Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families

Thomas Wicker and Beat Keller

Genome Res. 2007 17: 1072-1081 originally published online June 7, 2007
Access the most recent version at doi:10.1101/gr.6214107

Letter

Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families

Thomas Wicker[1] and Beat Keller

Institute of Plant Biology, University of Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland

Although *copia* retrotransposons are major components of all plant genomes, the evolutionary relationships between individual *copia* families and between elements from different plant species are only poorly studied. We used 20 *copia* families from the large-genome plants barley and wheat to identify 46 families of homologous *copia* elements from rice and 22 from *Arabidopsis*, two plant species with much smaller genomes. In total, 599 *copia* elements were analyzed. Phylogenetic analysis showed that *copia* elements from the four species can be classified into six ancient lineages that existed before the divergence of monocots and dicots. The six lineages show a surprising degree of conservation in sequence organization and other characteristics across species. Additionally, the phylogenetic data suggest at least one case of horizontal gene transfer between the *Arabidopsis* and rice lineages. Insertion time estimates for 522 high-copy elements showed that retrotransposons from rice were active at different times in waves of activity lasting 0.5–2 million years, depending on the family, whereas elements from wheat and barley had longer periods of activity. We estimated that half of the rice *copia* elements are truncated or otherwise rearranged after ~790,000 yr, which is almost twice the half-life of *Arabidopsis* elements. In contrast, wheat and barley *copia* elements appear to have a massively longer half-life, beyond our ability to estimate from the available data. These findings suggest that genome size can be explained by the specific rate of DNA removal from the genome and the length of active periods of retrotransposon families.

[Supplemental material is available online at www.genome.org.]

Retrotransposons are the predominant class of transposable elements in plants and are largely responsible for the vast differences in genome sizes. The relatively small and compact genome of *Arabidopsis* has a size of ~120 Mbp and contains only ~10% repetitive DNA (*Arabidopsis* Genome Initiative 2000), whereas the rice genome has 3.2 times that size (389 Mbp) and contains at least 35% of repetitive DNA (International Rice Genome Sequencing Project 2005). In contrast, species from the Triticeae tribe, which includes important crop plants such as wheat or barley, have diploid genomes that are almost 15 times the size of the rice genome and contain at least 80% repetitive DNA (~5500 Mbp) (Bennett and Smith 1976). So far, it is not known why plant genomes differ so strongly in their retrotransposons content.

copia retrotransposons are flanked by long terminal repeats (LTRs) that contain promoter and downstream control elements. Their internal domain usually contains the genes required for reproduction such as reverse transcriptase (*RT*), integrase (*INT*), and *gag*, which is probably involved in packaging RNA or DNA during replication (for review, see Wilhelm and Wilhelm 2001). The retrotransposon life cycle starts with the transcription of the DNA into mRNA. After translation, the reverse transcriptase produces a DNA copy of the mRNA, which is then integrated elsewhere in the genome by the integrase enzyme. Thus, each replication cycle produces an additional copy of the element (Wilhelm and Wilhelm 2001).

[1]Corresponding author.
E-mail wicker@botinst.unizh.ch; fax 41-44-634-82-04.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6214107.

Because only one of the two LTR sequences serves as template for reverse transcription, both LTRs are identical at the time of insertion. This characteristic can be used to estimate the age of retrotransposons (SanMiguel et al. 1998). Since retrotransposons are believed to be largely free from selection pressure, mutations are expected to accumulate randomly across the entire element (Petrov 2001). These mutations cause the two LTRs to diverge at a rate that is proportional to the age of the element. The absolute age of the element can therefore be estimated by applying a basic substitution rate. This basic rate was estimated to be 1.3×10^{-8} substitutions per site per year in grasses (Ma and Bennetzen 2004) and a very similar value of 1.5×10^{-8} in *Arabidopsis* (Koch et al. 2000). An earlier study proposed the basic substitution rate to be about half the above value in grasses (6.5×10^{-9}) (Gaut et al. 1996), and it is not yet established which rate should be used. However, since this method for estimating insertion times is a linear function, results can be easily converted by multiplying or dividing with the factor 2.16. In this study, we applied the rate of 1.3×10^{-8} for all species used (wheat, barley, rice, and *Arabidopsis*) to provide equal treatment of all analyzed retrotransposons. However, we do not know whether this rate is equally valid in all species.

The vast majority of plant retrotransposons studied so far were estimated to be <3 million years old (SanMiguel et al. 1998, 2002; Wicker et al. 2003a, 2005a; Gao et al. 2004; Ma et al. 2004; Du et al. 2006; Piegu et al. 2006) when the basic substitution rate of 1.3×10^{-8} is applied. This finding was surprising and led to the conclusion that repetitive DNA is constantly deleted from the genome through unequal crossing-over and illegitimate recom-

17:1072–1081 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

bination (Vicient et al. 1999; Devos et al. 2002; Pereira 2004). This permanent removal of repetitive sequences results in a complete reshuffling of intergenic regions within only a few million years (SanMiguel et al. 2002; Wicker et al. 2003a). If an individual retrotransposon is affected by a deletion (e.g., the loss of one LTR), it is likely to become defective and will not be able to replicate anymore. A recent study in *Arabidopsis* revealed an exponential age distribution among LTR retrotransposons. From this, the half-life for complete elements (i.e., copies that contain both LTRs) of ~470,000 yr outside of centromeric regions was estimated (Pereira 2004). This finding indicated that the rate of DNA removal from the *Arabidopsis* genome has remained constant for a long time. However, the number of retrotransposons in a genome can grow within a relatively short evolutionary time span if a particular family becomes active. Such bursts of activity can cause a rapid genome expansion, as was shown in barley (Kalendar et al. 2000) and *Oryza australiensis*, a wild relative of rice (Piegu et al. 2006).

Many retrotransposons are nonautonomous elements that rely for their replication completely or in part on proteins expressed by other elements elsewhere in the genome (Vitte and Panaud 2005). For example, *BARE1* elements from barley have a defective *gag* domain and probably use the protein encoded by the closely related *BARE2* elements for their replication (Sabot and Schulman 2006). Nonautonomous elements with highly degenerated coding regions or no coding capacity at all have been found in all TE classes (e.g., MITES, CACTA transposons, or LTR retrotransposons (Bureau and Wessler 1994; Wicker et al. 2003b; Kalendar et al. 2004). Additionally, many copies of retrotransposons that contain defective coding regions with frameshifts and in-frame stop codons may actually be the offspring of functional elements. This can be due to the high error rate of reverse transcription, producing so-called "death-on-arrival" copies (Gabriel et al. 1996; Keulen et al. 1997), or due to mutations that were accumulated after the element inserted into the genome.

