Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution

Clémentine Vitte and Jeffrey L. Bennetzen*

Department of Genetics, University of Georgia, Athens, GA 30602-7223

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved September 21, 2006 (received for review July 5, 2006)

Analysis of LTR retrotransposon structures in five diploid angiosperm genomes uncovered very different relative levels of different types of genomic diversity. All species exhibited recent LTR retrotransposon mobility and also high rates of DNA removal by unequal homologous recombination and illegitimate recombination. The larger plant genomes contained many LTR retrotransposon families with >10,000 copies per haploid genome, whereas the smaller genomes contained few or no LTR retrotransposon families with >1,000 copies, suggesting that this differential potential for retroelement amplification is a primary factor in angiosperm genome size variation. The average ratios of transition to transversion mutations (Ts/Tv) in diverging LTRs were >1.5 for each species studied, suggesting that these elements are mostly 5-methylated at cytosines in an epigenetically silenced state. However, the diploid wheat *Triticum monococcum* and barley have unusually low Ts/Tv values (respectively, 1.9 and 1.6) compared with maize (3.9), medicago (3.6), and lotus (2.5), suggesting that this silencing is less complete in the two Triticeae. Such characteristics as the ratios of point mutations to indels (insertions and deletions) and the relative efficiencies of DNA removal by unequal homologous recombination compared with illegitimate recombination were highly variable between species. These latter variations did not correlate with genome size or phylogenetic relatedness, indicating that they frequently change during the evolutionary descent of plant lineages. In sum, the results indicate that the different sizes, contents, and structures of angiosperm genomes are outcomes of the same suite of mechanistic processes, but acting with different relative efficiencies in different plant lineages.

indel frequency | mutation | point mutation | sequence evolution

Flowering plants (angiosperms) exhibit exceptional levels of variation in nuclear genome size and frequency of genic rearrangement (1, 2). The smallest angiosperm genomes, like *Arabidopsis* (≈125–160 Mb; refs. 3–6), contain 15–20% repetitive DNA that is largely limited to knobs, pericentromeres and other gene-poor heterochromatic regions (3, 6). The mid-size angiosperm genome of maize (≈2,700 Mb) contains >80% repetitive DNA, much of it intermixed with genes (7–9). Very little is known about the larger plant genomes, like that of bread wheat (≈17,000 Mb), although it has been argued that most wheat genes are sequestered in gene-rich islands that are reasonably well separated from the ≈90% of the genome that is repetitive in nature (10–12). In comparisons between fairly closely related species, like sorghum and maize (13), or even between different maize haplotypes (14), the DNA between genes is highly variable. Most of this intergenic variability is caused by the differential insertion of transposable elements in different plant lineages. Although transposable elements of both class I (retroelements) and class II (DNA elements) are abundant and mutationally significant in angiosperms, a specific class I element type (the LTR retrotransposons) comprises the greatest percentage of most flowering plant genomes (8, 15–19).

Although polyploidy and amplification of repetitive DNAs are the major forces behind genome size increase in flowering plants, less is known about the processes for DNA removal. Unequal intrastrand recombination between the two LTRs that terminate LTR retrotransposons often generates solo LTRs (20), with the associated loss of one LTR and the internal sequences of the element. This can attenuate genome growth, but does not fully reverse the growth in genome size created by the original insertion. Unequal recombination between homologous LTR retrotransposons at different genomic locations can cause net deletion or duplication of nuclear DNA between the elements, or genomic rearrangements like inversions and reciprocal translocations, depending on the chromosomal location and orientation of the participating elements (21). In plants, apparent intrastrand recombination events that remove the chromosomal DNA between two original LTR retrotransposons have been inferred by the absence of target site duplications (TSDs) (22, 23), but the precise nature of the deleted intervening sequences has not been determined.

Comparative sequence analysis of LTR retrotransposons in *Arabidopsis* and rice has shown that illegitimate recombination is associated with a high frequency of genomic DNA loss by the accumulation of small deletions (22, 23). The primary mechanism(s) of illegitimate recombination has not been defined, although both repair of double-strand breaks and slipped-strand replication have been proposed (22–25). Illegitimate recombination also appears to be the major process for DNA removal in animals, as indicated by studies in insects and mammals (26–28). It is clear that the slow and steady process of DNA removal by illegitimate recombination works on all sequences in plant genomes (29), and that it can slow or reverse overall genome growth.

Very little is known, however, about the reasons for different genome sizes, genome compositions and genic arrangements in plants. Why does the *Arabidopsis* genome contain only ≈12 Mb of LTR retrotransposons (≈9% of the genome) (6), whereas the maize genome contains >1,800 Mb of these elements (>70% of the genome) (9)? One answer for this may be different rates of DNA removal in different taxa. For instance, different rates of LTR retrotransposon removal by unequal recombination is suggested by the higher ratio of solo LTRs to intact elements in rice compared with *Arabidopsis* (23). In insects, the small *Drosophila melanogaster* genome (≈175 Mb) removes nuclear DNA by illegitimate recombination >40 times faster than in the ≈11-fold larger genome of *Laupala* crickets (27). Similarly, the relative frequency and size of deletions associated with double strand break repair have been found to be greater in *Arabidopsis* than in tobacco's ≈5,100 Mb genome (25). Alternatively, some lineages of angiosperms may have had more active transposon amplification in recent times, as suggested by the great number and diversity of LTR retrotrans-

Author contributions: C.V. and J.L.B. designed research; C.V. performed research; C.V. and J.L.B. analyzed data; and C.V. and J.L.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviation: TSD, target site duplication.

