thaliana*. The maize *Uwum* elements were clustered in a single family and in three distinct subfamilies as expected. Similarly, *RLC_Gmr6* and *RLC_Gmr18* elements in soybean were clustered into one family as expected but clearly separated into two different subfamilies. Moreover, all the previously identified families in other genomes were correctly identified using our procedure (data not shown). In particular, the three known rice LARD families, that is, *Dasheng*, *Spip,* and *Squiq* were correctly classified as subfamilies within the same family as their autonomous counterpart *RIRE2*, *RIRE3,* and *RIRE8*, respectively.

The Activity of Few Families Contribute to Genome Size Variations in Plants

We then first examined the distribution of the copy number for each family in the eight species that we analyzed. Surprisingly, these distributions exhibit a similar L-shaped pattern for all species (fig. 2). One could hypothesize that large genomes, similar to that of maize, should harbor a high number of repeated families, whereas smaller ones, similar to that of *Arabidopsis*, would harbor a smaller number. This is not the case. For each genome, regardless their genome size, only few LTRs-RT families are repeated, whereas the majority are single- or low copy. The main difference lies in the copy number of the largest LTR-RT families: If the 10 most highly repeated families are considered, the maize genome harbors a total of 52,357 elements that represent 524 Mbp of genomic DNA (1,022 Mbp if we consider deleted, nested, and solo-LTRs; see supplementary data S1, Supplementary Material online), whereas that of *A. thaliana* harbors 194 elements that represent 1.8 Mbp (3.85 Mbp including fragmented and noncomplete elements; supplementary data S1, Supplementary Material online) of genomic DNA. We tested the correlation between genome size and the total size of the 10 most repeated families in each genomes including all "intact copy" and fragmented, nested, and solo-LTRs (see Materials and Methods). As expected, a strong positive correlation (Pearson's $\rho = 0.98$, P value $= 1.31e^{-05}$) is obtained, demonstrating the clear correlation between genome size and the extent of retrotranspositional activity of the few most highly repeated families.

These observations clearly show that retrotransposition occurs in all plant genomes regardless their size but that only few families have undergone transpositional bursts in a recent past. This is in accordance with the recent discoveries made in the epigenetic control of transposition. Most TEs are strictly controlled by several distinct epigenetic pathways involving either transcriptional gene silencing or post-transcriptional gene silencing (Lisch and Slotkin 2011). These multiple pathways are similar to the multiple layers of a "mille feuilles" (Rigal and Matthieu 2011), and their accumulation ensures the efficiency of the control of transposition. However, unexpectedly, the impediment of one of these pathways leads to the transpositional reactivation of one or very

Table 1
Copy Numbers of Intact LTR-RTs Families

Species	Genome Size (Mb)	Total No. of Intact Copies	No. of Families	No. of Copies of 10 Most Repeated Families	*S/I*
Maize	2,500	68,462	2,873	52,357	0.14:1
Soybean	975	13,038	1,144	8,484	1.18:1
Sorghum	697	17,022	1,077	10,758	0.49:1
Grapevine	430	4,672	821	2,348	0.84:1
Rice	382	3,663	340	1,823	1.39:1
Brachypodium	272	2,162	466	1,104	1.29:1
Arabidopsis lyrata	207	2,134	481	986	0.23:1
Arabidopsis thaliana	135	475	210	194	0.7:1
Total		111,628	7,412	78,054	

NOTE.—Solo-LTR to intact elements ratio (*S/I*) in the eight plant genomes was calculated based on the most repeated families (for details see supplementary data S2, Supplementary Material online).