The objective of this study was a comparative analysis of *copia* elements from plant genomes that differ greatly in size and also represent different clades of the phylogenetic tree. We used the sequences from Triticeae repeat database (TREP) as a starting point for a homology search in rice and *Arabidopsis*. At the time this study was done, 32 families of *copia* retrotransposons were represented at TREP. We were able to identify rice homologs for all Triticeae families and traced the evolutionary lineages back to the divergence of monocots and dicots by comparison with *Arabidopsis* elements. We identified six ancient lineages of *copia* elements that existed before the divergence of monocots and dicots. Each of the six lineages shows a surprising degree of conservation and has distinct characteristics. Our analysis also demonstrated that specific retrotransposon families invade genomes in waves of high activity that are followed by long periods of relative silence.

Results

Of the 32 families of *copia* elements deposited at TREP, 20 have at least one copy that is complete, whereas 12 families are only represented by fragments. A "complete" element is defined as a copy that has intact ends, but does not make any statement about whether the element is actually functional or whether it contains internal deletions. In fact, the majority of the complete elements deposited at TREP have degenerated coding regions that contain frameshifts, stop codons, or deletions. For this analysis, we used the 20 families for which complete elements were available.

For eight of the 20 *copia* families, only a single complete element is deposited at TREP. Nine families have two to four members (*Ale*, *Barbara*, *Bianca*, *HORPIA2*, *Ikeros*, *Inga*, *Maximus*, *Oref*, and *TAR1*) and most abundant are the closely related *Angela*, *BARE1*, and *WIS* families for which 15, 17, and 26 complete copies are available, respectively (Supplemental Table 1). Multicopy elements were used in multiple sequence alignments to obtain a consensus sequence. In most cases, this consensus sequence gave rise to an intact open reading frame (ORF), because frameshifts and stop codons are usually the result of mutations in one specific copy and are eliminated in the consensus sequence. For single-copy elements, the hypothetical protein sequences were deduced by comparison with proteins from similar *copia* elements, and frameshifts were removed manually to obtain the putative ORFs.

Most Triticeae *copia* elements have homologs in the rice genome

Rice homologs were identified by using consensus sequences or single-copy elements from Triticeae for BLASTN searches against the rice genome. All except *HORPIA2* and *Oref* produced strong DNA alignments (*E*-values $<10 \times 10^{-10}$) at multiple loci in the rice genome. In all cases, the conserved region corresponded to the reverse transcriptase domain. The conserved regions, including 6 kb of flanking sequence on each side, were excised from the rice genome sequence (in silico) and screened for the presence of long direct repeats flanking the BLASTN hit (i.e., LTRs that flank the coding region). To isolate large numbers of full-size complete elements from the rice genome, specific software was designed that checked the putative LTRs for the conserved TG-CA termini, as well as for the presence of a 5-bp target-site duplication. If multiple complete elements were isolated, a consensus sequence was constructed, which was then used to deduce hypothetical protein sequences.

For this study, we defined a "family" based on common LTR DNA identity. LTRs are among the most rapidly evolving parts of retrotransposons, because they do not encode any proteins. Often, conserved parts of the coding region of retrotransposons can be 85%–90% identical, while their LTRs are highly divergent. Thus, we considered two elements as belonging to the same family if their LTRs are at least 80% identical. This was for a practical reason, as we wanted to be able to derive consensus sequences for families with multiple copies. Using these criteria, we could classify 31 of the 46 identified rice *copia* families as previously described elements. Sixteen families are novel. Five of them have an internal domain that is very similar to that of previously described elements (>90% DNA sequence identity), but have highly divergent LTRs (Supplemental Table 2).

In some cases, different *copia* elements from Triticeae identified the same family (or groups of families) of rice elements, indicating that these Triticeae elements diverged after the Triticeae/rice separation. Analogously, some families from Triticeae identified multiple rice families.

Arabidopsis homologs of grass *copia* elements

When the Triticeae and rice *copia* sequences were used in BLASTN searches against the *Arabidopsis* genome, most elements did not

identify sequences with significant DNA homology (*E*-value $<1 \times 10^{-10}$). Only *Adena* and *Osr8* elements from rice showed DNA sequence homology over several hundred base pairs with *Arabidopsis* sequences. The *Arabidopsis* homolog of *Adena* (referred to as *Anika*) was found in multiple copies in the *Arabidopsis* genome, whereas the *Osr8* element was found in one single copy in *Arabidopsis*. Curiously, the *Osr8* element plus 780 bp of its flanking regions is located between two gaps in the *Arabidopsis* genome. This exact sequence (*Osr8* plus flanking sequence) is found on chromosome 10 in rice. Thus, it is possible that this particular sequence fragment represents a (probably in silico) contamination.

Due to the low level of DNA sequence conservation, all other *Arabidopsis* homologs of rice or Triticeae families were identified by a TBLASTN search of the predicted protein sequences against the *Arabidopsis* genome. This search identified 22 families of *Arabidopsis copia* elements, 17 of which were previously described. Again, in some cases, multiple Triticeae elements identified the same *Arabidopsis copia* family (or group of families), indicating a divergence of the grass elements after the monocot/dicot separation. Isolation and characterization of *Arabidopsis copia* elements was done in the same way as for rice elements. However, due to the lower copy number of most elements, consensus sequences could not be obtained for all of them. For some, the protein sequences had to be predicted from one single element by comparison with similar protein sequences.

Most *copia* elements have a low copy number, while few are very abundant

The consensus sequences from rice and *Arabidopsis copia* families were used to identify further complete copies by BLASTN against their respective genomes. The *Arabidopsis copia* families all have a low-copy number, whereas in rice, some are present in hundreds of copies in the genome. Most abundant in rice is the *Houba* family, for which 151 complete copies were identified. The second most abundant family is *Osr8*, with 77 full-length copies. The actual copy numbers of these families are much higher, as many copies are fragments or solo-LTRs, which were not considered in this study. Six additional families from rice were found in 20 or more copies (*Adena*, *Ostonor1*, *Osr1*, *Osr10*, *SC3*, and *Tara*). In contrast, the most abundant *Arabidopsis* family is *AtCopia78*, with only eight copies, followed by *AtCopia58* and *Anika* (five and three copies, respectively). For the majority of *Arabidopsis copia* families, we found only one or two full-length copies.

Due to the limited data set for Triticeae, one can only speculate about the actual copy number of retrotransposons in the entire genome. The fact that for the majority of retrotransposons only one or a few copies were available, indicates that the Triticeae genomes must contain a vast number of different families with moderate copy numbers. In this study, we considered every Triticeae *copia* family for which more than one copy was available as high-copy. The only *copia* elements that are present in high-copy numbers in the available Triticeae sequences are the closely related *Angela*, *BARE1*, and *WIS* families.

copia elements from grasses and their *Arabidopsis* homologs can be separated into six major evolutionary lineages

Approximately 500 amino acids covering the reverse transcriptase domain (the most conserved region of the *copia* elements) from 77 families were used for the construction of a phylogenetic tree. A *copia* protein from yeast was used as outgroup. For the tree in Figure 1, a subset of 52 sequences was used (for some closely related families, just one representative was used) due to space constraints, while the complete tree is available as Supplemental Figure 1.