*To whom correspondence should be addressed. E-mail: maize@uga.edu.

Table 1. Species investigated

Species	Phylogeny	Ploidy level	Genome size (1C, Mb)*	Genome sample analyzed		
				No. of regions	No. of contigs	Total Mb
Arabidopsis thaliana	Eudicot (Brassicaceae)	2x	160	NA	NA	NA
Oryza sativa	Monocot (Poaceae)	2x	390	NA	NA	NA
Lotus japonicus	Eudicot (Fabaceae)	2x	470	100	100	10.21
Medicago truncatula	Eudicot (Fabaceae)	2x	470	100	100	11.42
Zea mays	Monocot (Poaceae)	2x	2,700	10	10	1.44
Hordeum vulgare	Monocot (Poaceae)	2x	5,400	10	10	1.18
Triticum monococcum	Monocot (Poaceae)	2x	6,100	5	9	1.34

NA, not available.
*From the plant C value database, www.rbgkew.org.uk/cval, except rice (52).

posons in maize (8, 15) and by the recent lineage-specific amplification of some LTR retrotransposons in *Gossypium* (19). However, all studied plant genomes appear to contain mainly young LTR retrotransposons (30). This observation has been misinterpreted as evidence of an exceptional burst of LTR retrotransposon activity in the last few million years in all angiosperms, but is actually caused by the fact that LTR retrotransposon removal by homologous and illegitimate recombination has made it difficult or impossible to discern the older elements (22, 23, 31). It is true that genomes like that of maize would be 2-fold or more smaller without LTR retrotransposon amplifications over the last 2–3 million years (15), but that does not mean that the maize genome was 2-fold smaller than its current size 2–3 million years ago. In rice, for instance, unequal homologous recombination and illegitimate recombination have removed >190 Mb of LTR retrotransposon DNA in the last 2–5 million years (23), but the rice genome seems to have grown during at least part of that time period due to a high rate of LTR retrotransposon amplification (29).

Because retroelements do not excise when transposing, there is no organismal selection for transpositional function on an element after it has inserted. Most LTR retrotransposons also have no effect on the genes nearby, largely because they are usually epigenetically silenced (32, 33). If plant genes were not well insulated from most of the transposable elements that surround them, then it would be impossible for different haplotypes of maize (e.g., around the *bronze* locus) (14) or orthologous genes in different cereal species to have the same conserved gene regulation and function despite having completely dissimilar local genomic compositions (13). Hence, the majority of LTR retrotransposons come as close to selectively neutral as any DNA within a plant nuclear genome. Investigating sequence change in these elements can thereby indicate the actual genomic phenomena associated with DNA sequence change, both their nature and relative frequency, without the severe filter applied by natural selection. Moreover, comparisons of the rates and natures of LTR retrotransposon sequence divergence and of other intergenic DNA divergence in two subspecies of rice (29) indicated a complete concurrence in properties. Hence, the relative neutrality and great abundance of LTR retrotransposons in angiosperms makes them the perfect molecules for analysis of the properties and propensities in genome evolution across a wide range of species.

Here, we present analysis of the nature of LTR retrotransposon distribution and divergence in five angiosperms (barley, lotus, maize, *Medicago truncatula*, and a diploid wheat relative, *Triticum monococcum*), and compare these results with similar studies from *Arabidopsis* and rice (22, 23). The observations indicate that all of these genomes are highly dynamic, but with different mechanisms of instability contributing more in some species than others. The data also demonstrate that the described approach for sample sequence analysis is an efficient way to determine genome evolutionary dynamics that can be applied across a wide range of plant species.

Results

Supporting Information. For further details, see Tables 4–10, which are published as supporting information on the PNAS web site.

Selection and Analysis of Angiosperm Genomic Sequences. Comprehensive analysis of the structure and evolution of LTR retrotransposons requires high quality genomic sequence data as the starting point. Five flowering plant species were chosen for the current study (barley, lotus, maize, medicago and a diploid wheat) because they represent two major classes within the angiosperms (monocots and eudicots) and because previous sequencing studies have generated data for a sufficient number of BAC inserts from these species. The large inserts in BACs, usually >100 kb, are required to minimize edge effects where an LTR retrotransposon is only partially contained within a contiguous sequence. In fact, the numerous LTR retrotransposons found in this study that appeared to be truncated by the cloning process were excluded from the results because the complete structure of the element could not be determined.

