Letter

Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization

Xiyin Wang,[1,2] Haibao Tang,[1,3] John E. Bowers,[1] and Andrew H. Paterson[1,3,4]

[1]*Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA;* [2]*College of Sciences, Hebei Polytechnic University, Tangshan, Hebei 063000, China;* [3]*Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA*

Whole-genome duplication produces massive duplicated blocks in plant genomes. Sharing appreciable sequence similarity, duplicated blocks may have been affected by illegitimate recombination. However, large-scale evaluation of illegitimate recombination in plant genomes has not been possible previously. Here, based on comparative and phylogenetic analysis of the sequenced genomes of rice and sorghum, we report evidence of extensive and long-lasting recombination between duplicated blocks. We estimated that at least 5.5% and 4.1% of rice and sorghum duplicated genes have been affected by nonreciprocal recombination (gene conversion) over nearly their full length after rice–sorghum divergence, while even more (8.7% and 8.1%, respectively) have been converted over portions of their length. We found that conversion occurs in higher frequency toward the terminal regions of chromosomes, and expression patterns of converted genes are more positively correlated than nonconverted ones. Though converted paralogs are more similar to one another than nonconverted ones, elevated nucleotide differences between rice–sorghum orthologs indicates that they have evolved at a faster rate, implying that recombination acts as an accelerating, rather than a conservative, element. The converted genes show no change in selection pressure. We also found no evidence that conversion contributed to guanine-cytosine (GC) content elevation.

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

Genetic recombination is important for DNA repair and for crossovers between homologous sequences. As a driving force of biological evolution, it is a major source of genetic novelties, such as new alleles and combinations of alleles (Puchta et al. 1996), which may permit adaptation to environmental changes. During meiosis, homologous chromosomes may recombine reciprocally, while during mitosis in somatic cells recombination can be induced by DNA damage. However, recombination, especially illegitimate recombination between paralogous loci, may produce severe chromosomal lesions characterized by various DNA rearrangements, which are often deleterious, but may also contribute to elimination of deleterious mutations (Khakhlova and Bock 2006). In plants, both meiotic and mitotic recombination outcomes can be transferred to the offspring, due to the lack of a predetermined germline. Paralogous recombination can be reciprocal or nonreciprocal. Reciprocal recombination involves symmetrical exchange of genetic information between paralogous loci. Nonreciprocal recombination involves unidirectional transfer of information from one locus to its paralogous counterparts, resulting in gene conversion (Datta et al. 1997).

In model organisms, much research has been performed to better understand how sequence divergence affects the frequency of recombination. Sequence divergence may limit the frequency, length, and stability of early heteroduplex intermediates formed during recombination, dramatically reducing the recombination frequency (Stambuk and Radman 1998). Research with a reporter system in *Arabidopsis* indicated that, relative to the homologous sequences, there was a fourfold to 20-fold decrease in the recombination frequency in lines with constructs containing 0.5%–9% sequence divergence (Li et al. 2006). In maize, the bronze (*bz*) gene is a recombination hotspot, and analysis of meiotic recombination between heteroallelic pairs of *bz* mutations reveals both insertion mutation and sequence divergence to affect the distribution of intragenic recombination events (Dooner and Martinez-Ferez 1997). Adjacent retrotransposons abruptly decrease recombination rates in the *bz* locus (Fu et al. 2002). With seven tomato lines, recombination frequency at two adjacent intervals on chromosome 6 was characterized (Liharska et al. 1996). When the entire chromosomal arm of tomato (*Lycopersicon esculentum*) was replaced with chromatin of *Lycopersicon pimpinellifolium*, a related species, or with that of *Lycopersicon peruvianum*, a relatively distant species, up to a sixfold decrease in recombination frequency was observed.

The vast quantity of duplicated sequences in plants produces many opportunities for nonreciprocal recombination or gene conversion. Paleopolyploidy is one of the main sources of duplicated sequences in plants. With the accumulation of genome sequences, recurrent polyploidies were uncovered in many plant genomes (Tang et al. 2008a). It was inferred that most if not all angiosperms were affected by whole-genome duplication (WGD) (Bowers et al. 2003). A WGD ~ 70 million years ago (Mya) is common to key cereals, including the sequenced grasses, rice, and sorghum, which diverged ~20 My after the WGD (Paterson et al. 2004). Soon after WGD, multiple homologous chromosomes could compete to pair and recombine with one another. "Genome turmoil" including massive DNA loss and restructuring (Bowers et al. 2005; Wang et al. 2005) might contribute to divergence among once-homologous chromosomes, perhaps also causing chromosomal rearrangements (Paterson et al. 2004).

[4]Corresponding author.
E-mail paterson@uga.edu; fax (706) 583-0160.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.087288.108.