The sequences cluster into six major evolutionary lineages (*Angela*, *Ale*, *Bianca*, *Ivana*, *Maximus*, and *TAR*) (Fig. 1). We defined the lineages primarily as large groups that are in a common branch of the tree with a very high bootstrap value (<95) and secondarily by shared characteristics (see below). Some branches that separate the six major lineages have lower bootstrap values, indicating that the precise relationships between the lineages are difficult to assess and that they had been evolving independently for a long time.

Relationships within the *Bianca* and *TAR* lineages reflect the expected pattern with elements from rice and Triticeae more closely related and *Arabidopsis* elements placed in a separate branch. The *Bianca* lineage contains only three families, one for each Triticeae, rice, and *Arabidopsis*, whereas the *TAR* lineage contains three families of rice elements, indicating that they diverged after the rice/Triticeae separation (Fig. 1).

Within the *Ale* lineage, the relationships are less clear, which is reflected in lower bootstrap values. This could indicate that either the families within these lineages diverged within a short evolutionary time or that sequence exchange has occurred between families. Specifically in *Arabidopsis*, a large number of families were identified. In contrast, both the *Ivana* and *Maximus* lineages show a wide variety in Triticeae and rice families, whereas in *Arabidopsis*, only one and two families were identified, respectively.

For the *Ikeros* family, a rice homolog (*Ostonor1*), but no *Arabidopsis* homologs, could be identified. This could either indicate that these particular families became extinct in *Arabidopsis* or that the actual homolog is simply too divergent to be identified by the criteria used (Fig. 1). Similarly, there is no clear *Arabidopsis* homolog for the group containing the *Inav* and *Ale* elements from Triticeae.

Elements from the same lineage have similar characteristics in Triticeae, rice, and *Arabidopsis*

The six evolutionary lineages have distinct characteristics that are common to all of their member families from Triticeae, rice, and *Arabidopsis*. Families from the same lineage are in a relatively narrow size range and most of them also have LTRs of similar sizes (Table 1). There is also a general tendency that lineages either contain mostly high-copy or mostly low-copy families.

The *Ale* lineage is the most diverse, as it contains 36 families. Interestingly, for 28 families, only one single full-length copy was found, and the most abundant family (*rn154–162* from rice) was found in only five complete copies. Representatives of the *Ale* lineage are the smallest of all *copia* elements studied, and range in size from 4.4 to 5.5 kb. Especially in *Arabidopsis*, *Ale* elements diverged massively, as 14 families were identified, most of them with only one complete copy in the genome (Supplemental Figure 1; Supplemental Table 1).

The largest elements (8.7–14.4 kb) were found in the *Maximus* lineage. Their unusual size is caused by long LTRs, large stretches of noncoding sequences in the internal domain, as well as by the presence of a unique second open reading frame (*ORF2*) downstream of the gag-*INT*-*RT* domain. *ORF2* is very divergent among different *Maximus* families, and no information as to the function could be found. However, its position suggests it to be

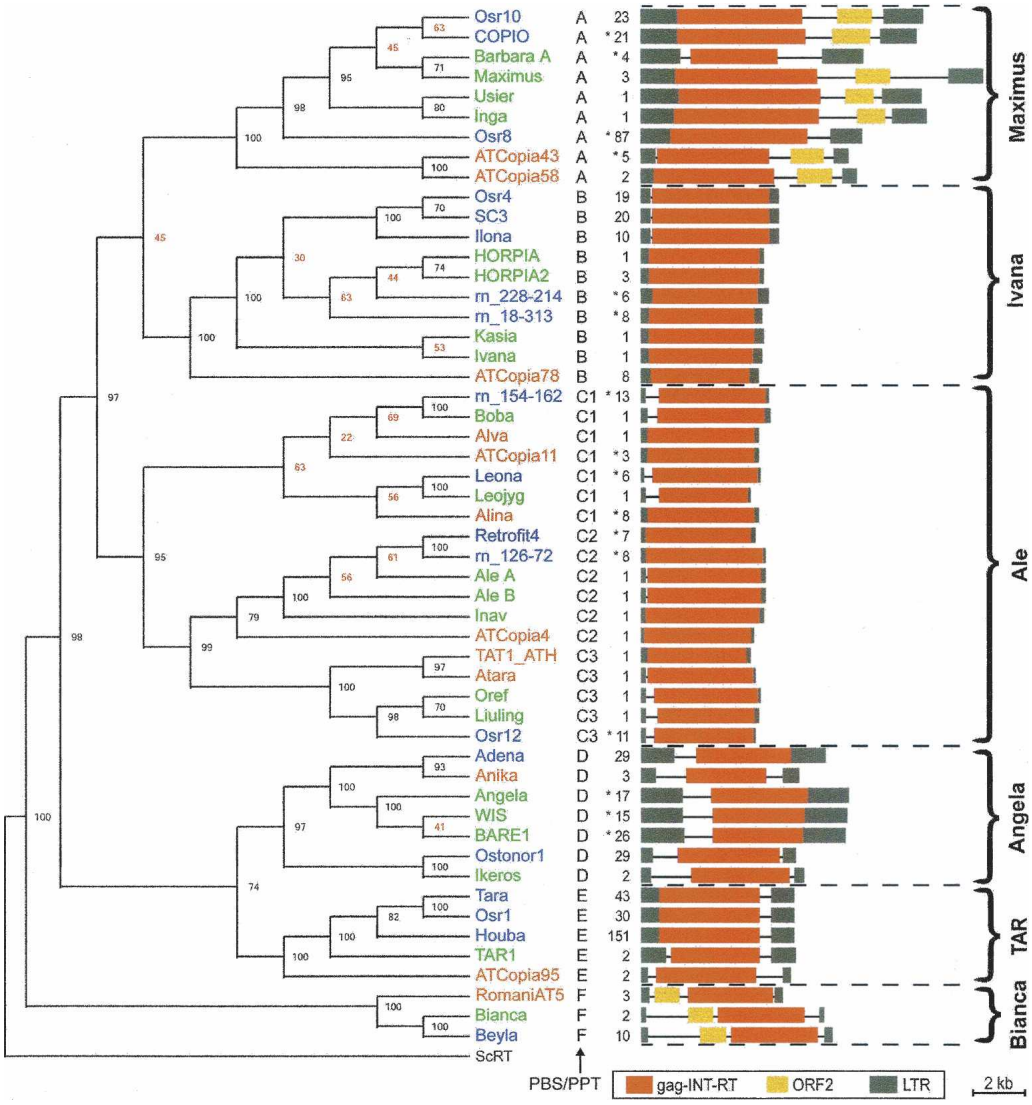

Figure 1. Phylogenetic tree of 52 *copia* families from Triticeae, rice, and *Arabidopsis*. Names of *copia* families from *Arabidopsis* are printed in red, those from rice in blue, and those from Triticeae in green. Subfamilies are indicated by a capital letter at the end of the name. Black uppercase letters refer to the type of primer binding site (PBS) and polypurine tract (PPT) detailed in Figure 2. Asterisks indicate the presence of additional, closely related families that had been omitted due to space constraints. Copy numbers in these cases refer to the total of all represented families (see Supplemental Fig. 1). Bootstrap numbers at the forks indicate how many times the sequences to the *right* of the fork occurred in the same group of 100 trees. Strong bootstrap values of at least 80 are shown in black. The copy numbers and sequence organization of all families are displayed next to the respective names. Major evolutionary lineages are indicated by curly brackets. A reverse transcriptase sequence from yeast (ScRT) served as outgroup.

homologous to the *env* gene of retroviruses (Wilhelm and Wilhelm 2001).