To provide comparable data sets, the initial screening for LTR retrotransposons in each of the five targeted genomes was identical. In each case, 7–30 BAC sequences that contained genes were randomly selected from the entire genomic data set for that species. Although BACs that were randomly selected by the initial investigators would have been superior for this purpose (34, 35), none were available for barley, lotus, medicago, or *T. monococcum*.

The 7–30 BAC sequences chosen for each species were analyzed with a modification of LTR_STRUC (36) to find LTR retrotransposons by purely structural criteria. Hence, the study was limited to those LTR retrotransposon families that contained at least one element with two intact LTRs. After intact elements were identified, five different elements were chosen from each species, each serving as a reference copy for a particular element family. The reference copy for each family was then used in a homology search of the entire BAC data set, consisting of 7–100 different gene-containing BAC sequences, the number depending on the species investigated (Tables 1 and 7–9).

Identified LTR Retrotransposon Families. For the smaller genomes analyzed, 10 (lotus) and 30 (medicago) BAC sequences were barely enough to identify the minimum of five LTR retrotransposons families that were targeted for analysis. In the larger genomes (barley, maize, and wheat), 7–10 BACs were sufficient to identify many LTR retrotransposon families. Not surprisingly, the five first-found LTR retrotransposons families in the three large grass genomes were all elements that had previously been named and characterized, for instance *Huck*, *Opie*, *Ji*, *Cinful*, and *Zeon* of maize (Table 2). These elements are the most abundant in the maize genome, and were first identified on BACs or other clones that were isolated because they contained genes (7, 37).

Of the 25 LTR retrotransposon families identified by structural criteria, copy numbers of the intact elements in each family ranged

Table 2. LTR retrotransposon types and abundance

Species	LTR retro. family	Type	Name	No.	Genomic no. est.
L. japonicus	1	*copia*	*mable*	5	230
	2	*copia*	*susa*	2	90
	3	*gypsy*	*vygy*	6	280
	4	Unknown	*giube*	1	–
	5	Unknown	*wooput*	7	320
M. truncatula	1	*copia*	***MIRE-1***	28	1150
	2	Unknown	*ufabi*	5	200
	3	Unknown	*pesuwe*	11	450
	4	*gypsy*	*ofuw*	11	450
	5	*copia*	*pude*	2	80
Z. mays	1	*gypsy*	***Huck***	15	28,100
	2	*copia*	***Opie***	11	20,600
	3	*copia*	***Ji***	27	50,600
	4	*gypsy*	***Cinful***	7	13,100
	5	*gypsy*	***Zeon***	10	18,800
H. vulgare	1	*gypsy*	***BAGY-1***	5	13,100
	2	*copia*	***BARE-1***	32	83,900
	3	*copia*	***Claudia***	1	–
	4	*copia*	***Leojyg***	2	5,240
	5	*gypsy*	***Wham***	8	21,000
T. monococcum	1	*copia*	***Angela***	37	168,000
	2	*gypsy*	***Sabrina***	13	59,200
	3	*gypsy*	***Jeli***	3	13,700
	4	*gypsy*	***Romani***	3	13,700
	5	*gypsy*	***Latidu***	4	18,200

retro., retrotransposon; est., estimate. Preexisting names are in bold.

from one (eight families) to 21 (for *Angela* of *T. monococcum*) (Table 4). Counting all family members, including intact elements, solo LTRs, and clear truncation products, family sizes ranged from one (two families) to >25 (for *Angela* of wheat, *BARE-1* of barley, *MIRE-1* of medicago, and *Ji* of maize) (Table 4). Given the small percentages of these genomes investigated, any LTR retrotransposon found repeatedly on these BACs is likely to be highly abundant within its resident genome. Although an element family with only one representative in the analyzed BACs might be a low-copy-number DNA found by chance in the data set, families with two or more copies must represent a substantial part of their host genome. For instance, the most abundant element found in this analysis of lotus, with seven copies in the data set, is predicted to have >300 copies because only ≈1/46th of the lotus genome sequence was analyzed (10.2 of 470 Mb; Table 2). Hence, this study is restricted to the properties of the most abundant LTR retrotransposons in each of these genomes.

Comparison of the predicted family sizes for each of these five species shows a very good correlation between overall genome size and the presence of a few highly abundant LTR retrotransposon families. In *T. monococcum*, maize, and barley, the most abundant LTR retrotransposon families are predicted to have >40,000 copies per haploid genome, whereas the smaller genomes of lotus and medicago have few or no element families with >1,000 copies (Table 2).

Insertion Times. For intact LTR retrotransposons with flanking TSDs, one can date the time of element insertion by analysis of the degree of sequence divergence between the two LTRs of that element (38). Intact LTR retrotransposons without TSDs were not analyzed for insertion dates because they could be rare products of interelement recombination that yields two LTRs from different initial elements (22, 23). Different families within a species can show different times of activity (31, 39–41), so an average for several families is needed to acquire some information about overall genome dynamics. As reported in ref. 30, our results show that different angiosperms have very different average LTR retrotransposon insertion dates (bootstrapped ANOVA, $P < 0.001$), with most of the legume elements appearing to be particularly recent insertions (Table 3).