19:1026–1032 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

This may typically lead to diploidization, with neo-homologous chromosomes being formed. Thereafter, recombination could occur mainly between the neo-homologous chromosomes, while being restricted between paralogous chromosomes/chromosomal segments due to previous rearrangements, and gradual sequence divergence. Nonetheless, recombination, literally termed as illegitimate, might still infrequently occur between paralogs, perhaps accounting for occasional multivalent chromosome pairings observed in some extant diploids. However, an evaluation of the frequency and pattern of possible recombination between paralogous sequences produced by WGD has not been available.

Using a comparative and phylogenomic approach, the present research explores possible illegitimate recombination between duplicated genes in rice and sorghum produced by the WGD in their common ancestor. Here, we report our findings about the pattern and frequency of illegitimate recombination by answering the following questions: (1) Has there been illegitimate recombination since rice–sorghum divergence, occurring about 50 Mya? (2) Is there any on-going illegitimate recombination in extant genomes? (3) If there are such recombinations, what genes have been affected and have they been affected along their whole or partial sequence length? (4) What factors may have contributed to retain illegitimate recombination? (5) Sequence base composition variation is often related to gene conversion; do these grass genes provide any supportive evidence?

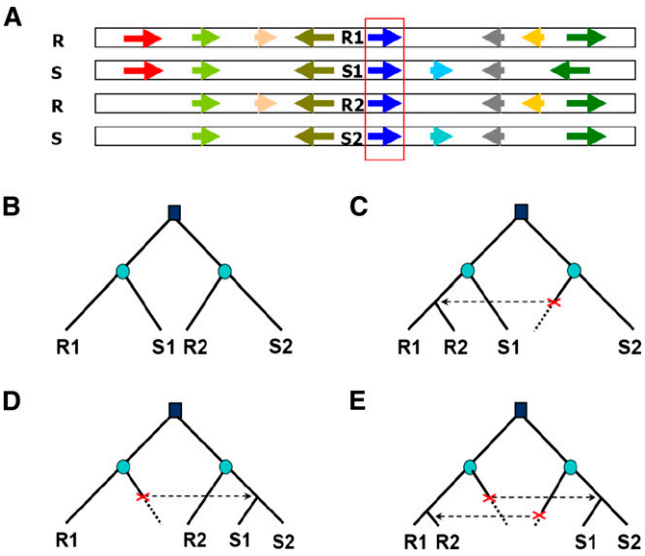

Figure 1. Definition of homologous gene quartets and inference of conversion based on phylogenetic topology changes. (*A*) Arrows show genes and the like-colored ones show homologous genes. Paralogous gene quartets formed by rice (R) paralogous genes R1 and R2, and their respective sorghum (S) orthologs S1 and S2. (*B–E*) Squares symbolize a duplication event in the common ancestral genome and circles symbolize species divergence. The expected phylogenetic relationship of the homologous quartets is displayed in (*B*) if no conversion has occurred; (*C*) if gene R2 is converted by R1; (*D*) if gene S1 is converted by S2; and (*E*) if both the above conversions occurred.

Results

Homologous gene quartets

To detect possible gene conversion between duplicated genes on paralogous chromosomes, we defined quartets of homologous genes by exploiting gene colinearity information. Supposing that a pair of duplicated chromosomal segments had been produced by the WGD in the common ancestor of rice and sorghum, then a homologous gene quartet is formed by two rice paralogous genes R1 and R2, and their respective sorghum orthologs S1 and S2 (Fig. 1A). If no conversion (nonreciprocal recombination) between the duplicates occurred after species divergence, the orthologs should be more similar to one another than either is to the paralogs (Fig. 1B). However, if there has been conversion, we may find aberrant gene tree topology changes (Fig. 1C–E). Gene tree topology is measured based on homologous sequence similarity, and is further checked by bootstrap tests. Since gene sequences may be wholly or partially converted, we employed different methods to detect whole gene conversion and partial gene conversion, respectively (see Methods for details).

Rice–sorghum conversion

We detected 1811 rice–sorghum quartets (Supplemental Table 1), involving 3622 (12.9% and 13.1% of all) rice and sorghum genes, respectively. These quartets reside in nine large duplicated blocks in both species (Fig. 2A,B), distributed unequally among the chromosomes.

We removed highly divergent quartets from further analysis when the gaps in their alignment account for >50% of the alignment length or amino acid identity between any two homologs is <40%, since such divergent sequences would result in problematic alignment, and consequently lead to false inference of conversion. We successfully aligned the sequences for 1721 quartets and characterized the sequence similarity between homologs. Alignment at a more stringent criterion, requiring >70% protein sequence similarity, yielded a similar result, but only permitted analysis of approximately one-third of the quartets. The following analyses are based on the relatively lenient criteria.