The three families of the *Bianca* lineage also contain a unique ORF, which is located upstream of the gag-*INT-RT* domain (in contrast to *ORF2* of *Maximus* families, which is located downstream of gag-*INT-RT*). *Bianca* probably represents an ancient lineage, as it is placed on a separate branch most distantly related to all other lineages in the phylogenetic tree.

Table 1. Characteristics of main evolutionary lineages of *copia* elements

Lineage	Size (kb)	LTR (bp)	Families	Comment
Ivana	5–5.7	321–519	14	mostly high-copy
Maximus	8.7–14.4	549–1746	12	3′ ORF2, mostly high-copy
Ale	4.4–5.4	88–403	36	mostly low-copy
TAR	6.2–6.5	274–1028	5	mostly high-copy
Angela	6.4–8.9	375–1822	9	mostly high-copy
Bianca	6.9–8.0	175–341	3	5′ ORF2

The *Angela* and *TAR* lineages are similar in size and overall sequence organization with long LTRs and several hundred base pairs of noncoding DNA in their internal domain. *TAR* contains the most abundant *copia* family identified in this study (*Houba* from rice), whereas *Angela* contains three of the most abundant Triticeae families (*Angela*, *BARE1*, and *WIS*). The families from the *Ivana* lineage also have relatively high-copy numbers. In comparison with *TAR* and *Angela*, they are compact, with short LTRs, and almost the entire internal domain is comprised of coding region.

Primer binding sites and polypurine tracts reflect the relationships between lineages

Primer binding site (PBS) and polypurine tract (PPT) are specific sequence motifs of ~20–25 bp, which are located immediately downstream of the 5′LTR and immediately upstream of the 3′ LTR, and which are necessary for the replication of retrotransposons. PBS and PPT were isolated from all *copia* families and used in multiple alignments. The 11–15 bp at the 5′ end of the PBS were found to be highly conserved within evolutionary lineages, allowing the deduction of consensus sequences for all six lineages in that region (Fig. 2); the bases further downstream were too variable to obtain reliable consensus sequences (Supplemental Fig. 2). The PPT was found to be more variable than the PBS, which is why for some lineages only short stretches of consensus sequences could be obtained (Fig. 2; Supplemental Figs. 2 and 3). The consensus sequences of PBS and PPT matched the data from the phylogenetic analysis very well, as all six evolutionary lineages have distinct motifs within their PBS (Fig. 2; Supplemental Fig. 3). They also reflect the overall structure of the phylogenetic tree, as PBS from more closely related lineages share more common bases (Fig. 2). Indeed, in some cases PBS and PPT showed some variations that even reflected subgroups that had lower bootstrap values in the phylogenetic tree (Fig. 1; Supplemental Fig. 3). For example, the *Ale* lineage could be divided into three sublineages (Figs. 1, 2). Interestingly, the PBS of the *TAR* lineage shows basically no similarity to any of the others, indicating an ancient rearrangement specific to the *TAR* lineage.

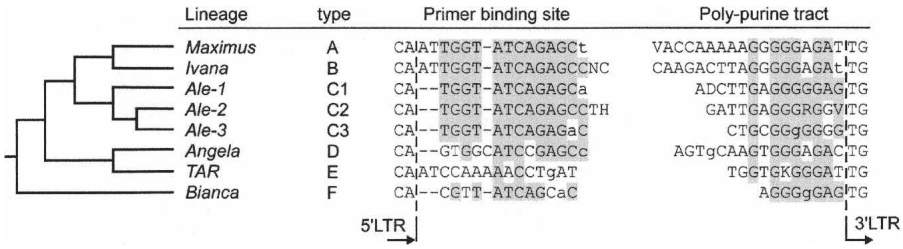

Figure 2. Comparison of primer binding sites (PBS) and polypurine tracts (PPT) of the six evolutionary *copia* lineages. The tree diagram at the *left* illustrates the relationship between the six evolutionary lineages as determined through the phylogenetic analysis shown in Figure 1.The terminal two bases at the 5′ end of PBS and at the 3′ end of PPT belong to the 5′ and 3′ LTRs, respectively. Shown are consensus sequences for five main lineages as well as the *Ale* lineage, which is split into three sublineages due to strong divergence within the *Ale* lineage. Each PBS/PPT pair is given a "type" index that corresponds to the one in Figure 1. The full-sequences alignments are available as Supplemental Figure 2.

The *Maximus* lineage contains multiple populations of deletion derivatives

Many *copia* elements analyzed in this study contain deletions, but most of them are unique to one specific copy and probably resulted in disruption of functionality of the element. However, some retrotransposons that have lost some parts in a deletion are obviously still able to replicate by relying on proteins encoded by full-length elements. Thus, a population of deletion derivatives (which we refer to as "subfamily") is established in the genome, in addition to the population of autonomous elements. Curiously, all four deletion-derivative populations identified belonged to families of the *Maximus* lineage.

Osr9 from rice does not have an ORF2 region, but is otherwise very similar in size and sequence to other *Maximus* elements from rice (e.g., *COPIO*) (Fig. 3). It is likely that the *Osr9* family originated from a single deletion event that eliminated the ORF2 region.

Barbara_A and *Barbara_B* from wheat show a more complex pattern, as both subfamilies share common sequences only in the LTRs, whereas their internal domains are completely different (Fig. 3). It appears that *Barbara_A* has lost its *gag* and ORF2 regions in two independent deletions. Interestingly, the *Barbara_B* family almost perfectly complements *Barbara_A*, as it lacks only the *INT-RT* region but still contains both *gag* and ORF2. *Barbara_A* and *Barbara_B* have very similar LTRs (~83% identity), indicating that the divergence of these two subfamilies occurred relatively recently.

The high-copy family *Osr8* consists of two populations (*Osr8* and *Osr8B*). *Osr8* represents a putative autonomous element and is more abundant (61 complete copies) than *Osr8B* (16 complete copies). *Osr8B* appears to be a deletion derivative of *Osr8*, which contains only the *gag* domain but lacks the *INT-RT* region. Interestingly, the LTRs plus the 5′ part of the *gag* region is very well conserved between *Osr8* and *Osr8B*, whereas the 3′ part of *gag* is divergent (Fig. 3). One explanation is that one of the two is a recombined element that carries part of the coding region of a different subfamily of an *Osr8*-like element. Du et al. (2006) suggested that such a sequence exchange could occur, for example, through template switching during the replicative process. Alternatively, the observed pattern could reflect different degrees of selection pressure on the different regions of the two elements. Interestingly, *Osr8* itself appears to be a deletion derivative, as it does not contain the ORF2 characteristic for the *Maximus* lineage.