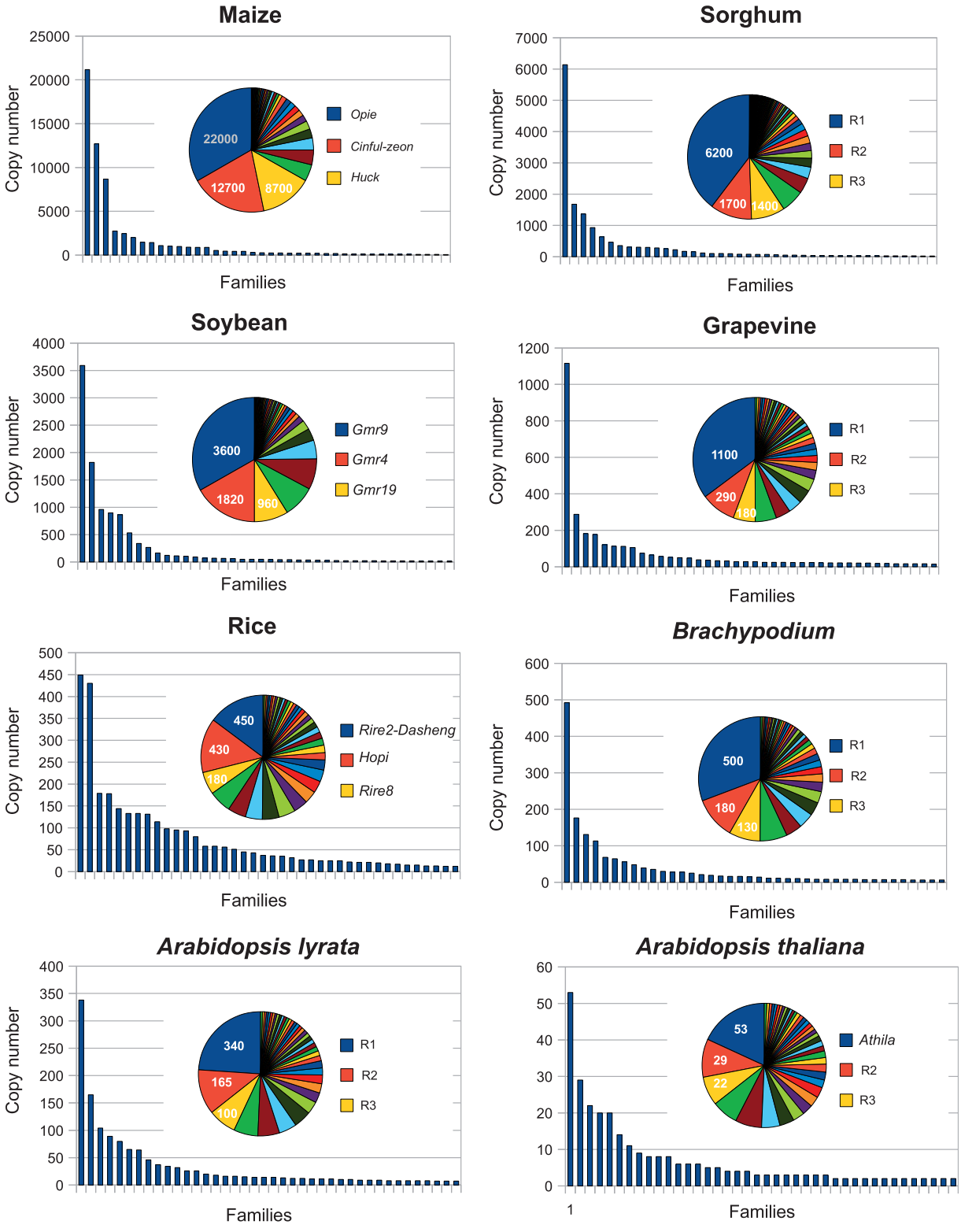

Fig. 2.—Distribution of copy number per LTR-RTs families in eight plant genomes. *x* axis represent different families and *y* axis the copy number per family. For each species, a pie chart represents the proportion of all LTR-RT families in the genome. The copy number of the three most repeated families is given in each corresponding pie chart. Only 40 families are represented for each species.

few TE families among the many that populate plant genomes. For example, only the LTR-RT *Evade* is activated in *met1* mutants in *A. thaliana*, whereas only the LTR-RT *Onsen* is activated in siRNAs mutants and heat stress in the same species (Mirouze et al. 2009; Ito et al. 2011). This, together with our results, suggests that the activity of only few families at a time in the recent evolutionary past could result from the temporary relaxation of only one or few silencing pathways. Whether such relaxation is caused by genetic impediment or environmental stimuli (e.g., physiological stresses) remains to be elucidated.

