Patching gaps in plant genomes results in gene movement and erosion of colinearity

Thomas Wicker,[1] Jan P. Buchmann, and Beat Keller

Institute of Plant Biology, University Zurich, CH-8008 Zurich, Switzerland

Colinearity of genes in plant genomes generally decreases with increasing evolutionary distance while the actual number of genes remains more or less constant. To characterize the molecular mechanisms of this "gene movement," we identified non-colinear genes by three-way comparison of the genomes of *Brachypodium*, rice, and sorghum. We found that genomic fragments of up to 50 kb containing the non-colinear genes are duplicated to acceptor sites elsewhere in the genome. Apparent movement of genes may usually be the result of subsequent deletions of genes in the donor region. Often, the duplicated fragments are precisely bordered by transposable elements (TEs) at the acceptor site. Highly diagnostic sequence motifs at these borders strongly suggest that these gene movements were the result of double-strand break (DSB) repair through synthesis-dependent strand annealing. In these cases, a copy of the foreign DNA fragment is used as filler DNA to repair the DSB linked with the transposition of TEs. Interestingly, most TEs we found associated with gene movement have a very low copy number in the genome and for several we did not find autonomous copies. This suggests that some of these elements spontaneously arose from unspecific interaction with TE proteins that are encoded by autonomous elements. Additionally, we found evidence that gene movements can also be caused when DSBs are repaired after template slippage or unequal crossing-over events. The observed frequency of gene movements can explain the erosion of gene colinearity between plant genomes during evolution.

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

The genes of closely related species are usually found in similar order along chromosomes. This "colinearity" reflects genome descent from a common ancestor. While gene order in animal genomes is well-conserved over hundreds of millions of years (Murphy et al. 2004; Benarafa and Remold-O'Donnell 2005; Li et al. 2010), colinearity in plants blurs more rapidly (Salse et al. 2009). That is, with increasing phylogenetic distance, the number of colinear genes is decreasing. For example, the family of the grasses, which arose ~70 million yr ago (Mya), shows extensive conservation in the linear order of genetic markers, allowing the determination of which chromosomes or chromosome segments correspond between species (Gale and Devos 1998). Corresponding chromosomes or chromosome segments are called "syntenic."

At the molecular level, colinearity of genomic sequences has many exceptions. Between rice and sorghum, only ~57% of genes are found in colinear blocks (Paterson et al. 2009). Between monocotyledonous ("monocots") and dicotyledonous plants ("dicots"), two major plant lineages that diverged ~200 Mya (Wolfe et al. 1989; Yang et al. 1999), only traces of colinearity are detectable (Tang et al. 2010). Intergenic regions and noncoding parts of genes evolve even more rapidly because they are completely reshuffled within just a few million years by transposable element (TE) insertions and deletions caused by illegitimate recombination and unequal crossing-over (Devos et al. 2002; SanMiguel et al. 2002; Wicker et al. 2003a).

With rice (International Rice Genome sequencing Project 2005), sorghum (Paterson et al. 2009), and the recently released genome of *Brachypodium distachyon* (International Brachypodium Initiative 2010), there are now three relatively small and compact high-quality grass genome sequences publicly available. Rice and *Brachypodium* diverged ~40 Mya, while sorghum diverged from the other two ~50 Mya (International Brachypodium Initiative 2010). This allows three-way comparisons of gene order and colinearity and thus the identification of specific gene movement events in *Brachypodium* and rice. For example, if a gene is non-colinear in *Brachypodium* but is found in colinear positions in rice and sorghum, that gene was most likely moved in *Brachypodium*. For sorghum, this approach does not work, as the gene could have been moved already in the common ancestor of *Brachypodium* and rice.

The molecular mechanisms responsible for this "gene movement" have largely been the subject of speculation. Leister (2004) suggested that such "ectopic duplications" originate from recombination between unlinked homologous sequences or from TE activity. Indeed, several studies showed that some TEs occasionally "capture" short fragments of genic sequences and move them across the genome. This phenomenon was described for elements of the superfamilies *Mutator* (Jiang et al. 2004), *Helitron* (Lai et al. 2005; Morgante et al. 2005), *CACTA* (Wicker et al. 2003b; Paterson et al. 2009), *Harbinger* (International Brachypodium Initiative 2010), and LTR retrotransposons (Jin and Bennetzen 1994). The capture of the foreign fragments might be due to readthrough events where neighboring TEs and the fragment between them are joined into one element (Kapitonov and Jurka 2007).

As an alternative mechanism of gene capture in *Helitron* elements, it was suggested that the foreign fragments are introduced as filler DNA for the repair of a double-strand break (DSB) inside a TE (Kapitonov and Jurka 2007). It was recently shown in yeast that one of several mechanisms for repair of DSBs can be the use of foreign filler DNA fragments (Agmon et al. 2009). The underlying molecular mechanism is "synthesis-dependent strand annealing" (SDSA) (Nassif et al. 1994; Puchta 2005; Hartlerode and Scully 2009), where short sequence motifs at the breakpoint serve as a template to invade a foreign DNA strand and initiate strand synthesis. The result is a copy of the filler DNA at the site of the DSB.

Although ectopic duplication, TE-driven exon shuffling, and SDSA have been suggested as important factors in genome

[1] **Corresponding author.**
E-mail wicker@botinst.uzh.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.107284.110.

evolution, there has not been any quantitative assessment of their actual contribution to gene movement.

The objective of this study was to unravel the molecular mechanisms that lead to the movement of functional genes in genomes and the erosion of gene colinearity that ultimately results from it. We performed a three-way genome-wide comparison of *Brachypodium*, rice, and sorghum genes and could show that gene movement is largely a copy-and-paste process in which large genomic fragments are duplicated to other locations in the genome. Many recently duplicated fragments are bordered by transposable elements (TEs) or breakpoints of duplication or unequal crossing-over events. We found highly diagnostic sequence signatures that indicate that the foreign fragments were introduced as "filler DNA" to repair DSBs that occurred upon TE insertion or a recombination event.

Results

Three-way comparison of *Brachypodium*, rice, and sorghum allows identification of genes that were moved specifically in one species

It is not trivial to determine if two genes are in colinear positions in two species (i.e., whether they are still in the position they had in the common ancestor). Many times, colinearity is blurred due to the presence of clusters of duplicated genes (paralogs) or local, sometimes overlapping, inversions. Thus, both the gene's overall location (e.g., similar region on syntenic chromosome arms) as well as its immediate neighbors have to be determined to establish colinearity. For this study, we considered a gene colinear if it is found in a syntenic chromosomal region and four out of its eight closest neighboring genes also have their closest homologs in the same location and order in the other two species.