The *Arabidopsis* genome contains a small population of *ATCopia58* elements, which consists of one putatively autonomous copy of *ATCopia58*, four deletion derivatives (*ATCopia58B*), and four solo-LTRs. All four deletion derivatives have the same structure, namely, four different deletions compared with the full-length *ATCopia58* element, indicating that they originated from the same ancestor element (Fig. 3).

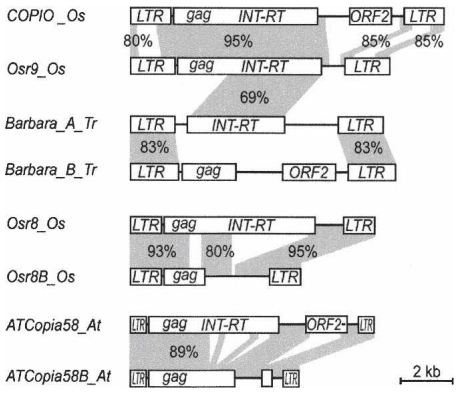

Figure 3. Deletion derivatives in *copia* elements of the *Maximus* lineage. Gray areas represent regions conserved between elements with the degree of DNA sequence identity indicated. The low degree of sequence conservation (69%) between the rice *Osr9* elements and the Triticeae *Barbara* elements shows that the loss of the ORF2 region in each family was the result of an independent deletion event.

Waves of genome invasion

The large number of *copia* elements available from Triticeae and rice inspired an analysis of the history of *copia* activity in these genomes. We measured the divergence of LTRs of 522 identified complete elements from high-copy families (i.e., families for which at least 10 complete copies were found). These included 58 copies of the three closely related *Angela*, *BARE1*, and *WIS* families from Triticeae, as well as 464 copies from 14 rice *copia* families. To obtain estimates when these elements have inserted into the genome, we applied a basic substitution rate of 1.3×10^{-8} per site per year (Ma and Bennetzen 2004). The element with the most divergent LTRs (86%) was thus estimated to have inserted ~5.7 million years ago (Mya), whereas the vast majority have LTRs that are 93%–100% identical, corresponding to <3 million years of age (Fig. 4). This is in agreement with earlier findings from Triticeae, rice, and maize (SanMiguel et al. 1998, 2002; Wicker et al. 2003a; Gao et al. 2004; Ma et al. 2004; Du et al. 2006). The 10 elements that were estimated to be older than three million years were inspected for possible gene conversion events or other rearrangements that could explain the strong divergence of their LTRs. Only one of them showed statistically significant indication for such an event (data not shown).

Graphical display of LTR divergence of members of *copia* families revealed that different families were active at different times during evolution (Fig. 4). Statistical Bonferroni-corrected Kolomogorov-Smirnov tests showed the insertion distributions of all but four families to be significantly different from a uniform distribution (Fig. 4; Supplemental Table 3). Most of the analyzed families apparently had active periods of 1–2 million years, which were followed by periods of relative silence. The most abundant family, *Houba*, shows two active periods, one within the last 500,000 yr and a second one ~1–3 Mya. The rice *Adena* family had a highly active period 1–2.5 Mya and apparently became silenced after that, since no copies younger than one million years were found.

In contrast, the *Ostnor1*, *Osr8*, and *Osr10* families were obviously active for more than two million years, but at a lower level, producing fewer copies than the *Houba* family. The *Beyla*, *COPIO*, and *Echidne* families each had one wave of moderate activity during roughly one million years.

Osr8 elements were very active for a long period ~0.5–3 Mya, whereas the deletion derivative *Osr8B* was active in a more narrow time span of 1–2.5 Mya. It makes sense that the time of activity of *Osr8B* falls within the active wave of *Osr8* elements, if one assumes that *Osr8B* depended on *Osr8* for its replication. Thus, it seems that the deletion derivative *Osr8B* was very successful for some time during evolution, but ceased to replicate about 1 Mya.

The three *copia* families from wheat and barley have waves of activity that are much more spread out in time than those of the rice elements. All three families had a period of relatively high activity 0.5–1.5 Mya, but also produced copies during the entire last three million years (Fig. 4). The youngest element (*Angela_AF326781-6*) inserted only ~45,000 yr ago, whereas the oldest (*Angela_AF497474-1*) inserted almost 3 Mya.

copia elements are truncated or removed from the rice genome at a half-life rate of about 790,000 yr, and more slowly from the wheat genome

To estimate the rate at which full-size copies are (at least partially) removed from the genome, we divided the estimated insertion times of all 506 available full-length *copia* elements from rice, wheat, and barley into bins of 100,000 yr (Fig. 5). Here, we included all elements from high- and low-copy families to obtain an overall distribution of insertion times. The largest number of retrotransposons inserted within the past 100,000 yr, and the number of insertions, is generally declining in the more distant past.

Assuming that repetitive sequences are removed from the genome at a constant rate, insertion time distribution can be described by an exponential function with a constant half-life rate. To estimate the average half-life of *copia* elements, we fitted an exponential function with half-life and starting value as variables to the observed data by minimizing the sum of distances between the calculated and the observed value for each bin. The calculations resulted in an average half-life of 794,000 yr. Thus, on average, half of the full-length elements are at least partially

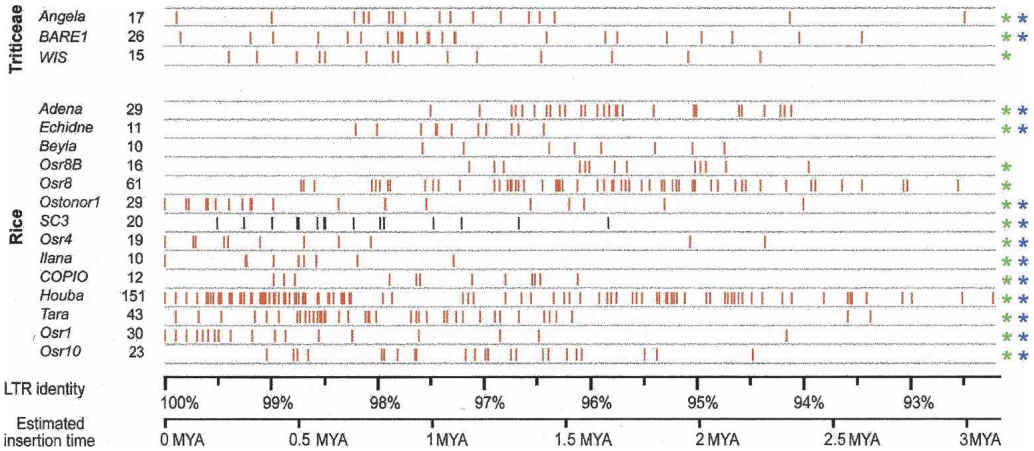

Figure 4. Estimated insertion times of high-copy *copia* families from Triticeae and rice. The families studied are listed in the *left*most column. Numbers of elements analyzed for each family are given in the second column. Individual insertion events are indicated as vertical red lines. The x-axis scale indicates the DNA sequence identity between LTRs of a particular element (*top*) and the estimated insertion time derived from it using a basic substitution rate of 1.3×10^{-8} per site per year (*bottom*). Asterisks indicate whether the distribution is significantly different from a uniform distribution ($P < 0.05$), before (green) and after (blue) Bonferroni correction.