LTR-RTs: Recent Activity and Rapid Elimination

The insertion date of each of the 111,628 LTR-RTs that were mined out of the eight plant genomes was calculated using previously described methods (SanMiguel et al. 1998; see Materials and Methods). Figure 3 shows the comparative age distribution of these elements for each species. Whatever the species considered, these distributions clearly show that the majority of the elements transposed recently, with almost no insertion older than 3 Myr. This cannot be the result of an experimental bias, because the threshold that we used for LTR similarity prediction is approximately 85% between the two LTRs. This value corresponds to 5.7 Myr, which is far older than the oldest elements identified in our study. In the case of the large genome of maize, our data set may be slightly biased toward younger elements because of a tendency of the LTR-RTs to build nested insertions (SanMiguel et al. 1998). However, for the seven others, nested insertions are not as frequent and can therefore not be the cause of a bias in the data mining process. The distributions also show that the transposition bursts that occurred in this time range are not strictly concomitant among species. For example, the maximum retrotranspositional activity in rice appears to be younger than in *Brachypodium*. In addition, several waves of transpositional bursts can be observed in soybean (fig. 3). These results thus suggest that the increase in genome size caused by retrotransposition does not depend on the age of the latest retrotranspositional bursts but rather from the accumulation of all the bursts that occurred within the last 3 Myr. Obviously, this raises the question of the elimination of TE-related sequences from plant genomes, which has been subjected to debates among evolutionary genomicists over the past 10 years: The paucity of LTR-RTs older than 3 Myr in all the genomes that we investigated can only be explained by their quick elimination from their host genomes. The interesting point raised by our comparative study is that this elimination process does not appear to be genome size dependent because the distribution curves of all eight species reach their asymptote (value of 0) between 3 and 4 Myr (fig. 3). We nevertheless further analyzed the LTR-RTs elimination rate among the eight plant species to confirm whether the LTR-RTs elimination, either through ectopic recombination (formation of solo-LTRs) or through deletions, is not lineage specific.