To minimize the number of annotation artifacts (e.g., TEs that were mistakenly annotated as genes), we used only the 20,468 of the total 25,532 predicted *Brachypodium* genes for which we found genes with homology at the DNA level in both rice and sorghum. Using these stringent criteria, we identified 14,181 genes (69%) that are colinear in all three grass species. Between *Brachypodium* and rice, we found 16,814 genes (82%) to be colinear.

More interestingly, we could use this data set to identify genes that are non-colinear specifically in either *Brachypodium* or rice, but colinear in the two other species. We identified a total of 1406 genes that were non-colinear (i.e., moved) in *Brachypodium* and 1625 genes for which this is the case in rice. Furthermore, having information on colinearity of all genes in the three species, we could determine where each non-colinear gene actually originated from (i.e., the location that contains the homologs of the neighbors of the non-colinear gene in the two other species). We refer to that location as the "donor region" and the current location as the "acceptor site" (an example is shown in Fig. 1).

For 1247 of the 1406 non-colinear *Brachypodium* genes, we could identify such a donor site. Interestingly, in 973 (69.2%) donor sites, we found a homolog of the moved gene, while 274 only contain its neighbors. In rice, we found donor regions for 1415 of the 1625 non-colinear genes, and 1073 (75.8%) contained a homolog of the non-colinear gene. This indicates the phenomenon studied is predominantly a "copy-and-paste" process in plant genomes and that the term "gene movement" is largely a misnomer.

Gene movement appears to be a constant process

For those genes where a homolog in the donor region was identified, we aligned the two coding sequences, if DNA sequence conservation allowed it. The duplicated genes show a wide range of sequence similarity with their donor regions, ranging from near complete identity (98.2%) to ~70%. Many donor/acceptor gene pairs were too divergent as to allow a reliable alignment at the DNA level (DNA identity below ~70%). A total of 643 could be aligned, and the distribution of DNA sequence identity is given in Figure 2. The genes with higher identities are likely representing more recent duplications. Indeed, alignment of these genes from the donor region and acceptor sites plus their flanking genomic regions showed that the duplicated region extends past the coding region (Fig. 2B). In contrast, in ancient duplications, homology was limited to protein-coding regions of genes (Fig. 2C), because intergenic and noncoding sequences have diverged to a degree that they cannot be aligned anymore. In some recent duplications, the duplicated fragment also contained conserved TE sequences (i.e., TEs that inserted before the duplication event). Because TE sequences are largely free from selection pressure (Petrov 2001), they accumulate mutations at a background rate that was estimated to be 1.3×10^{-8} substitutions per site per year (Ma and Bennetzen 2004, see Methods). Therefore, such sequences are especially suitable to estimate how long ago the duplication event occurred. The example in Figure 2D shows a duplicated fragment that contains two genes plus a conserved non-LTR retrotransposon (LINE). The duplicated sequences are 97% identical. Based on the divergence of the LINE element, we estimate that this duplication occurred ~1.3 Mya. This example also illustrates the rapid turnover of intergenic sequences as the duplicated units differ already in several major insertion/deletions, despite their relatively recent divergence (Fig. 2D).

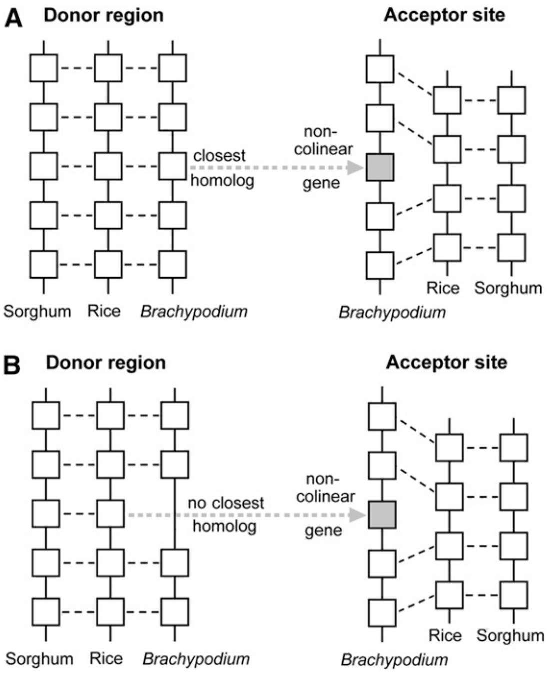

Figure 1. Example of the identification of a non-colinear gene in *Brachypodium* and its putative donor region. Homologs from the three species are connected by dashed lines. A non-colinear gene (acceptor site) has its closest homologs in rice and sorghum in a different genomic region than its neighbors. The locus where the non-colinear gene originated contains colinear neighbors in rice and sorghum. (*A*) A homolog of the moved gene is found in the donor region. The gene was copied to its acceptor site. (*B*) The donor region does not contain a homolog. Apparently, the gene was moved or the homolog in the donor region was deleted later on.