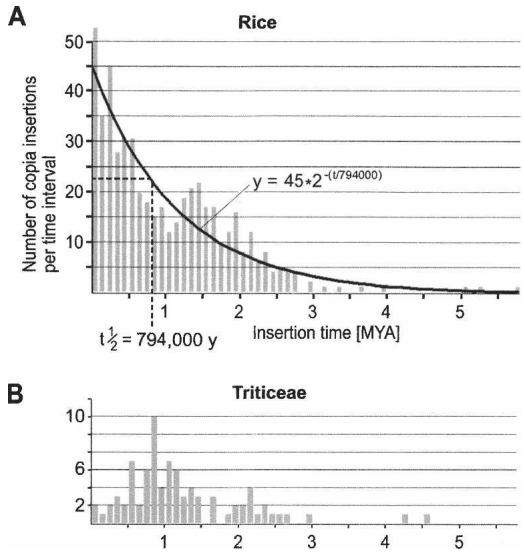

Figure 5. Distribution of *copia* element insertion times. Estimated insertion times were divided into bins of 100,000 yr. (*A*) The distribution of rice *copia* elements can be described as a hyperbolic function based on the assumption that retrotransposon sequences are at least partially removed from the genome at a constant rate. The curve represents the best fit for the starting value (43.6) and the half-life (796,000). (*B*) Insertion distribution of 87 *copia* elements from Triticeae (wheat and barley). The distribution is not similar to a hyperbolic distribution.

removed from the genome after 794,000 yr. The observed distribution deviates to some degree from the calculated distribution, especially for the period of 1–2 Mya, where multiple families were active (Figs. 4, 5A).

It is important to emphasize that this analysis only included complete elements. If a retrotransposon is only partially deleted (e.g., a part of one LTR is missing), it is not considered in the analysis. Thus, our estimate does not make a statement on how long it takes until repetitive elements are completely removed from the genome as previous studies have done (Vitte and Bennetzen 2006), but provides an estimate for the average survival time of complete elements.

Despite the smaller sample size for Triticeae elements, an attempt was made to obtain an estimate of their half-life. Indeed, the distribution of insertion times of 86 *copia* elements from barley and wheat does not resemble an exponential distribution at all, making it meaningless to find the best-fitting exponential function (Fig. 5B). Although the available data set is too small to draw definite conclusions, the observed distribution may reflect a very long half-life of *copia* elements in the Triticeae genomes and, thus, a fundamental difference of small and large genome plants.

Evidence for horizontal gene transfer between *Arabidopsis* and rice

Of all Triticeae and rice families studied, only the *Adena* element from rice showed strong DNA sequence conservation with *Arabidopsis* sequences. Indeed, the rice *Adena* is more closely related to its *Arabidopsis* than to its Triticeae homologs: the coding region of the rice *Adena* is 76.3% identical to *Anika* from *Arabidopsis* over more than 2600 bp. Interestingly, both are equally distantly related to the wheat *Angela* family and can only be aligned over ~2100 bp with 70.2% identity. Additionally, the phylogenetic analysis of the reverse transcriptase protein sequence clearly places the rice sequence closer to the *Arabidopsis* sequence. We interpret this data as evidence for a horizontal gene transfer from *Arabidopsis* to rice or vice versa. Two alternative explanations (which we consider less likely) are: first, both the *Arabidopsis* and rice elements are under stronger selection pressure than their Triticeae homologs, and thus evolve more slowly; second, the actual homologous lineage of the rice and *Arabidopsis* elements became extinct in Triticeae, and elements that are actually much more distantly related were compared. However, this would not explain why the rice and *Arabidopsis Anika* are the only elements that show such a high degree of conservation at the DNA level. More plant species would have to be analyzed to answer this question.

Discussion

The presented comparative analysis of *copia* elements from wheat, barley, rice, and *Arabidopsis* provided a broad insight into the evolution of these retrotransposons. Our analysis compared *copia* elements from plant genomes with very different characteristics. First, they represent the major clades of the monocots and dicots and, second, they differ greatly in genome size, as *Arabidopsis* represents a small, rice, a medium-sized, and barley and wheat, large genome species. Our data showed that *copia* elements in these three genomes are descendants of only a few ancient evolutionary lineages, each with distinct conserved characteristics. Individual families were shown to be active for a few million years and are then silenced for similarly long periods. Additionally, our data suggest that at least in one case, horizontal gene transfer might have occurred between the *Arabidopsis* and rice lineages.

Is genome size the product of a species-specific rate of DNA removal and the length of waves of retrotransposon activity?

Our data show that *copia* elements are active in waves, rather than continuously. We could distinguish three patterns of activity in the rice *copia* families. First, families that are present in very high-copy numbers show a relatively short burst of intense activity of a few 100,000 yr, in which dozens or hundreds of copies are produced. Examples of such elements are the *Houba*, *Tara*, and *Osr1* families. Second, some high-copy families are permanently active over several million years, but at a lower level. Third, some families produce only a moderate activity that can also be spread out over relatively long periods of time (~1–2 million years).

Interestingly, the three families of high-copy elements from wheat and barley show much longer waves of activity than all of the rice families studied. This could reflect one of the fundamental differences between rice and Triticeae species and contribute to the vast difference in genome sizes between the two. A permanent high activity of certain retrotransposon families could, therefore, have caused the expansion of the Triticeae genomes. This expansion phase presumably started after the divergence from its closest relative with a small compact genome, *Brachypodium*, ~35 Mya (Bossolini et al. 2007).

The question of whether the Triticeae genomes are still expanding remains unanswered. The fact that all Triticeae have similar genome sizes implies that their genome sizes have been constant at least since their divergence 11–14 Mya (Wolfe et al. 1989). However, it is also possible that all Triticeae genomes have

been growing at the same rate since their divergence. The second possibility seems to be supported by the observation that the *copia* families *Angela* and *WIS* from wheat have activity patterns that are not distinguishable from the one of the barley *BARE1* family, based on the available data.