LTR Retrotransposon Structures Indicate Different Modes and Rates of Sequence Removal in Different Angiosperms. Analysis of the structures of the LTR retrotransposons within each family indicated dramatic differences in the distribution of element types within families and between species. In lotus, maize, and diploid wheat, intact elements with a retained TSD far outnumbered solo LTRs or any other type of element rearrangement (Table 3). Three classes of rearranged elements are expected to be the outcome of unequal homologous recombination. These are the solo LTRs with TSDs (intraelement recombination), plus the solo LTRs and apparently intact elements without TSDs. The presence of both solo LTRs and intact elements without TSDs in each of these species except lotus (where the data set was smallest) suggests that unequal recombination within individual elements or between different members of the same LTR retrotransposon family occurs in all angiosperms, and may be a significant factor in genomic DNA removal and other forms of chromosomal rearrangement (21).

The ratio of intact elements to apparent unequal recombination products (both types of solo LTRs plus the apparent intact elements without TSDs) is quite different between the five genomes studied, and in comparison to previous data for *Arabidopsis* and rice (Table 3). This characteristic does not exhibit strong correlation with genome size or with the apparent insertion dates of intact elements (Table 3). For instance, barley and medicago show very similar ratios of intact to recombined elements, despite their great difference in genome size (Table 3). The ratio of recombined to truncated elements also varies a great deal between species, suggesting different relative efficiencies of DNA removal by unequal homologous recombination or the small deletions associated with illegitimate recombination.

Table 3. Overview of LTR retrotransposon structure and age diversity

Species	% Intact	Intact: Rec.	Trunc.: Rec.	No. of intact elements dated	Insertion date, Mya* Range	Insertion date, Mya* Average	Ts:Tv	(Ts + Tv): Indel
A. thaliana†	30	0.5	0.9	87	0.2–4.4	1.9	ND	ND
O. sativa‡	27	0.8	0.5	260	0–5.9	1.3	ND	ND
L. japonicus	57	6.0	3.5	13	0–3.0	0.3	2.5	7.3
M. truncatula	37	1.6	1.8	19	0–0.6	0.2	2.4	3.6
Z. mays	61	5.4	2.4	47	0–2.3	0.7	3.9	8.2
H. vulgare	35	1.5	1.8	17	0–3.8	1.3	1.6	12.0
T. monococcum	50	4.3	3.3	30	0–3.3	1.0	1.9	8.7

Rec., recombined; Trunc., truncated; Mya, million year ago; Ts, transition; Tv, transversion; Indel, insertion or deletion; ND, not determined.
*Insertion times of LTR retrotransposons using LTR divergence method (38), with a subsitution rate of 1.3×10^{-8} substitutions per site per year (29).
†Data from ref. 22.
‡Data from ref. 23.

LTR Retrotransposon Sequence Diversity. Because the two LTRs of an LTR retrotransposon are usually identical at the time of element insertion, the degree and nature of sequence divergence that accumulates in the LTRs is a reflection of the processes that mutate the host genome. In each of these five species, the ratio of transitions to transversions in LTRs was found to be >1.5:1 (Tables 3 and 5). Because genic sequences, including introns, exhibit ≈1:1 ratios of transitions (Ts) to transversions (Tv), it has been argued that a higher Ts:Tv ratio is evidence of extensive cytosine 5-methylation, because this epigenetic DNA modification increases the C to T transition rate (38). Therefore, the >1.5:1 Ts:Tv ratio seen in plant LTR retrotransposons suggests that most or all of these elements are in an epigenetically silenced state associated with extensive cytosine 5-methylation, especially in the sequences CG and CNG (42). This finding further supports the choice of these molecules as neutrally evolving sequences, but also necessitates the use of a faster molecular clock for dating insertion times (29) (Table 3).

Although the Ts to Tv ratios in all of the examined species are greater than ≈1:1, the ratios differ dramatically between species, from 1.6 (barley) to 3.9 (maize). Statistical analysis using bootstrapped ANOVA indicates that the differences in Ts:Tv ratios are highly significant between species ($P < 0.001$). Indels (sequences that may be insertions or deletions) are also found at different frequencies across these species (Table 5). In addition, the ratios of point mutations (Ts plus Tv) to indels is quite variable, ranging from 3.6 to 12, but does not correlate with either phylogenetic relatedness or genome size (Table 3).

Among the 126 LTR pairs analyzed, 376 indels were detected. The two largest classes, about equal in number, are indels of unknown origin (50%) and indels associated with simple sequence repeats (44%) (Table 5). In the three cereals, a third class of indels is associated with LTR retrotransposon insertions into the LTRs, a common phenomenon in these large genome species. Unknown indels range in size from 1 to 116 bp, with a mean size of 10 bp. For indels of size >1 bp, 146 of 204 show sequence repetition at their termini (ranging in size from 2 to 53 bp, with an average size of 9 bp and a median at 7 bp) (Table 6), suggesting that they are the products of illegitimate recombination (22), as is also expected for many or all of the indels associated with the simple sequence repeats.