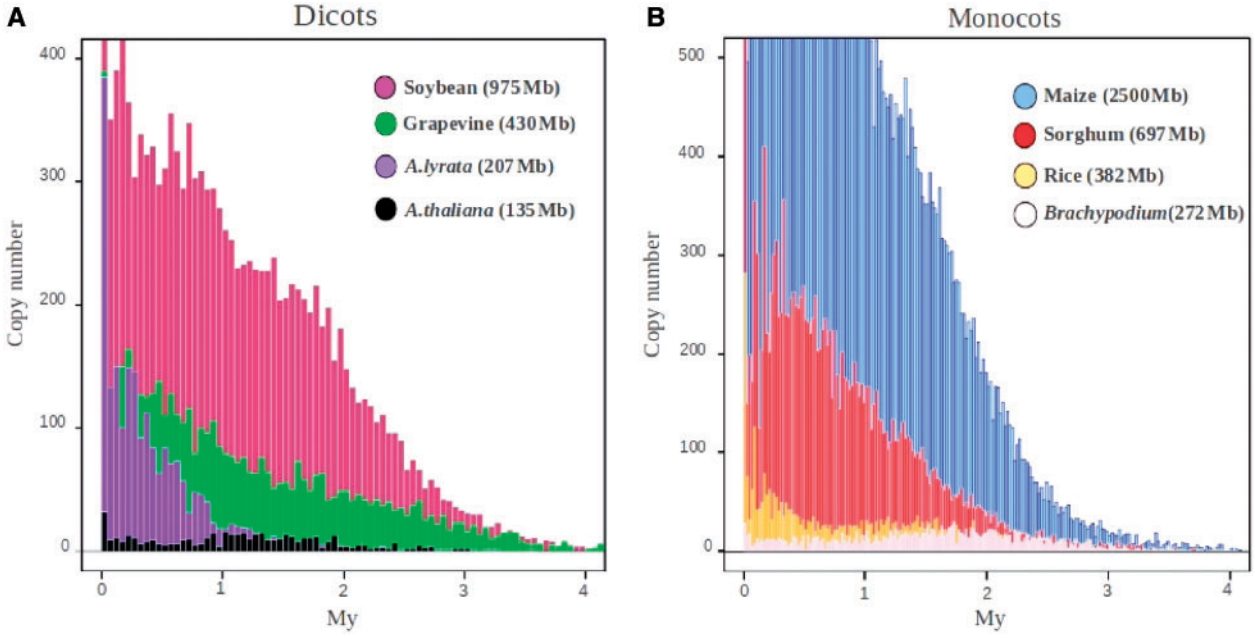

Fig. 3.—Distribution of estimated insertion ages (in Myr) of LTR-RTs in maize, sorghum, rice, *Brachypodium* for monocots (*A*) and soybean, grapevine, *Arabidopsis Lyrata,* and *A. thaliana* for dicots (*B*). Note that for monocots, the number of copies represented is limited to 500 to facilitate interspecies comparison.

Solo-LTR Formation Is Function of Both LTR and Internal Region Length but Is Not Lineage Specific

Solo-LTR formation through ectopic recombination between the LTRs of the same element or distant elements is considered as the prevalent mechanism for LTR-RT elimination in plants (Vitte et al. 2007). Hence, we estimated the ratio of solo-LTR to intact elements (*S/I*) for each LTR-RT family in these eight genome species (see Materials and Methods). As listed in table 1, this ratio varies greatly from one species to another. For instance, it is 0.14:1 in maize and significantly higher in rice (i.e., 1.39:1). These values are in accordance with previously published ones (SanMiguel et al. 1998; Ma et al. 2004) and may indeed suggest that a lower elimination rate through ectopic recombination in maize could explain its larger genome size. However, this is not correct for two reasons: first, our results for other species show that there is no correlation between the overall *S/I* and genome size, for example, 1.29:1 for *Brachypodium*, 0.23:1 for *A. lyrata,* and 1.18:1 for soybean (Spearman's $rs = -0.19$, P value $= 0.67$). Second, a more in-depth analysis of *S/I* for the most highly repeated families of each of the eight genomes clearly shows that it varies greatly among families of the same species (supplementary data S2, Supplementary Material online). Previous reports have demonstrated the positive correlation between the LTR length and solo-LTR formation rate in different species (Du et al. 2010). We calculated the *S/I* of 343 different families of LTR-RTs belonging to the eight species and confirmed that there exists a positive correlation between the LTR length and *S/I* (Spearman's $rs = 0.3$, P value $= 5.42e^{-09}$). Interestingly, we found that the (LTR length /internal region size) ratio is more strongly correlated with *S/I* (Spearman's $rs = 0.43$, P value $= 2.58e^{-17}$) than LTR length suggesting that the distance between the two LTR may also impact solo-LTR formation. This may be explained by the fact that ectopic recombination can be favored by longer LTR, because of more stable pairing of repeats but only if they are not too distant. We can therefore conclude from this analysis that there is no lineage-specific elimination rate of LTR-RTs through ectopic recombination between LTRs. The differences in the overall *S/I* observed among species may thus simply originate from the structural characteristics of the LTR-RT families that have undergone the largest transpositional bursts in the most recent past.