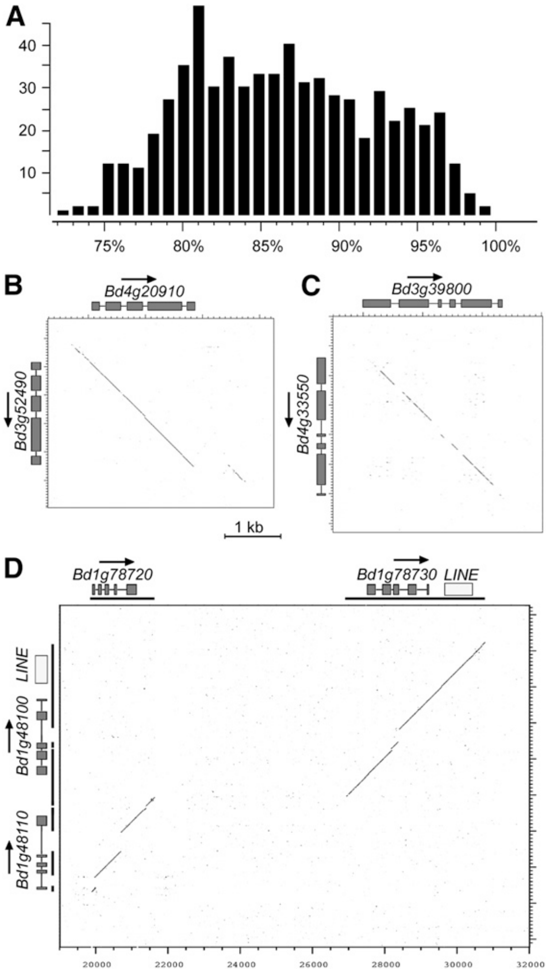

Figure 2. (*A*) Levels of DNA sequence identities between 534 gene copies from the donor region and acceptor sites where DNA sequence conservation allowed a reliable alignment. (*X*-axis) Level of DNA sequence identity; (*y*-axis) the number of gene pairs in that identity range. (*B*) DotPlot alignments of donor region and acceptor site. In recent duplications, the gene and its flanking regions are well conserved, and the borders of the duplicated region can clearly be identified. (Gray boxes) Exons of the predicted gene. (*C*) With time, duplicated regions diverge and sequence conservation is basically limited to the protein-coding exons. (*D*) DotPlot alignment of a duplicated fragment containing two genes. The duplicated fragments are indicated as black bars underneath the genes. The interruptions in the conserved parts are caused by insertions or deletions that occurred after the duplication. Molecular dating of the conserved LINE (white box) indicates that the duplication occurred ~1.3 Mya.

Most donor–acceptor gene pairs were 80%–85% identical (Fig. 2A). We believe that this range of sequence identity represents the "steady state" that is reached after a few million years of evolution, because some DNA sequence conservation is essential in protein-coding regions, while promoters, introns, and downstream regions evolve more rapidly. These data indicate that gene movement is a constantly occurring process.

The acceptor site triggers gene duplication

To study the possible mechanisms that led to the duplication of the genes and their integration elsewhere in the genome, we focused on high-scoring pairs where donor and acceptor gene sequences are at least 95% identical. For these, we extracted (in silico) donor region and acceptor sites including 25 kb of their flanking sequences from the genomes. Corresponding acceptor sites and donor regions were aligned to identify the precise borders of the duplicated fragments. Suspecting that TEs might be involved in the gene duplication process, we focused this analysis mainly on the 58 high-scoring pairs from *Brachypodium*, because of the low repeat content of the *Brachypodium* genome and the high quality of TE annotation. For comparison, we randomly picked 42 high-scoring pairs from rice (to bring the total number of analyzed regions to 100). The size of the duplicated fragments ranged from 114 bp to 50.6 kb with an average size of 6.4 kb (Fig. 3).

Special attention was paid to the precise borders and immediate flanking sequences of the duplicated region. Curiously, we found TE sequences or other peculiar motifs immediately flanking the duplicated region only at the acceptor sites. This led to the hypothesis that the cause for the gene movement was to be found at the acceptor and not at the donor site.

Transposable element-driven gene capture and retroposition do not explain gene movement in plants

We first searched for the previously described events of gene capture by TEs (Wicker et al. 2003b; Jiang et al. 2004; Lai et al. 2005; Morgante et al. 2005; Paterson et al. 2009). In the 100 high-scoring pairs, we identified 10 cases from *Brachypodium* and 20 cases from rice where the duplicated gene sequence was clearly captured by a TE (i.e., located inside the borders of a the TE that was flanked by a target site duplication). In *Brachypodium*, they were associated with *Mutator*, *Helitron*, and *Harbinger* elements, while in rice all but three gene fragments were captured by *Mutator* elements (Table 1). In both *Brachypodium* and rice, we found multiple examples where the entire TE including the captured fragment was present in multiple copies in the genome. Additionally, for most of these TEs we also found very similar elements (>80% identical) that did not contain the foreign gene. For example, for the *Mutator* element that captured the gene fragment *Os2g15050*, we found nine copies that contained the fragment and 20 that did not. This demonstrates that the respective TEs continued to proliferate after capturing the foreign fragment. Most captured foreign fragments were <1000 bp and only one was >2000 bp (Fig. 3). The captured

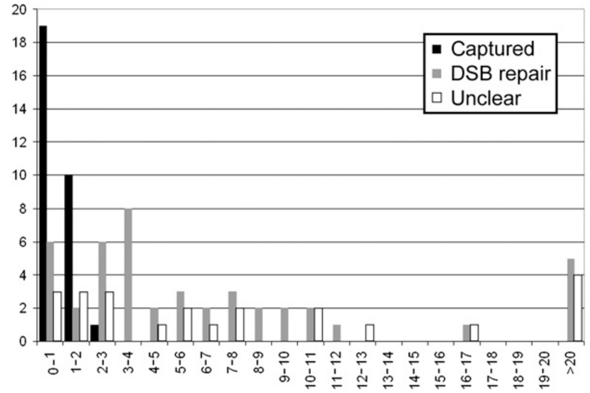

Figure 3. Size distribution of 100 duplicated fragments, originating from recent duplications. The fragments captured by transposable elements and moved across the genome are mostly <1 kb in size and none contained an entire gene. Fragments that were presumably duplicated during DSB repair are usually several kilobases in size. Those duplication events for which no specific signatures could be identified show a similar size distribution as the ones triggered by DSB repair.

Table 1. Gene capture events in *Brachypodium* and rice

Superfamily	*Brachypodium*	Rice
Mutator	2	17
Harbinger	1	0
Helitron	4	2
Unclassified[a]	3	1

[a] Found in several copies, contains terminal repeats and/or target site duplication, but has no homology with known transposable element superfamilies.

fragments covered only parts of genes and none represented a full-length gene.

In animals, retroposition of mRNA is an important source of novel and non-colinear genes (for review, see Kaessmann et al. 2009). Since these genes are derived from mRNA, they usually do not contain introns. We aimed at estimating the contribution of retroposition to gene movement in plants by comparing the number of exons of gene copies at the acceptor site with the one in the donor region. In total, 46.4% of donor/acceptor gene pairs had the same number of predicted exons. Slightly more genes (33.3%) had more predicted exons in the donor region, while 20.2% had more exons at the acceptor site. The number of single-exon genes was virtually identical for donor (20.7%) and acceptor genes (20.9%).

From these data, we concluded that neither TE-driven gene capture nor retroposition alone are a sufficient explanation for gene movement in plants. From here, we proceeded to study in detail the borders of the duplicated regions in the remaining 70 high-scoring pairs.