Discussion

Limitations to the Analysis. The study of repeats in any genome has the advantage that small data sets can provide large amounts of appropriate data. However, the manner in which the data are generated and the sequences are chosen may bias the conclusions that can be drawn. In this study, five angiosperms were chosen for study based on their variation in genome size, their phylogenetic dispersal, and the availability of reasonably long contiguous sequences of genomic data. Even with the five species chosen, only one (maize) has a significant data set of randomly selected BACs (34). Hence, the present study used only those BACs from maize that contained genes, so that the results could be compared with those for the gene-containing BACs like those available from the other targeted organisms. This method creates the caveat to the results in this work that they pertain only to gene-containing regions. Any LTR retrotransposons, or LTR retrotransposon properties, that are limited to largely gene-free regions like pericentromeric heterochromatin (43) would be missed in this study.

A second caveat to these studies is that they pertain only to the most abundant LTR retrotransposon families in each species. By chance, the structure-based search of a few BACs is most likely to first find the most abundant elements. However, a previous study on randomly chosen LTR retrotransposons in rice (23) found that the low-, medium-, and high-copy-number elements did not exhibit significant differences in structural properties or patterns of diversity.

Finally, elements that contain few or no intact family members would be missed or under-represented in the study undertaken. The first selection criterion for data analysis, that an intact element be found on one of the first 7–30 BACs analyzed, was instituted to avoid a sequence-based screen (e.g., a BLAST screen) that would bias for elements that are homologous to already-known LTR retrotransposons. Although this structure-based bias is significant in this study, its potential effects are limited only to interpretations of average element ages and to the relative distribution of element types, as discussed below.

Large Angiosperm Genomes Are Primarily Enlarged by the Activity of a Few Families of LTR Retrotransposons and Different Ones in Each Species Studied. All flowering plant genomes appear to have large quantities of LTR retrotransposons, but only the larger genomes contain any families with 10,000 or more members. Hence, very large plant genomes do not expand by having huge numbers of extra LTR retrotransposon families, but by having a few families with very high copy numbers. However, the mechanistic reasons for this difference are not clear. Do larger genomes somehow maintain a greater diversity of LTR retrotransposon families such that one or more of these have the chance to amplify to very high copy numbers? Or is there some deficient aspect of epigenetic silencing in the larger genome plants that allows certain element families to reach exceptional copy numbers? A third alternative is that some plant lineages or species have recently acquired, perhaps by a chance interspecific mating or other horizontal event, LTR retrotransposons that have the potential for massive amplification in the recipient genome. None of these possibilities are exclusive, of course, nor are they particularly well supported or undermined by current data. However, the differential levels of amplification of different families that occur within genomes suggest that variation in the reproductive potential of the elements themselves plays a larger role than interhost variation in the effectiveness of overall silencing mechanisms.

The most abundant LTR retrotransposons analyzed in this study do not fall into any particular small subset of LTR retrotransposon types. Some are *gypsy*-like elements, whereas others are *copia*-like elements. Even within these superfamilies, there is no obvious phylogenetic clustering of the most abundant elements between or within species. Hence, many different types of LTR retrotransposons have the potential to become highly repetitive and thus greatly influence overall plant genome size. The fact that plant genome size can vary so quickly (44, 45), and sometimes due primarily to the action of a small number of LTR retrotransposon families (19), indicates that activation of these elements can be a very dynamic process.

LTR Retrotransposon Insertion Dates. As noted (30), most detectable LTR retrotransposons appear to have inserted within the last 0–4 million years in all plant species analyzed. This does not mean that LTR retrotransposons have only recently been active, because all of the older elements are likely to no longer be available for dating because they have undergone deletions that removed all or most of at least one LTR. When truncated elements that contain small segments of two shared LTRs are dated, some are found to be old insertions (e.g., >10 million years; ref. 23).

The LTR retrotransposon insertion dates in this study are from too few families and too few elements to allow any major interpretation of the significances of the differences that were observed. Moreover, the calculation of relative insertion dates across species required the simplifying assumption that the rate of point mutation is identical in each of these species, an assumption that is unlikely to be consistently valid. Use of the sequence change rate of 1.3×10^{-8} substitutions per site per year, acquired from analysis of LTR retrotransposon sequence divergence in rice (29), provides a more accurate and consistent set of insertion dates within this five-genome analysis.

It is noteworthy that both genome size and phylogenetic relationships trend in the same direction as differences in average LTR retrotransposon insertion date for these five angiosperms. The species calculated to have the oldest average insertion dates are closely related species with the largest genomes, *T. monococcum* and barley. At the other end of the spectrum, the species with the youngest average LTR retrotransposon insertions are the two smallest genome species and are both legumes, medicago and lotus. This finding suggests that DNA removal may be more rapid in legumes than in the Triticeae, although more data are needed to further investigate this possibility.