The Removal Rate of LTR-RTs through Deletions Is Not Lineage Dependent

The removal rate was estimated based on a data set of LTR-RT that harbors both their LTR to date their insertion (see Materials and Methods for technical discussion about removal rate estimation). *Arabidopsis thaliana* and *Brachypodium* were not included in this analysis because their genome does not harbor enough paralogs even for the most repeated families to conduct relevant statistical tests. For each family of the remaining six genomes, we separated the elements into four classes according to their insertion date: 0–0.5 Myr, 0.5–1 Myr, 1–1.5 Myr, and >1.5 Myr. Figure 4 shows the histograms of the average of DNA loss for these four classes in the six genomes. The first observation is that in most cases, as expected, DNA loss is higher for older elements. However, there are some exceptions for maize, grapevine, and *A. lyrata*. In maize, deletions appear to decrease with time (during the first 1.5 Myr). For the latter two species, the fourth class (elements older than 1.5 Myr) exhibit a lower deletion rate than the third one. Figure 4 also shows that there is a substantial variation in removal rate among the six genomes. This was confirmed using a one-way analysis of variance (ANOVA) (P value < 0.001). However, no correlation could be established between genome size and removal rate: For instance, even if the maize removal rate appears to be lower than that of rice (except for the 0–0.5 Myr class) as one would expect if there exists a negative correlation between removal rate and genome size, the sorghum removal rate is lower than that of soybean, although it has a smaller genome size. We further investigated each of the six genomes by comparing the removal rate of five randomly chosen LTR-RT families. The results are shown in figure 5. Surprisingly, each genome exhibits a large variation of removal rate among the various families that it harbors (for all the six genomes, the one-way ANOVA test is highly significant with a P value < 0.001). This leads us to conclude that, like in the case of the rate of solo-LTR formation, the extent of variation of small deletions rate among families within genomes is so large that it makes the comparison between genomes and even between time range within genomes irrelevant. Although it should be pointed out that our data set is biased toward young and complete elements, our study suggests that the LTR-RT removal rate through small deletions is family, rather than lineage specific. The cause of such variation remains unclear and will necessitate more comparative analyses between the families with high deletion and those with low deletion rates. In addition, Ma et al. (2004) showed that small deletions are always flanked by small direct repeats. Whether families exhibiting high deletion rate harbor more such direct repeats remains to be investigated.

Conclusion

Unlike transposons, RTs never get excised once inserted into the chromosomes. They can therefore be exploited as "genomic fossils" to unravel the evolutionary history of genomes. Here, we focused on the recent history of eight plant genomes that contrast in size and belong to independent lineages and thus tried to understand how LTR-RT could contribute to such structural diversity. This, however, necessitated to first conduct some conceptual and technical developments that lead us to propose a new automated and robust classification procedure of LTR-RT families that we applied on the eight species. An exhaustive survey of both the content in LTR-RTs and their

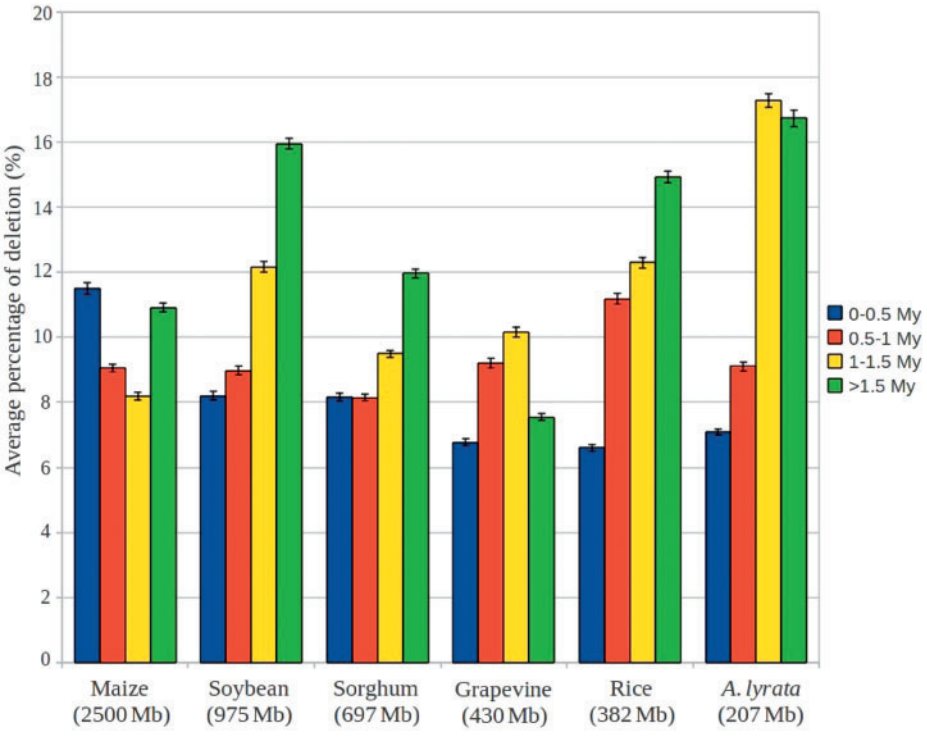

Fig. 4.—Comparison of average percentage of deletion of LTR-RTs between different plant species for four different time range (Myr). Different colors represent different age class. Brackets represent standard deviation of each age class.