By checking aberrant tree topology, at bootstrap percentage ≥80% we found that 14.2% (244 pairs) of rice paralogs have been converted after rice–sorghum divergence, including 5.5% whole and 8.7% partial gene conversions (Table 1). Paralogous pairs on different chromosomes have been unequally affected by gene conversion (Fig. 2C). The most affected chromosomes are Os11 and Os12, with 55.7% of paralogs being affected. In contrast, no paralogs on Os03 and Os12 have been affected. Paralogs on some chromosomes, e.g., Os02, Os04, and Os06, have been more affected by partial than whole gene conversion, while those on the other chromosomes, e.g., Os03 and Os10, have been more affected by whole gene conversion.

Fewer sorghum paralogs have been converted than their rice counterparts. At bootstrap percentage ≥80%, 12.2% (210) of sorghum paralogs have been converted, including 4.1% whole and 8.1% partial gene conversions (Table 1). Conversion rates also show an unequal distribution among sorghum chromosomes (Fig. 2D). Interestingly, sorghum chromosomes show similar conversion patterns to their rice orthologs. Orthologous to Os11 and Os12 in rice, Sb05 and Sb08 are the most converted chromosomes in sorghum (45%). None of Os03–Os12 paralogs has been converted, and only one of their orthologs on Sb01 and Sb08 has been converted. Like Os02, Os04, and Os06, the paralogs on their sorghum orthologs Sb04, Sb06, and Sb10 have been more affected by partial, rather than whole, gene conversion, while those on other sorghum chromosomes have been more affected by whole gene conversion. The rice and sorghum paralogs in 69 quartets (4.4%) have been wholly converted in both species after their divergence,

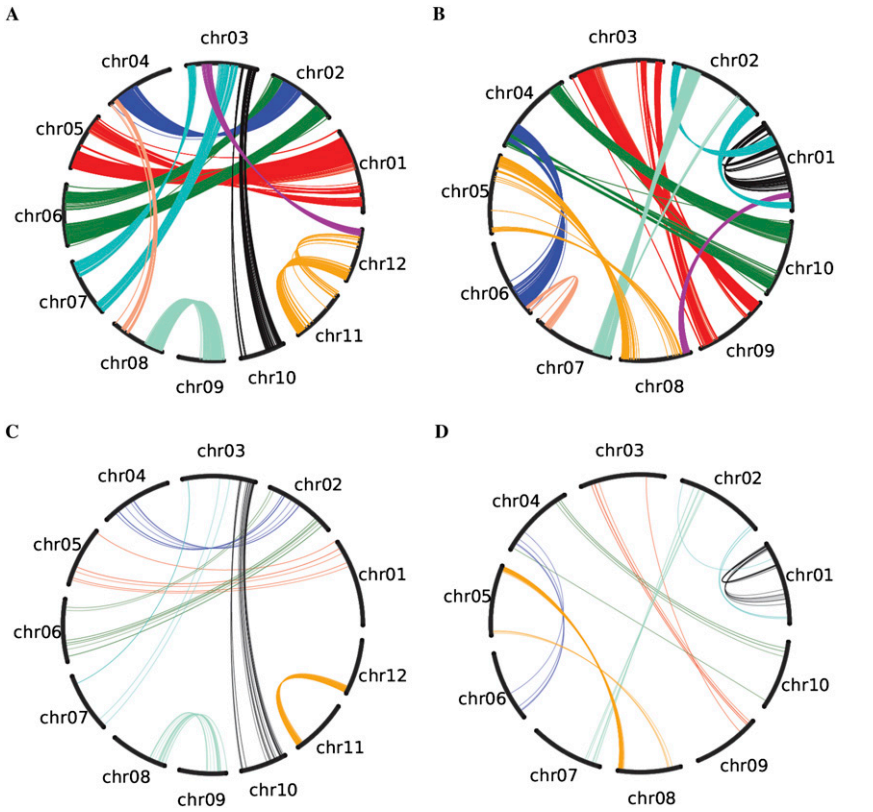

Figure 2. Genome duplications and conversion patterns in rice sorghum. We show in *C* and *D* all the wholly converted genes in largest duplicated blocks. Colored lines are adopted to show homologous regions between chromosomes in two genomes. (*A*) Duplication in rice; (*B*) duplication in sorghum; (*C*) conversion in rice; and (*D*) conversion in sorghum.

more than the expected 43 quartets (*P*-value = 0.03), indicating that some genes are more prone to gene conversion in both species. Some examples of converted genes are shown (Supplemental Fig. 1), including genes encoding 60s ribosomal protein, phosphatidate cytidylyltransferase, and esterase.