Unequal Homologous Recombination and Illegitimate Recombination Are Major Agents of DNA Removal in Angiosperms, but Appear to Differ in Their Relative Significance Between Species. An earlier analysis of LTR retrotransposons suggested that unequal homologous recombination was a more quantitatively important DNA removal activity in rice than in *Arabidopsis* (23); this was evidenced by the fact that the ratio of intact to recombined to truncated LTR retrotransposons was ≈0.8:1:0.9 in *Arabidopsis* and 0.5:1:0.5 in rice. In these studies, the calculated number of truncated elements includes only those with at least one partially intact LTR, because annotation of tiny truncation remnants can be highly subjective. This limitation does not affect the accuracy of the calculation of the ratio of intact elements to recombined elements, but it does massively under-predict the number of truncated element fragments.

The comparisons across species are biased by the fact that LTR retrotransposons with few intact elements will be missed in a small data set. The larger data sets analyzed for *Arabidopsis* and rice uncovered some LTR retrotransposons with very rare intact elements and very low ratios of intact to recombined or truncated elements (22, 23). However, the similar size data sets for maize, barley, and diploid wheat (for instance) are all comparable within this study, so that the major differences observed in element type distribution points to major differences in genomic mechanisms for DNA removal.

It is particularly interesting that the relative abundance of intact LTR retrotransposons does not directly correlate with genome size in the five species studied. Maize and lotus (a large and a small genome) have the highest percentages of intact elements (61% and 57%, respectively), whereas barley and medicago (a large and a small genome) have the lowest percentages of intact elements (35% and 37%, respectively) (Table 3). This finding suggests that differences in rates of DNA removal have been less significant across these lineages in the determination of genome size, and that relative differences in LTR retrotransposon amplification have been a more significant factor.

Although the ratio of intact elements to other element types could be biased by our structure-based approach to LTR retrotransposon discovery, the ratio of truncated elements to recombined elements should not be affected. Hence, the results for the five angiosperms in this study suggest very different relative efficiencies of DNA removal by unequal homologous recombination or the small deletions associated with illegitimate recombination. There is no apparent correlation of the relative efficiencies of DNA removal by unequal homologous recombination or illegitimate recombination with either phylogenetic relatedness or genome size (Table 3). This finding suggests that the relative aggressiveness of these DNA removal mechanisms frequently changes during the evolutionary descent of plant lineages, and that it is not the primary determinant of genome size. The high relative levels of unequal homologous recombination in *Arabidopsis* (22) and, especially, rice (23) (Table 3) may be an indication of investigator-based differences in detection of truncated elements, although every effort was made to exactly replicate the annotation procedure.

Substitutions and Indels in Diverging LTRs. Because of their relatively neutral status in natural selection, and their identity at the time of insertion, paired LTRs within an intact element are excellent indicators of the primary qualitative and quantitative nature of sequence divergence. Compared with genes and genic regions in plants, all of these species indicated a greater number of indels relative to point mutations. Once again, this is a likely indicator of the degree to which the LTR retrotransposons evolve in a selectively neutral manner.

Most of the nucleotide substitution mutations in each of the studied species are transitions. This has been a routine observation for LTR retrotransposons in plants, and has been argued to be an outcome of their high degree of 5-methylation of cytosines in the sequences CG and CNG (38). The methylation of cytosine at the 5 position causes an increase in its rate of deamination, and the deamination leads to a direct 5meC to T transition. Therefore, 5meC is consistently observed as a hotspot for transition mutations (46). The relatively low ratios of transitions to transversions in *T. monococcum* and barley is an intriguing observation. Perhaps the two Triticeae are unusually rich in LTR retrotransposons that are not epigenetically silenced.

Most of the indels found in the LTR retrotransposons in all of these species are small and of unknown origin, many within simple sequence repeats, and exhibit terminal direct duplications. Therefore, it seems likely that these are the legacies of illegitimate recombination. This process, and unequal homologous recombination, are the major factors responsible for DNA removal in all plant species examined (22, 23), and the results indicate that this is also true in these five angiosperms. With the data available, it is impossible to determine whether DNA removal processes are stronger than the factors behind genome growth (primarily LTR retrotransposon amplification) in these five species. Comparisons across orthologous regions in closely related species would be needed to draw these quantitative conclusions, as has been done for some members of the genus *Oryza* (29). For now, it is clear that all of the molecular processes of genome growth and shrinkage are continuously modifying the structures of all five of these angiosperm genomes.

Genome Dynamics. LTR retrotransposon analysis has uncovered several mechanisms and impressive rates of nuclear genome change in flowering plants. Several properties, including the relative activity of unequal homologous and illegitimate recombination and the potential for very high copy numbers of an LTR retrotransposon family, appear to be very different across angiosperm species. A broader sampling of plant species will be needed to track down the actual points in the evolutionary descent of different plant lineages where these apparent changes in the relative activity of various mutational mechanisms arose (47). With this information in hand, the causes and outcomes of the different genome dynamics in each plant lineage might be ascertained.