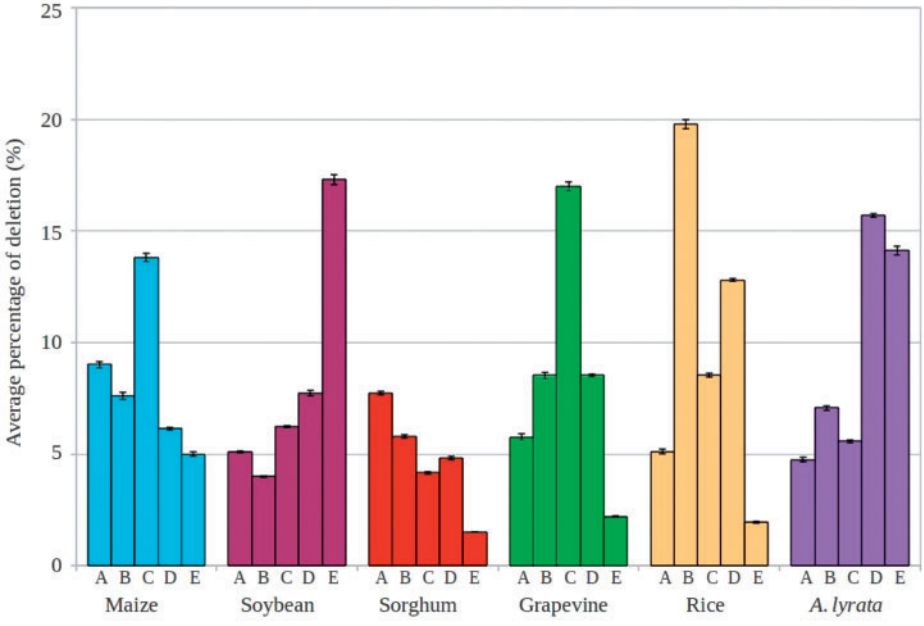

Fig. 5.—Comparative deletion pattern of LTR-RTs families within different plant species (0.5–1 Myr). The horizontal line shows the average percentage of deletion of different families belonging to six plant species (see supplementary data S3, Supplementary Material online).

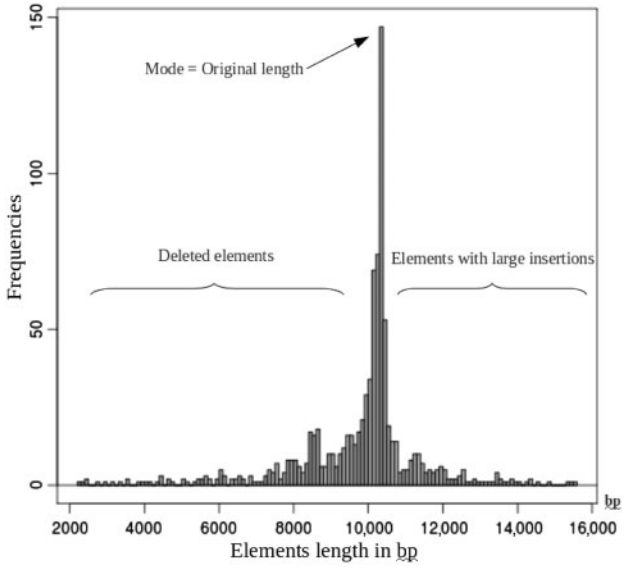

Fig. 6.—Distribution of copies length within one LTR-RT family (*Gmr19* from soybean genome) (see Materials and Methods).