It was repeatedly speculated that the balance of deletion and generation of DNA through retrotransposition is responsible for the differences in genome sizes (for review, see Vitte and Panaud 2005). In a recent study, the half-life of complete *copia* elements in the *Arabidopsis* genome was found to be ~470,000 yr for elements outside of centromeric regions (Pereira 2004). Applying the substitution rate of 1.3×10^{-8} described by Ma and Bennetzen (2004) would slightly lower that half-life estimate to about 407,000 yr. That is, within ~400,000 yr, half of all *copia* elements are at least partially deleted or otherwise rearranged. This is roughly half of our estimate of 790,000 yr for *copia* elements in rice. The fact that the insertion time distribution of Triticeae *copia* elements did not resemble at all an exponential distribution suggests that the half-life of repetitive elements in Triticeae may be much higher than in rice or *Arabidopsis*. Thus, it seems that the rate at which complete *copia* elements are truncated or deleted from the Triticeae genomes is much lower, resulting in retrotransposons residing for a longer time in the genome. Considering that the Triticeae *copia* families studied also have longer periods of activity, the combination of the two factors (DNA removal and retrotransposon activity) can, in fact, explain the large genome of the Triticeae.

From these data, we suggest that genome size is largely the product of DNA removal rate and the length of periods of retrotransposon activity. Already, small differences in the rates of DNA removal and/or retrotransposon activity could have a profound effect on genome size due to the additive nature of the two factors. One can imagine that a genome can grow rapidly if the rate of DNA removal is decreased slightly, while periods of retrotransposon activity are prolonged slightly. Analogously, the genome can contract quickly when the situation is reversed. Thus, the genome sizes we observe today are probably very different from what they were in the recent evolutionary past or from what they will be in the near future.

Conserved characteristics of *copia* lineages

It is intriguing how strongly the main evolutionary lineages and sublineages of *copia* elements are conserved between species that diverged ~50 Mya (rice/Triticeae) (Paterson et al. 2004) and 140–150 Mya (monocots/dicots) (Chaw et al. 2004), respectively. In addition to a general similarity in sequence organization, families from the same lineage show additional common characteristics. For example, the *Ale* lineage has the tendency to evolve a wide variety of mostly very low-copy families. In contrast, the most ancient lineage (*Bianca*) consists of only one family in each species, but is always found in multiple copies. An additional and very puzzling characteristic of *Maximus* elements is to produce populations of deletion derivatives in Triticeae, rice, and *Arabidopsis*. There are many examples for deletion-derivative populations in a wide range of species; often they turn out to be more numerous than their autonomous "mother" elements, e.g., MITEs in grasses (Bureau and Wessler 1994), *En1* from maize (Pereira et al. 1986), or *Galluhop* from chicken (Wicker et al. 2005b). Nevertheless, it is surprising that a particular lineage seems to be prone to give rise to such elements.

Repetitive DNA is generally believed to be free from selection pressure and often referred to as "selfish" DNA (Petrov 2001). If retrotransposons were indeed purely selfish genomic parasites, the situation would be comparable to a host–pathogen relationship, where the pathogen permanently evolves new strategies to overcome the host defense and the host responds with the evolution of new resistance genes. Such a "genetic arms race" is believed to lead to the observed rapid and dynamic evolution of resistance genes in plants (Holub 2001). A genetic arms race between retrotransposons and their hosts might result in high-retrotransposon diversity, with lineages and families basically randomly being eliminated or flourishing within species. However, our data show the opposite. Different lineages and families of *copia* elements show a surprising degree of conservation across species, and individual families become extinct only very rarely in a species. This suggests that individual families or lineages are under balancing selection and have distinct functions that were conserved at least since the monocot/dicot divergence 140–150 Mya.

Concluding remarks

If copia elements are indeed not purely selfish, but are under specific selection pressure, our view of transposable elements would, of course, have to be fundamentally revised. Thus, one must consider other possible explanations for the observed strong conservation of evolutionary lineages.

In our view, the most viable alternative explanation is frequent horizontal transfer of these elements across species boundaries, which would lead to an overall homogenization of the retrotransposon gene pool. A previous study actually suggested that conservation of lineages across distantly related species is due to frequent horizontal transfer between species (McCarthy et al. 2002). Indeed, we found one case that strongly suggests a horizontal transfer between rice and *Arabidopsis*. However, except for that one case, our analysis of 77 *copia* families produced a surprisingly consistent phylogenetic tree, in which sublineages and groups of families show the expected pattern that homologs from rice and Triticeae are more closely related to each other than to those of *Arabidopsis*. Conservation through homogenization would, in particular, require frequent gene transfer between monocots and dicots, otherwise the strong conservation of elements in plant genomes that diverged 140–150 Mya could not be explained.

Between closely related cross-breeding species, one can expect that transfer of transposable elements is frequent, and thus provides a stabilizing force that prevents individual families from extinction. However, this can only have an effect for relatively short evolutionary time spans, as long as freshly diverged species can still cross-breed.

One could argue that *copia* elements are mostly selfish, but that the presence of certain types of elements (e.g., representatives of the six evolutionary lineages described) can somehow have beneficial effects on the host population. Recently, it was shown that levels of retrotransposon silencing vary in a species-specific manner, indicating that permanent background activity occurs at least in some species (Vitte and Bennetzen 2006). Additionally, certain families could be activated as a specific response to environmental factors such as drought or heat stress. For example, *BARE1* elements in barley were shown to be activated by drought stress (Kalendar et al. 2000). Thus, different families might be activated by different environmental conditions, leading to the observed pattern of activity waves. If that is

the case, genomes would carry a record of past environmental conditions in their retrotransposon populations.

Methods

Sequence analysis tools

For sequence analysis, BLAST (Altschul et al. 1997), CLUSTALW (Thompson et al. 1994), and DOTTER (Sonnhammer and Durbin 1995) were used. Stand-alone Blast software was obtained from NCBI (http://www.ncbi.nih.gov). Repetitive elements from Triticeae were obtained from the TREP database (http://wheat.pw.usda.gov/ITMI/Repeats). Pairwise sequence alignments were done with the EMBOSS program WATER (http://emboss.sourceforge.net/). For analysis of rice and *Arabidopsis* sequences, datasets from TIGR rice genome (rice version 4 and *Arabidopsis* version 5, www.tigr.org) were used. Previously described *copia* elements from rice and *Arabidopsis* were obtained from Repbase (http://www.girinst.org/server/RepBase). Statistical analysis was done with the R-package (http://www.r-project.org). Possible gene-conversion events were detected by testing whether substitutions are uniformly distributed between LTRs.

Multiple alignments were done with CLUSTALW (Thompson et al. 1994) using a gap creation penalty of 1.0 and a gap extension penalty of 0.01 for protein and default values for DNA alignments. Phylogenetic analysis was performed with the PHYLIP package (http://evolution.genetics.washington.edu/phylip/) using the protein sequence parsimony method (PROTPARS) on 100 bootstrap replicates with jumbling the order of sequences five times for each replicate. The alignments used for the analysis are provided (Supplemental Figs. 2, 3). Initial multiple alignments of PBS and PPT sequences were done with CLUSTALW using a gap creation penalty of 6.0 and a gap extension penalty of 0.06. Alignments were adjusted manually, and consensus sequences were also deduced manually using the following rules: uppercase letters were used for positions where at least half plus one of the sequences showed the respective base (e.g., five of seven), whereas lowercase letters indicate positions where more than half of the sequences had the displayed base (e.g., three of five). The IUPAC code was used to describe base composition at that position where no base was present in at least half of the sequences.