Large duplicated fragments are often flanked by TEs

The clues about the nature of the gene movement mechanism came from 27 high-scoring donor/acceptor pairs. In these donor/acceptor pairs, one border of the duplicated region at the acceptor site was precisely or within a few base pairs of one end of a TE. We consider it extremely unlikely that this would happen by chance. In four cases, we found the target site duplication (TSD) that typically flanks the TE at the other border of the duplicated region (Fig. 4B,C,E,F), while in two cases, the TSD is flanking the TE but also overlaps with the duplicated region (Fig. 4A,D). Additionally, 17 cases were identified where a truncated TE fragment borders the duplicated region (examples in Supplemental Fig. 1). Interestingly, we found TEs belonging to multiple superfamilies including *CACTA*, *hAT*, *Harbinger*, *Mariner*, *Mutator*, *Helitron*, and SINE at the breakpoints of the moved regions.

Several additional TEs found at the breakpoints of the moved regions could not be classified into known superfamilies. They were considered to be TEs because they were present in multiple copies in the genome, flanked by a TSD, and/or contained terminal inverted repeats. Some of these elements have only a very low copy number (three to 20) in the genome. For none of them could we identify autonomous copies (e.g., copies that contain the genes necessary for transposition or replication), indicating that they rely entirely on enzymes encoded by other TE families for their transposition.

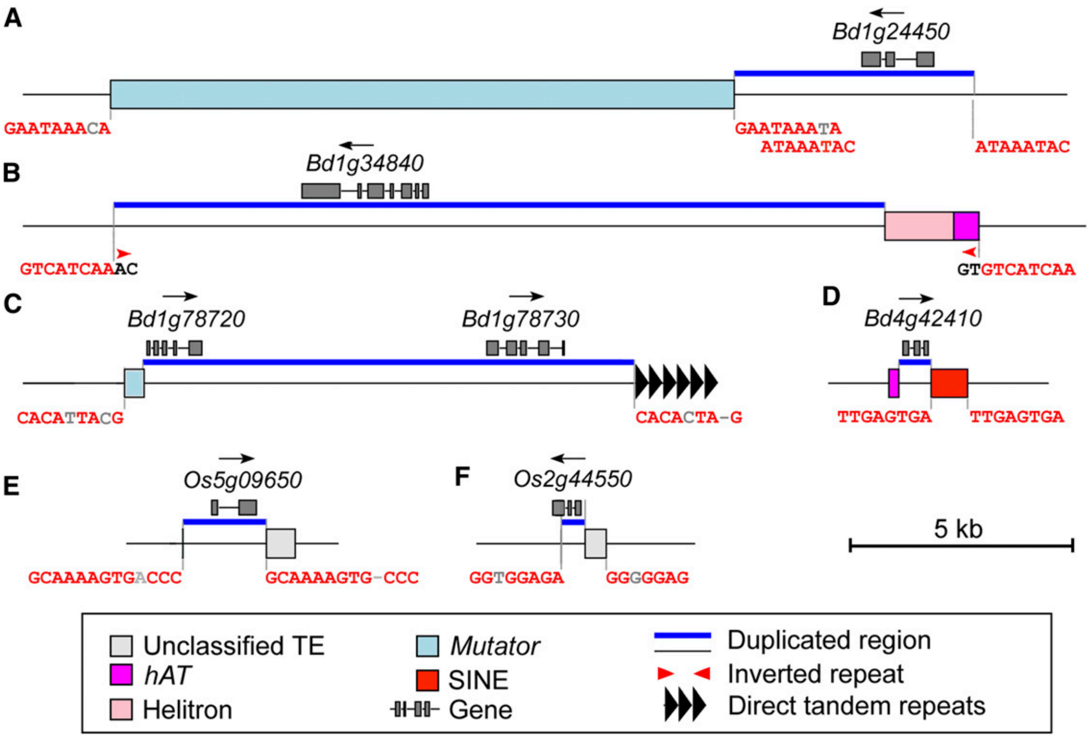

Figure 4. Molecular signatures that point to DSB repair as the cause of gene movement. (Red) Target site duplications (TSDs) that were produced upon the insertion of transposable elements; (gray) mismatches. The double-strand break (DSB) induced by the transposable element insertion occurred within a few base pairs of the terminus of the transposable element. A model for the molecular events that led to the situation in *A* is given in Figure 6. Depending on the precise location of the DSB, TSDs are flanking the transposable element (*C–E*), the duplicated region (*B,F*), or both (*A*). The example in *C* is flanked on one side by an array of direct repeats. It is possible that a combination of the TE insertion and the presence of the repeat array lead to the DSB (see also Fig. 5).

We found no correlation between the transcriptional orientation of the genes in the duplicated fragments and the orientation of the TE flanking it. In fact, nine of the 16 duplicated fragments in the high-scoring pairs that contained multiple genes had genes in the forward as well as reverse orientation. We also found no bias as to whether the filler DNA is inserted at the 5′ or 3′ end of the TE.

Double-strand breaks are the likely trigger of gene movement

Based on the above findings we propose a model where the fragments containing the foreign genes were used as filler DNA to repair DSBs. We propose that in most cases DSBs have occurred upon insertion of TEs. In the few cases where a TSD was still conserved, molecular events can be postulated in detail like in the case of gene *Bradi1g24450* (Fig. 5): In a first step, a *Mutator* element inserted into the genome. The transposase cutting the host DNA created the typical 9-bp overhangs bordering the termini of the element. Usually these gaps would be filled by cellular DNA repair enzymes, resulting in the characteristic TSD. We assume that during this process, a DSB can occur either precisely at the insertion point or a few base pairs away from it. The 3′ overhang produced by the transposase invades a complementary motif elsewhere in the genome. A filler strand is synthesized until a second matching motif is reached. The result is that the filler DNA is immediately adjacent to the TE insertion. Apparently, matching motifs of only a few base pairs in size are sufficient for strand invasion and priming of synthesis (Nassif et al. 1994).