transpositional history across these genomes then revealed that 1) retrotransposition occurs in all plant genomes regardless their size, but only few families have been active in a recent past. 2) Difference in genome size can be accounted for by the extent of the transpositional activity of these few families (e.g., 52,000 copies for the most highly repeated family in the 2,500-Mbp genome of maize compared with 194 copies of the most highly repeated family in the 110-Mbp genome of *A. thaliana*). 3) LTR-RTs are quickly eliminated from all genomes through ectopic recombination between LTRs and/or deletions (like it was shown by others), but the LTR-RT elimination rate is not lineage dependent although it varies significantly among families regardless the size of the genome. From these observations, we can tentatively draw a new and more complete model of TE-driven genome evolution than the one we proposed earlier (Vitte and Panaud 2005). This model posits that retrotransposition occurs recurrently in most (if not all) plant genomes, through the transposition burst of one (or very few) families at a time. The extent of such bursts varies greatly (from hundred to thousands) from one lineage to another. Following these short periods of transpositional activity, TE-related sequences are efficiently eliminated from the genomes at a rate which is not lineage specific.

One immediate consequence of this model is the stochastic nature of most TE-driven structural variations in plant genomes: Plant genome size at a given time and in a given lineage should be the result of a combination of several factors concerning the latest retrotranspositional burst, for instance, and not exhaustively its extent (high in the case of large genomes) or the LTR size of the family involved in the burst (LTR-RTs harboring large LTRs are eliminated faster through ectopic recombinations) together with the distance separating the two LTRs of a given element. Nevertheless, one has to keep in mind that the amplification process of LTR-RTs is also the result of complex interactions between the elements and their hosts. Future investigations on the epigenetic control of TEs by their hosts and the different strategies employed by TEs to escape such silencing pathways may help to understand their evolutionary success during plant genome evolution.

Supplementary Material

Supplementary data S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org).

Acknowledgments

M.E. performed programming, conducted data analyses, and wrote the manuscript. O.P. designed the research, supervised this work, participated to data analyses, and finalized the manuscript. The authors thank Cristian Chaparro and Vincent Miele for their technical help and Scott Jackson and Marie Mirouze for their critical comments on the manuscript. This work was supported by a CNRS/Région Languedoc Roussillon research grant and by the University of Perpignan Via Domitia.

Literature Cited

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant opsis thaliana. Nature 408:796–815.
Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 95:127–132.
Chevenet F, Brun C, Banuls AL, Jacq B, Christen R. 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics 7:439.
Du J, et al. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposon in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J. 63:584–598.
Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.
Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 9:397–405.
Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 3:329–341.
Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci. 12:543–548.
Gregory TR. 2003. Is small indel bias a determinant of genome size? Trends Genet. 19:485–488.
Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 16: 1252–1261.
Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–1428.
Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43:476–481.

International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463: 763–768.
International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature 436:793–800.
Ito H, et al. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature 472:115–119.
Jaillon O, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467.
Jones JM, Gellert M. 2004. The taming of a transposon: V(D)J recombination and the immune system. Immunol Rev. 200:233–248.
Kalendar R, et al. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166:1437–1450.
Katoh K, Misasa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. Science 304:982.
Lisch D, Slotkin RK. 2011. Strategies for silencing and escape: the ancient struggle between transposable elements and their hosts. Int Rev Cell Mol Biol. 292:119–52.
Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A. 101:12404–12410.
Ma J, Devos KM, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 14:860–869.
McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR-retrotransposons. Bioinformatics 19: 362–367.
Miele V, Penel S, Duret L. 2001. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.
Mirouze M, et al. 2009. Selective epigenetic control of retrotransposition in *Arabidopsis*. Nature 461:427–430.
Paterson AH, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556.
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. 16:1262–1269.
R Development Core Team. 2010. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rabinowicz PD. 2000. Are obese plant genomes on a diet? Genome Res. 10:893–894.
Rigal M, Mathieu O. 2011. A "mille-feuille" of silencing: epigenetic control of transposable elements. Biochim Biophys Acta. 1809: 452–458.
SanMiguel P, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768.
SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nat Genet. 20: 43–45.
Schmutz J, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.
Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. 10:908–915.
Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:1–10.
Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 37:7002–7013.
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.
Vitte C, Panaud O. 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res. 110:91–107.
Vitte C, Panaud O, Quesneville H. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8:218.
Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35:W265–W268.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Zedek F, Smerda J, Smarda P, Bures P. 2010. Correlated evolution of LTR retrotransposons and genome size in the genus *Eleocharis*. BMC Plant Biol. 10:265.

Associate editor: Ellen Pritham