Materials and Methods

Selection of Genomic Sequences for Analysis. Initial identification of LTR retrotransposons was performed by running a modified version of LTR_STRUC (36) on 7–10 BACs for each species. The major modification of LTR_STRUC was that the program now detects intact LTR retrotransposons within nested insertion arrays. The 7–10 initial BACs were selected as follows: for barley, of the 13 BACs publicly available in June 2006, all but three (GenBank accession nos. AY661558, AF427791, and AY642926) were analyzed. For *T. monococcum*, all available BACs were used (nine contigs corresponding to seven BACs). For medicago and lotus, 10 BACs were randomly selected from the gene-rich BACs available at http://medicago.org/genome/BACregistry.php and www.kazusa.or.jp/lotus after selection for size >80 kb. For medicago, 10 BACs were insufficient to discover five distinct LTR retrotransposon

families, so the sample was enlarged by screening 20 more BACs. Among the five species, maize is the only one for which sequenced BACs were randomly selected. To get a sample comparable to the other genomes, 10 gene-rich maize BACs were chosen. Three came from published results (GenBank accession nos. AF123535, AF448416, and AY555142) (7, 14, 48). The other seven correspond to randomly selected BACs (www.broad.mit.edu/annotation/plants/maize/randomclones.html) that were characterized as containing genes (6, 34). When several contigs were available for these BACs, only the longest was kept, to avoid truncation of elements due to contig assembly, except *T. monococcum*, for which the total number of contigs was limited. A list of the 7–30 initially analyzed BACs for each species is available in Table 7.

LTR Retrotransposon Discovery and Selection. For each species, the LTR retrotransposons found by modified LTR_STRUC were manually checked (presence of two well defined LTRs, clear boundaries) and compared by using an all-vs.-all BLASTN search to cluster the different copies corresponding to the same element. One copy was chosen as reference for each family. The number of copies found in the genome sample was computed, and the element class (*gypsy*-like, *copia*-like, unknown) was determined by BLASTN and T-BLASTX searches of Poaceae, Brassicaceae, and Fabaceae repeat databases (http://tigrblast.tigr.org/euk-blast/index.cgi?project=plant.repeats) and Psi-Blast searches (www.ncbi.nlm.nih.gov/blast) on all-possible-reading-frame translated sequences of each element. To select five element families for analysis, the following criteria were used: (*i*) select elements with the most copies and (*ii*) for elements with only one copy, choose elements that are located on different BACs (to maximize the number of genomic regions represented) and that maximize diversity in LTR retrotransposon types. For each of the five families, one copy was chosen as reference for subsequent analyses. The reference copies are listed in Table 10. Obvious insertions found in these reference sequences were removed before running the BLASTN search.

Characterization of LTR Retrotransposon Structures. The LTR retrotransposon reference copy was used in a BLASTN search (www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) on a total of 7–100 BACs, depending on the species analyzed (Table 1), with homologs identified by a maximum *e* value of e^{-10}. Copies were extracted and their truncation level (copies with two LTRs and internal region, copies with one LTR and part of the internal region, solo LTRs, and other fragmented elements; Table 4) were identified. Fragments corresponding to internal regions alone were not taken into account, as they could correspond to a cross-match with a different element family within the polyprotein sequence. TSDs were checked for all complete copies and solo LTRs. For lotus and medicago, a sample of 100 BACs (including the 10 or 30 BACs used for the original search) was screened, due to low copy numbers. A list of these additional BACs is available in Tables 8 and 9. The classification of the different types of structures is as presented in Ma *et al.* (23). Intact elements without TSD and solo LTRs (with and without TSD) were considered as "recombined," whereas elements showing truncation in at least one LTR and LTR remnants were classified as "truncated."

Estimation of LTR Retrotransposon Insertion Dates. Insertion dates were estimated for copies with two LTRs (intact or somewhat truncated) and TSDs following the method of SanMiguel *et al.* (38). Each LTR pair was aligned by using CLUSTALX (49) and manually inspected. Divergence between LTRs was computed by MEGA version 3.0 (50), using the Kimura 2 parameter distance (51) that corrects for homoplasy and differences in the rates of transition and transversion. LTR divergence data were converted into million years using a substitution rate of 1.3×10^{-8} substitutions per site per year, as discovered by Ma and Bennetzen (29).

Substitutions and Indels in LTRs. The number of transition and transversion mutations were estimated between the LTRs of copies harboring two LTRs (intact or partially truncated) and target site duplications using MEGA version 3.0 (50), whereas indels were identified and analyzed by manual inspection.

Statistical Analysis. Differences in the patterns observed between species were analyzed by using bootstrapped ANOVA, which accounts for possible nonnormal distribution of the data. Analyses were performed by using the software Resampling version 1.3 (Resampling Procedures, D. C. Howell, University of Vermont, Burlington, VT). A total of 5,000 bootstrap replicates were generated for each test.

We thank Drs. J. Ma, R. Liu, and G. Baucom for advice regarding data selection and analysis and Dr. E. McCarthy for assistance with initial LTR retrotransposon identification. This research was supported by Grant DBI-0501814 from the Plant Genome Program at the National Science Foundation.