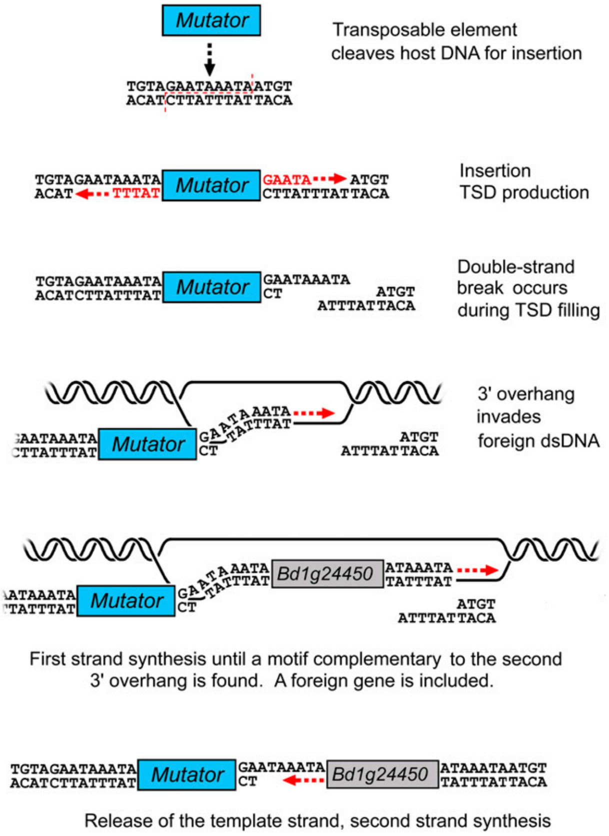

Figure 5. Model for molecular events that led to a duplication of gene *Bd1g24450*. A DSB is introduced after the insertion of a *Mutator* element in the genome. A sequence fragment from elsewhere in the genome containing gene *Bd1g24450* is used as filler DNA to repair the DSB.

There seems to be a fair bit of variation as to where exactly the DSB occurs. This may depend on whether the DSB that occurs during the TSD filling reaction generates blunt ends, 5′ overhangs, or 3′ overhangs. Only if the 3′ overhangs produced upon insertion of the TE are preserved will the TSD signature still be identifiable on one or both sides of the filler DNA (Fig. 5). In the case of gene *Bradi1g34840*, the breakpoint is actually 2 bp into the terminal inverted repeat of the *hAT* element, as the end of the element plus the TSD is found at the opposite side of the duplicated region (Fig. 4B).

Causes for DSBs are manifold

Previous studies showed that DSBs occur frequently in so-called "fragile sites" (Debrauwère et al. 1999; for reviews, see Pfeiffer et al. 2000; Lovett 2004). These are regions consisting of tandem repeated motifs such as micro- or minisatellites, which are hotspots of recombination either by unequal crossing-over or template slippage. Indeed, we identified 10 cases where the duplicated gene was on one or both sides flanked by direct repeats. Examples are given in Figure 6, A and B. Some are tandem arrays with repeat unit sizes of several hundred base pairs, while others have only a few-base-pair unit size (Fig. 6). The duplicated fragment that contains rice gene *Os3g30240* is located inside an array of tandem repeats, three units on its left and five units on its right side (Fig. 6B). The individual repeat units are a few dozen base pairs in size and very GC-rich. We assume that during a template slippage or unequal crossing-over event, a DSB occurred that was then repaired with the foreign fragment containing the gene. The duplicated fragment that contains genes *Bradi78720* and *Bradi1g78230* is flanked on one side by a *Mutator* element and on the other side by a large array of direct repeat units of 110 bp in size (Fig. 6B). We suggest that the DSB was caused by the insertion of the *Mutator* element into a region that is inherently unstable.

The duplicated fragment containing gene *Bradi2g19950* is flanked on both sides by a duplication of ∼1100 bp (Fig. 6C). We hypothesize that the DSB occurred during a replication slippage event between two 14-bp repeats that are separated by a little over 1100 bp. As described by Lovett (2004), such events are a common source of duplications. We propose that resolution of the slippage intermediate caused the DSB. The short motifs that served as a template for the recombination can still be found flanking the duplicated region (Fig. 6C), and the region containing gene *Bradi2g19950* lies precisely at the breakpoint of one direct repeat. The direct repeat unit at the right side is not immediately adjacent to the border of the duplicated region containing *Bradi2g19950*, suggesting a template switch during the filler synthesis (Fig. 6C). The formation of such "patchwork" filler sequences is observed frequently in plants (Gorbunova and Levy 1999). A detailed model for a possible sequence of events is proposed in Supplemental Figure 2.

There are many known sources that cause DSBs in plants including TE excisions, radiation, and reactive oxygen species (for review, see Pfeiffer et al. 2000). These can, in principle, occur at any position in the genome. Repair of such DSBs with gene-containing fragments as fillers could explain those cases where we found no particular sequence motifs (i.e., TE sequences) flanking the duplicated fragment.

Additionally, our data suggest that the mechanism of gene capture by TEs itself might be the result of DSB repair. We compared *Mutator* element copies with and without captured fragments. All members of the analyzed families have a very GC-rich

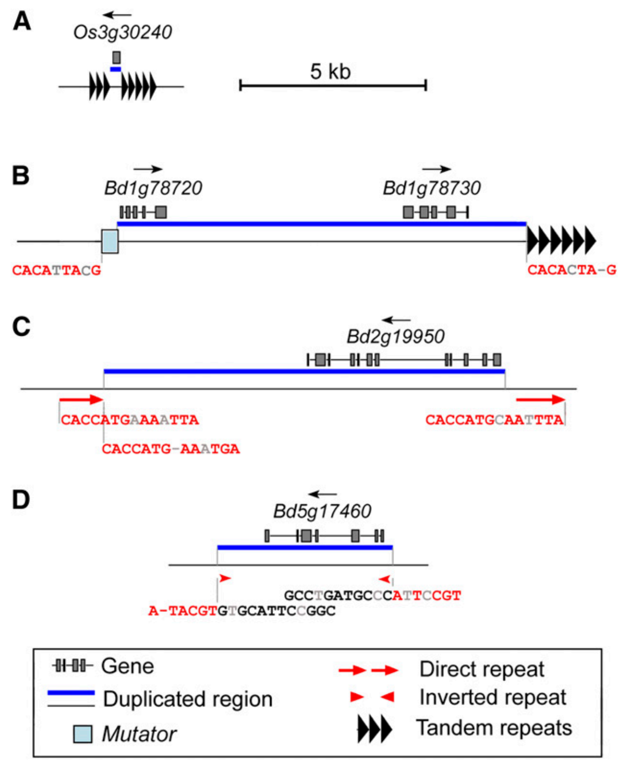

Figure 6. Examples of gene movements that were triggered by mechanisms other than transposable element insertions. (*A*,*B*) The duplicated region is flanked on one or both sides by an array of tandem repeats. The DSB was possibly introduced during a template slipping or unequal crossing-over event. The example in *B* is flanked on one side by a *Mutator* element. It is possible that a combination of the TE insertion and the presence of the repeat array led to the DSB (see also Fig. 4). (*C*) A duplication caused by a replication slippage event led to a DSB that was then repaired with the foreign fragment. (Red) Signature sequences that presumably served as templates for the template slippage; (gray) mismatches (see also Supplemental Fig. 2). (*D*) An accidental inverted repeat structure (black/gray letters) flanking the gene was possibly recognized by a "stray" transposase. The inserted fragment is flanked by a TSD (red/gray letters).

internal domain with many low-complexity motifs at the border of the captured fragment (Supplemental Fig. 1). We speculate that, whatever primary function these regions have for the transposon, they may accidentally trigger the uptake of foreign DNA sequences when DSBs occur because of template slippage or unequal crossing-over.