Isolation of complete *copia* elements form the rice and *Arabidopsis* genomes

For characterization of *copia* families from the rice and *Arabidopsis* genome, a series of Perl scripts were written. A first script excised a fragment of 6 kb to each side of all regions that gave strong BLAST hits (*E*-values $<1 \times 10^{-10}$). A second script used the program WATER to find direct repeats (i.e., LTRs) flanking the region of the BLAST hit, and extracted the terminal regions of the paired sequences for manual check for target-site duplications and the canonical LTR termini TG-CA. If available, 10–20 elements were isolated in this way to identify the presence of dominant subfamilies and to construct consensus sequences. In a next step, the LTRs of all identified subfamilies were used for a second round of BLASTN searches against the genome, and a third Perl script was used to identify occurrences of two LTRs within the same 20 kb (allowing for some nested insertions) and to extract these sequences (plus 5 bp to each side to include possible target site duplications).

Nested insertions were detected by alignment of the extracted sequences with the respective consensus sequence using the program WATER. This process was also automated by means of a Perl script that included the detection of target-site duplications to exclude elements that were the product of inter-element recombination. Alignments of LTR pairs were done with the program WATER using a gap creation penalty of 30 and a gap extension penalty of 0.5. A Perl script was written to extract the number of aligned bases and number of mutations (transitions and transversions) from large numbers of WATER outputs.

All consensus DNA and protein sequences, as well as the complete set of *copia* sequences used for this study are available upon request.

copia elements were classified by BLAST against publicly available repeat databases such as RepBase (http://www.girinst.org/repbase/update/), the TIGR repeat database (www.tigr.org), RetrOryza (http://www.retroryza.org), and the data set described by Pereira (2004).

Acknowledgments

We thank Dr. Vini Pereira for kindly providing his dataset of *Arabidopsis copia* elements and Adrian Roellin for his help with statistical analyses. This work was supported by the Swiss National Science Foundation (SNF) grant 3100AV-105620.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

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

Bennett, M.D. and Smith, J.B. 1976. Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **274:** 227–274.

Bossolini, E., Wicker, T., Knobel, P., and Keller, B. 2007. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: Implications for wheat genomics and grass genome annotation. *Plant J.* **49:** 704–717.

Bureau, T. and Wessler, S.R. 1994. *Stowaway*: A new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Proc. Natl. Acad. Sci.* **9:** 1411–1415.

Chaw, S.M., Chang, C.C., Chen, H.L., and Li, W.H. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J. Mol. Evol.* **58:** 424–441.

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12:** 1075–1079.

Du, C., Swigonova, Z., and Messing, J. 2006. Retrotranspositions in orthologous regions of closely related grass species. *BMC Evol. Biol.* **6:** 62. doi: 10.1186/1471-2148-6-62.

Gabriel, A., Willems, M., Mules, E.H., and Boeke, J.D. 1996. Replication infidelity during a single cycle of Ty1 retrotransposition. *Proc. Natl. Acad. Sci.* **93:** 7767–7771.

Gao, L., McCarthy, E.M., Ganko, E.W., and McDonald, J.F. 2004. Evolutionary history of *Oryza sativa* LTR retrotransposons: A preliminary survey of the rice genome sequences. *BMC Genomics* **5:** 18.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. 1996. Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci.* **93:** 10274–10279.

Holub, E.B. 2001. The arms race is ancient history in *Arabidopsis*, the wildflower. *Nat. Rev. Genet.* **2:** 516–527.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436:** 793–800.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A.H. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci.* **97:** 6603–6607.

Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A.H. 2004. Large retrotransposon derivatives: Abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* **166:** 1437–1450.

Keulen, W., Nijhuis, M., Schuurman, R., Berkhout, B., and Boucher, C. 1997. Reverse transcriptase fidelity and HIV-1 variation. *Science* **275:** 229–231.
Koch, M.A., Haubold, B., and Mitchell-Olds, T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*). *Mol. Biol. Evol.* **17:** 1483–1498.
Ma, J. and Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* **101:** 12404–12410.
Ma, J., Devos, K.M., and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14:** 860–869.
McCarthy, E.M., Liu, J., Lizhi, G., and McDonald, J.F. 2002. Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol.* **3:** RESEARCH0053.1–0053.11. doi: 10.1186/gb-2002-3-10-research0053.
Paterson, A.H., Bowers, J.E., and Chapman, B.A. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci.* **101:** 9903–9908.
Pereira, V. 2004. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biol.* **10:** R79.
Pereira, A., Cuypers, H., Gierl, A., Sommer, Z.S., and Saedler, H. 1986. Molecular analysis of the *En/Spm* transposable element system of *Zea mays*. *EMBO J.* **5:** 835–841.
Petrov, D.A. 2001. Evolution of genome size: New approaches to an old problem. *Trends Genet.* **17:** 23–28.
Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., et al. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16:** 1262–1269.
Sabot, F. and Schulman, A.H. 2006. Parasitism and the retrotransposon life cycle in plants: A hitchhiker's guide to the genome. *Heredity* **97:** 381–388.
SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.
SanMiguel, P.J., Ramakrishna, W., Bennetzen, J.L., Busso, C., and Dubcovsky, J. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5Am. *Funct. Integr. Genomics* **2:** 70–80.
Sonnhammer, E.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** 1–10.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W, improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Vicient, C.M., Suonemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., and Schulman, A.H. 1999. Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell* **11:** 1769–1784.
Vitte, C. and Bennetzen, J. L. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl. Acad. Sci.* **103:** 17638–17643.
Vitte, C. and Panaud, O. 2005. LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenet. Genome Res.* **110:** 91–107.
Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.-D., Dubcovsky, J., and Keller, B. 2003a. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell* **15:** 1186–1197.
Wicker, T., Guyot, R., Yahiaoui, N., and Keller, B. 2003b. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.* **132:** 52–63.
Wicker, T., Zimmermann, W., Perovic, D., Paterson, A.H., Ganal, M., Graner, A., and Stein, N. 2005a. A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: Recombination, rearrangements and repeats. *Plant J.* **41:** 184–194.
Wicker, T., Robertson, J.S., Schulze, S.R., Feltus, F.A., Magrini, V., Morrison, J.A., Mardis, E.R., Wilson, R.K., Peterson, D.G., Paterson, A.H., et al. 2005b. The repetitive landscape of the chicken genome. *Genome Res.* **15:** 126–136.
Wilhelm, M. and Wilhelm, F.X. 2001. Reverse transcription of retroviruses and LTR retrotransposons. *Cell. Mol. Life Sci.* **58:** 1246–1262.
Wolfe, K.H., Gouy, M.L., Yang, Y.W., Sharp, P.M., and Li, W.H. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA-sequence data. *Proc. Natl. Acad. Sci.* **86:** 6201–6205.

Received December 15, 2006; accepted in revised form April 12, 2007.