1. Leitch IJ, Soltis DE, Soltis PS, Bennett MD (2005) *Ann Bot* 95:207–217.
2. Bennetzen JL, Ramakrishna W (2002) *Proc Natl Acad Sci USA* 99:9093–9095.
3. The *Arabidopsis* Genome Initiative (2000) *Nature* 408:796–815.
4. Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H (2002) *DNA Res* 9:117–121.
5. Bennett MD, Leitch IJ, Price HJ, Johnston JS (2003) *Ann Bot* 91:547–557.
6. Liu R (2005) PhD dissertation (University of Georgia, Athens, GA).
7. SanMiguel P, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D, Melake Berhan A, Springer PS, Edwards KJ, Avramova Z, Bennetzen JL (1996) *Science* 274:765–768.
8. Meyers BC, Tingey SV, Morgante M (2001) *Genome Res* 11:1660–1676.
9. Emberton J, Ma J, Yuan Y, Bennetzen JL (2005) *Genome Res* 15:1441–1446.
10. Gill KS, Gill BS, Endo TR, Boyko EV (1996) *Genetics* 143:1001–1012.
11. Sandhu D, Gill KS (2002) *Plant Physiol* 128:803–811.
12. Li W, Zhang WP, Fellers JP, Friebe B, Gill BS (2004) *Plant J* 40:500–511.
13. Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) *Proc Natl Acad Sci USA* 96:7409–7414.
14. Fu H, Dooner HK (2002) *Proc Natl Acad Sci USA* 99:9573–9578.
15. SanMiguel P, Bennetzen JL (1998) *Ann Bot* 82:37–44.
16. Vicient CM, Suoniemi A, Anamthawat-Jónsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) *Plant Cell* 11:1769–1784.
17. Dubcovsky J, Ramakrishna W, SanMiguel PJ, Busso CS, Yan L, Shiloff BA, Bennetzen JL (2001) *Plant Physiol* 125:1342–1353.
18. Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) *Plant J* 26:307–316.
19. Hawkins JS, Kim HR, Nason JD, Wing RA, Wendel JF (2006) *Genome Res*, in press.
20. Roeder GS, Fink GR (1980) *Cell* 21:239–249.
21. Garfinkel DJ (2005) *Cytogenetic Genome Res* 110:63–69.
22. Devos KM, Brown JK, Bennetzen JL (2002) *Genome Res* 12:1075–1079.
23. Ma J, Devos KM, Bennetzen JL (2004) *Genome Res* 14:860–869.
24. Ehrlich SD, Bierne H, d'Alencon E, Vilette D, Petranovic M, Noirot P, Michel B (1993) *Gene* 135:161–166.
25. Kirik A, Salomon S, Puchta H (2000) *EMBO J* 19:5562–5566.
26. Petrov DA, Lozovskaya ER, Hartl DL (1996) *Nature* 384:346–349.
27. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) *Science* 287:1060–1062.
28. Bensasson D, Petrov DA, Zhang DX, Hartl DL, Hewitt GM (2001) *Mol Biol Evol* 18:246–253.
29. Ma J, Bennetzen JL (2004) *Proc Natl Acad Sci USA* 101:12404–12410.
30. Bennetzen JL, Ma J, Devos K (2005) *Ann Bot* 95:127–132.
31. Kapitonov VV, Jurka J (2003) *Proc Natl Acad Sci USA* 100:6569–6574.
32. Kashkush K, Feldman M, Levy AA (2003) *Nat Genet* 33:102–106.
33. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, *et al.* (2004) *Nature* 430:471–476.
34. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Birren B, *et al.* (2005) *Plant Physiol* 139:1612–1624.
35. Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) *Proc Natl Acad Sci USA* 102:19243–19248.
36. McCarthy EM, McDonald JF (2003) *Bioinformatics* 19:362–367.
37. Hu W, Das OP, Messing J (1995) *Mol Gen Genetics* 248:471–480.
38. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) *Nat Genet* 20:43–45.
39. Panaud O, Vitte C, Hivert J, Muszlak S, Talag J, Brar D, Sarr A (2002) *Mol Gen Genomics* 268:113–121.
40. Vitte C, Panaud O (2003) *Mol Biol Evol* 20:528–540.
41. Pereira V (2004) *Genome Biol* 5:R79.
42. Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981) *Nature* 292:860–862.
43. Ma J, Bennetzen JL (2006) *Proc Natl Acad Sci USA* 103:383–388.
44. Kellogg EA, Bennetzen JL (2004) *Am J Bot* 91:1709–1725.
45. Caetano-Anolle's G (2005) *Crop Sci* 45:1809–1816.
46. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) *Nature* 274:775–780.
47. Bennetzen JL (1999) *Trends Genet* 15:85–87.
48. Ma J, Lai J, Messing J, Bennetzen JL (2005) *Genetics* 170:1209–1220.
49. Thompson JD, Gibson TJ, Plewniak F, JeanMougin F, Higgins DG (1997) *Nucleic Acids Res* 25:4876–4882.
50. Kumar S, Tamura K, Nei M (2004) *Brief Bioinform* 5:150–163.
51. Kimura M (1980) *J Mol Evol* 16:111–120.
52. Sasaki T, Matsumoto T, Antonio BA, Nagamura Y (2005) *Plant Cell Physiol* 46:3–13.