Actual movement of genes is rare

We suspect that the apparent movement of a gene (i.e., the absence of a homolog in the donor region), is most often the result of deletions in the donor region after the duplication event. A good example is the gene movement that led to the duplication of a large fragment containing up to seven rice genes (*Os3g37960* through *Os3g38020*). Based on a conserved intergenic sequence located between *Os3g37984* and *Os3g38010*, we calculated that this duplication occurred ~5.5–6.6 Mya. Comparison of the donor regions with the acceptor site shows that numerous deletions and insertions occurred in both the donor and the acceptor region since the original duplication (Fig. 7). This resulted in the homolog of gene *Os3g37984* being completely absent from the donor site and the homolog of *Os3g38020* being partially deleted, thus creating the impression that some genes were actually moved rather than duplicated (Fig. 7).

Gene *Bradi5g17460* is the only case identified where the cause for the gene movement might be in the donor region. The donor gene is flanked by short inverted repeats. The acceptor site still contains those motifs but is also flanked by a putative TSD (Fig. 6D). The motifs are degenerate but are located precisely at the border of the moved region. We propose that these inverted repeats (which might not be of TE origin at all) have served as a target for a transposase encoded by a TE elsewhere in the genome. An explanation for the presence of a homolog of *Bradi5g17460* in the donor region could be that the parent chromosome that still contained the original gene was passed on to the offspring while the donor chromosome where the gene was excised was lost.

Discussion

The cause of gene movement and the erosion of gene colinearity that goes with it has been an unsolved riddle since the advent of comparative genomics. One reason is that intergenic regions evolve rapidly because they are largely free from selection pressure (Petrov 2001). Thus, diagnostic sequence motifs such as TSDs of TEs or the precise borders of the duplicated fragments are not recognizable anymore after a few million years of evolution (San-Miguel et al. 2002; Wicker et al. 2003a; Ma and Bennetzen 2004). Only by conducting genome-wide comparisons of three genomes were we able to focus specifically on the most recent gene movement events to identify the telltale signatures. Even in the entire *Brachypodium* genome, there were only a few dozen such cases. Nevertheless, these were sufficient to provide the highly informative sequence motifs that point to gene movement as the result of DSB repair. Our data indicate that DSBs can be caused by multiple mechanisms, the most predominant one being the insertion of TEs.

The progenitor of the grasses underwent a whole-genome duplication ~70 Mya (Salse et al. 2009). This was followed by a "diploidization" process in which differential gene loss occurred, resulting in a grass ancestor with nearly the same number of chromosomes as the original diploid. Therefore, apparent movement of genes in the three species studied might just be the result

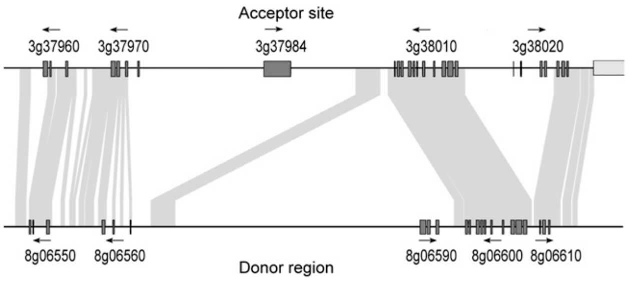

Figure 7. Example for differential deletion of genes after duplication. The fragment that originally contained six genes was duplicated 5.5–6.5 Mya. This estimate is based on comparison of a conserved intergenic sequence between *Os3g37984* and *Os3g38010*. (Gray) Conserved regions. Deletions after the initial duplication event in one or the other sequence lead to the complete or partial elimination of genes in either the donor or the duplicated region, creating the impression that some genes were actually moved rather than duplicated. For example, the homolog of gene *Os3g37984* is completely absent from the donor site while the homolog of *Os3g38020* is partially deleted.

of differential gene loss in different species. However, this can be excluded for two reasons: First, Paterson et al. (2009) showed by whole-genome comparisons that diploidization was practically complete by the time sorghum and rice diverged. Second, loss of a gene in either *Brachypodium* or rice would not be registered in our approach as we required the unique appearance of a gene in a new locus in one of the two species.

Translocation and duplications of genomic sequences must occur frequently

The observed level of gene movement is most likely only the tip of the iceberg, as we studied only the movement of genes that remained intact and became fixed during evolution. As our approach for the identification of gene movements was based on functional genes, we cannot draw any general conclusion on the actual number and frequencies of events that duplicate or translocate genomic sequences. In fact, one has to assume that the number of such events is much higher, but only a very small fraction will eventually become fixed in the form of translocated functional genes. This is for three reasons: First, many events are most likely deleterious, for example, if a DSB occurs in an exon of an essential gene and the insertion of filler DNA disrupts that gene. Such events will be immediately selected against. Second, many duplications and translocations may affect sequences that are not vital for the organism such as TEs. In these cases, the result is evolutionarily completely neutral, and all traces of such events are removed after a few million years of evolution through genomic turnover. Thus, they cannot be traced back by comparing species that diverged dozens of millions of years ago. Third, repair of DSBs often leads to deletions (Kirik et al. 2000). Again, if the deletion is deleterious, it will be selected against; if it is neutral, no traces will be left of the event.

In conclusion, the only events that can be recognized after long evolutionary periods are those in which the initial event was not deleterious and an intact gene that is under selection pressure was duplicated or moved.

About 25% of donor sites did not contain a homolog of the moved genes. Thus, the term "gene movement" is somewhat misleading because the absence of a gene in the donor region is probably in most cases the result of a deletion that happened after the initial duplication. These could be cases where there was no neofunctionalization and no selection for two copies. We found only one case (*Bradi5g17460*) that suggests that the gene was indeed moved through excision by a transposase.

Gene movement seems to be triggered mostly by low-copy transposable elements

Interestingly, we found TEs from multiple superfamilies associated with gene movements, suggesting that, in principle, all TEs can lead to gene movement events. Curiously, we did not find any LTR retrotransposons associated with such events. This is surprising because LTR retrotransposons are the most abundant TEs in all three genomes (International Rice Genome Sequencing Project 2005; Paterson et al. 2009; International Brachypodium Initiative 2010). In fact, most TE families we identified associated with gene movements have a very low copy number (three to 20). It is likely that elements that have a higher tendency to introduce DSBs are selected against because of the decrease in fitness of the host and are thus present only in very few copies. In contrast, high-copy elements such as LTR retrotransposons might have more sophisticated insertion mechanisms that cause less damage so that they can be tolerated in much greater numbers. An alternative explanation could be that LTR retrotransposons are usually most concentrated in centromeric regions while genes are mostly found in distal chromosomal regions. Thus, LTR retrotransposons would be less likely to insert into or near genes. However, in most plant genomes, many high-copy LTR retrotransposons are still found interspersed with genes.

Additionally, it seems that several of the mobile elements associated with gene movement are the result of accidental transposition of non-TE sequences. Indeed, most of them had fewer than five copies in the genome, and we did not find any autonomous elements of the respective type (i.e., copies that contain the genes necessary for transposition). And yet many were flanked by terminal inverted repeats and/or TSDs. We speculate that these elements actually originate from unspecific transposition events where "stray" transposases bind to sequence motifs that by chance resemble their targets. This is a novel and somewhat disconcerting aspect, because it implies that in addition to the main evolutionary TE lineages that have been around for hundreds of millions of years, completely new types of mobile DNA elements can occasionally spontaneously arise and vanish again. Such elements might be particularly deleterious and cause a disproportional number of DSBs because they do not contain sequence motifs that allow a proper excision or integration into the genome. An alternative explanation for the absence of autonomous elements is that the autonomous elements are no longer present in the accession analyzed.

Gene movement and exon shuffling are two distinct phenomena

We propose that TE-mediated gene capture plays only a minor role in gene movement. Although, as suggested by Kapitonov and Jurka (2007) and supported by our findings, some gene capture events might actually be the result of DSB repair, we found that only relatively small gene fragments and no entire genes were captured. It is possible that TEs are not able to transpose and proliferate when capturing large fragments. Thus, TE-driven gene capture may well contribute to exon shuffling as previously reported (Wicker et al. 2003b; Jiang et al. 2004; Lai et al. 2005; Morgante et al. 2005; Paterson et al. 2009), but the relatively small size of the captured fragments makes it unlikely that entire genes are moved. Therefore, the exact impact of gene capture on genome evolution remains unclear.

Interestingly, gene capture events appear to be much more frequent in rice than in *Brachypodium* (20 out of 42 vs. 10 out of 58). This may be a consequence of the differences in genome size. Indeed, also in sorghum and maize with their larger genomes, gene capture is found frequently (Paterson et al. 2009; Schnable et al. 2009). Also, retroposition of genes, a major mechanism for the creation of new and/or non-colinear genes in animals (Kaessmann et al. 2009), seems to be of less importance in plants.

Conclusion and outlook

According to our analysis, ~3600 genes are not colinear between rice and *Brachypodium*. Through three-way comparison of grass genomes that diverged within the past 50 Myr, we were able to trace back the fate of over 2400 of these genes. Gene movement seems to go on at very similar rates in both species, with rice having a slightly higher rate. If the observed pace of gene movement is constant, ~20% of all genes between two plant species would escape colinearity within 40 Myr of evolution. Thus, colinearity would be almost completely eroded within 200 Myr of evolution. In fact, this is what is observed between monocotyledonous and dicotyledonous plants (Tang et al. 2010).

It remains to be determined why, in animal genomes, genes remain at colinear positions for longer evolutionary periods. We found that many duplicated fragments in the studied grass genomes have sizes of several kilobases, large enough to carry most plant genes, which tend to be compact. Animal genes usually have longer introns and are spread out over much larger genomic regions. It is possible that gene size is the reason for the less frequent gene movement in animal genomes. Alternatively, the large genes could be an adaptation that prevents animal genes from being moved as frequently as plant genes.

Methods

Identification of closest gene homologs in *Brachypodium*, rice, and sorghum

FASTA files of all predicted coding sequences (CDS) of genes from *Brachypodium*, rice, and sorghum were kindly provided by the Institute for Bioinformatics and Systems Biology (MIPS, http://www.helmholtz-muenchen.de/en/mips/). Because multiple gene models exist for most genes, we created a condensed data set that contained only the largest gene model for each gene. The closest homologs of genes in *Brachypodium*, rice, and sorghum were identified by BLASTN. Since gene movement specific for one species could only be identified in rice and *Brachypodium*, BLAST searches were done with the *Brachypodium* and rice CDS data sets as follows: Each CDS was used as query in a BLASTN search against the CDS data set of both other species. BLASTN hits with *E*-values $< 10 \times 10^{-10}$ were recorded. The final result was one table each for *Brachypodium* and rice genes that contained the gene identifiers of the closest homologs from both other species as well as their base-pair position on the respective chromosomes and the description line of the gene.

Identification of colinear and non-colinear genes

Genes that are colinear in all three species were identified in a two-step analysis. First, a Perl script called gene_movement_mk_colinear_gene_list used what we refer to as the "ancestor chromosome index" (Supplemental Table 1): The 12 chromosomes of rice served as a reference chromosome numbering system, reflecting the structure of the hypothetical common ancestor of the grasses (Salse et al. 2009). For example, *Brachypodium* chromosome 1 is the result of two nested chromosome fusions that brought together rice chromosomes 3, 6, and 7 (International Brachypodium initiative 2010). Whenever possible, the same number as the ancestor chromosome was assigned to the chromosomal region in *Brachypodium* and sorghum (Supplemental Table 1).

This first Perl script created a table with information on whether genes are in syntenic chromosomes. However, this table still included a few genes that have moved within the same ancestral chromosome. The program gene_movement_mk_colinear_genes_list2 checked all colinear genes from this table and their four immediate neighbors on both sides. If there were at least four genes with gene numbers that differ by less than 200 in *Brachypodium* and 400 in rice, the gene is considered truly colinear (genes are generally numbered in intervals of 10). We allowed the larger range in rice because the annotation of the rice genome contains more artifacts (e.g., TEs or pseudogenes) than the one of *Brachypodium*. The same Perl script was used to identify genes that were specifically non-colinear in either *Brachypodium* or rice (an indication that a gene has moved specifically in either *Brachypodium* or rice).

Characterization of donor regions and acceptor sites

We then searched for a locus that contains the next neighbors of a non-colinear *Brachypodium* gene in rice or Sorghum. This should be the putative donor region of the moved gene. The program Bdis_gene_movement_donor_sites identified the closest neighbors of the non-colinear genes and checked if the whole group lies on a chromosome colinear between rice and *Brachypodium* and sorghum. The criterion was that genes are from colinear chromosomes and that gene numbers differ by less than 100 from that of the gene movement candidate. The same procedure was carried out vice versa with non-colinear rice genes.

If a homolog of the duplicated gene was found in the donor region, the two CDS were aligned with the program WATER from the EMBOSS package (http://emboss.sourceforge.net/), and high-scoring donor–acceptor gene pairs (>95% DNA identity) were selected for further analyses. For all high-scoring pairs, a Perl program was written that specifically excised (in silico) donor regions and acceptor sites from the chromosomes for further analysis. We extracted (in silico) donor region and acceptor sites including 25 kb of their flanking sequences from the genomes. Corresponding acceptor sites and donor regions were aligned with DOTTER (Sonnhammer and Durbin 1995) to identify the precise borders of the duplicated fragments.

Identification of TE sequences flanking the duplicated regions

The flanking 500 bp on both sides of the duplicated region were used as queries against databases containing repeat sequences of *Brachypodium* and rice, respectively, in order to identify known transposable elements. If the flanking regions of the duplicated segment did not contain known repeats, they were used in BLAST searches against the respective genome. This was done to identify possible additional copies of thus far unknown transposable elements in the flanking regions. For low-copy elements, all additional copies plus 5 kb of flanking sequences were excised from the genome (in silico). This allowed us to determine the borders of the novel transposable elements by DotPlot alignment of the different copies. For high-copy elements, the top 15–20 copies were used. After having characterized transposable element sequences flanking the duplicated fragment, target site duplications on either side of the transposable element and the duplicated fragment were searched with DOTTER. Novel TEs were identified based on the following criteria: First, the sequence has to be present in multiple copies in the respective genome with at least 80% sequence identity at the DNA level. Second, structural features such as terminal inverted or direct repeats and/or target site duplications have to be present. Additionally, for *SINE*s, a poly-A tail served as diagnostic feature.

Molecular dating of transposable elements that were conserved in the donor and acceptor sequence was done according to SanMiguel et al. (1998) and Wicker et al. (2003a), applying a synonymous substitution rate of 1.3×10^{-8} (Ma and Bennetzen 2004).

The contribution of retroposition was estimated by comparing the number of exons of gene copies at the acceptor site with the one in the donor region. In case of multiple gene models, we compared the largest ones.

Acknowledgments

This study was financed by the Swiss National Science Foundation Grants 3100A0-122242 and 3100A-127061/1.

References

Agmon N, Pur S, Liefshitz B, Kupiec M. 2009. Analysis of repair mechanism choice during homologous recombination. *Nucleic Acids Res* **37:** 5081–5092.

Benarafa C, Remold-O'Donnell E. 2005. The ovalbumin serpins revisited: Perspective from the chicken genome of clade B serpin evolution in vertebrates. *Proc Natl Acad Sci* **102:** 11367–11372.

Debrauwère H, Buard J, Tessier J, Aubert D, Vergnaud G, Nicolas A. 1999. Meiotic instability of human minisatellite CEB1 in yeast requires DNA double-strand breaks. *Nat Genet* **23:** 367–371.

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

Gale MD, Devos KM. 1998. Comparative genetics in the grasses. *Proc Natl Acad Sci* **95:** 1971–1974.

Gorbunova VV, Levy AA. 1999. How plants make ends meet: DNA double-strand break repair. *Trends Plant Sci* **4:** 263–269.

Hartlerode AJ, Scully R. 2009. Mechanisms of double-strand break repair in somatic mammalian cells. *Biochem J* **423:** 157–168.

International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463:** 763–768.

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

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569–573.

Jin Y-K, Bennetzen JL. 1994. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell* **6:** 1177–1186.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat Rev Genet* **10:** 19–31.

Kapitonov VV, Jurka J. 2007. Helitrons on a roll: Eukaryotic rolling-circle transposons. *Trends Genet* **23:** 521–529.

Kirik A, Salomon S, Puchta H. 2000. Species-specific double-strand break repair and genome evolution in plants. *EMBO J* **19:** 5562–5566.

Lai J, Li Y, Messing J, Dooner HK. 2005. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. *Proc Natl Acad Sci* **102:** 9068–9073.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet* **20:** 116–122.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463:** 311–317.

Lovett ST. 2004. Encoded errors: Mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* **52:** 1243–1253.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci* **101:** 12404–12410.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by *Helitron*-like transposons generate intraspecies diversity in maize. *Nat Genet* **37:** 997–1002.

Murphy WJ, Pevzner PA, O'Brien SJ. 2004. Mammalian phylogenomics comes of age. *Trends Genet* **20:** 631–639.

Nassif N, Penney J, Pal S, Engels WR, Gloor GB. 1994. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Mol Cell Biol* **14:** 1613–1625.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457:** 551–556.

Petrov DA. 2001. Evolution of genome size: New approaches to an old problem. *Trends Genet* **17:** 23–28.

Pfeiffer P, Goedecke W, Obe G. 2000. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis* **15:** 289–302.

Puchta H. 2005. The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution. *J Exp Bot* **56:** 1–14.

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C. 2009. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci* **106:** 14908–14913.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The palaeontology of intergene retrotransposons of maize. *Nat Genet* **20:** 43–45.

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A^{m}. *Funct Integr Genomics* **2:** 70–80.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The *B73* maize genome: Complexity, diversity, and dynamics. *Science* **326:** 1112–1115.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–GC10.

Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci* **107:** 472–477.

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci* **86:** 6201–6205.

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu ZD, Dubcovsky J, 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, Keller B. 2003b. *CACTA* transposons in Triticeae—a diverse family of high-copy repetitive elements. *Plant Physiol* **132:** 52–63.

Wicker T, Yahiaoui N, Keller B. 2007. Contrasting rates of evolution in *Pm3* loci from three wheat species and rice. *Genetics* **177:** 1207–1216.

Yang YW, Lai KN, Tai PY, Li WH. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J Mol Evol* **48:** 597–604.

Received March 4, 2010; accepted in revised form June 1, 2010.