On-going conversion

Some duplicated genes have quite small synonymous and nonsynonymous nucleotide substitution difference (p_S and p_N, respectively) values, suggesting the possibility of having recently been converted. These young duplicates are distributed across all rice and sorghum chromosomes. Outstanding evidence of on-going gene conversion is from the duplicated region near the termini of Os11 and Os12, which is shared with corresponding sorghum orthologs Sb5 and Sb8, implying a duplication event in the common ancestor of rice and sorghum (Paterson et al. 2009). However, the duplicated genes appear especially young, with the initial paralogous rice sequences showing 99% identity, and indicating frequent and on-going recombination between duplicated regions, which results in an L-shaped p_S distribution pattern that is not observed in the other blocks (Supplemental Figs. 2 and 3).

Conversion and evolution

Conversion homogenizes paralogous gene sequence. This makes the affected paralogs appear younger than expected, based on sequence divergence with one another. p_N and p_S values between the affected paralogs are often much smaller than those not affected (Table 2). The converted rice paralogs have an averaged p_S = 0.15 and p_N = 0.06, significantly smaller than those (0.49 and 0.20, respectively) of nonconverted paralogs. The converted sorghum paralogs have an averaged p_S = 0.24 and p_N = 0.08, also significantly smaller than those (0.50 and 0.19, respectively) of nonconverted paralogs. Do the converted genes evolve slowly? We could not find the answer based on the paralogs themselves, the pairwise distance between what could have been distorted by gene conversion. Comparative analysis of the corresponding orthologs can provide some insight. We estimated the nucleotide differences of orthologs classified into two groups, those whose paralogs were affected by whole gene conversion in both species and those whose paralogs were not affected in either species. If conversion did not affect evolution, we would find the p_S and p_N values in the two groups to be similar, or we would find different p_S and p_N. Our analysis with the rice–sorghum orthologs indicates that the conversion-related group has a little larger, rather than smaller, p_N and p_S (Table 2), showing that converted paralogs evolve faster (significantly for p_S) than those not affected.

No evidence suggests a change in selection pressure in the converted genes. The converted rice paralogs have an averaged p_N/p_S ratio = 0.34 and nonconverted paralogs have a ratio of 0.44, suggesting a significant selection pressure difference (Table 2). The converted sorghum paralogs also have a smaller average ratio than that of the nonconverted (0.31 vs. 0.42). However, we note that based on the paralogs themselves we could not find the actual selection pressure difference. Because paralogous nonsynonymous and synonymous nucleotide substitution differences have been distorted to be smaller by possible gene conversion, their ratios have also been distorted. Therefore, we turn to rice–sorghum orthologs for help again, whose divergences have not been (directly) affected by gene conversion, and their p_S seem to approximately reflect the time of rice–sorghum divergence. A comparative analysis with the rice–sorghum orthologs indicates that the ratios are basically the same between the converted and nonconverted groups (0.35 and 0.34, respectively; *P*-value = 0.60). This shows that gene conversion has not caused obvious changes in selection pressure.

Conversion and physical location

The physical location of genes may correlate with their chance of being converted. Quartets are often distributed on the distal regions on chromosomes (Fig. 2). We investigated gene conversion rates in relation to proximity to chromosomal termini. Our analysis indicates that genes near the chromosomal termini are more

Table 1. Statistics of converted paralogs in rice and sorghum

Chromosomes of quartets	Quartet1[a]	Quartet2[b]	Rice paralogs wholly converted[c]		Sorghum paralogs wholly converted		Rice paralogs partially converted		Sorghum paralogs partially converted	
			All	B.P. > 0.8[d]	All	B.P. > 0.8	All	B.P. > 0.8	All	B.P. > 0.8
Os01_Os05_Sb03_Sb09	476	457	5 (0.011)	1 (0.002)	2 (0.004)	1 (0.002)	141 (0.308)	64 (0.14)	122 (0.267)	60 (0.131)
Os02_Os04_Sb04_Sb06	270	262	7 (0.027)	1 (0.004)	3 (0.011)	2 (0.008)	75 (0.286)	39 (0.149)	53 (0.202)	39 (0.149)
Os02_Os06_Sb04_Sb10	303	283	5 (0.0177)	1 (0.004)	5 (0.018)	1 (0.004)	94 (0.332)	46 (0.163)	61 (0.216)	39 (0.138)
Os03_Os07_Sb01_Sb02	215	203	6 (0.030)	1 (0.005)	2 (0.010)	1 (0.005)	0 (0)	0 (0)	0 (0)	0 (0)
Os03_Os10_Sb01_Sb01	117	116	53 (0.457)	14 (0.121)	27 (0.233)	5 (0.043)	1 (0.009)	1 (0.009)	1 (0.009)	1 (0.009)
Os03_Os12_Sb01_Sb08	50	46	0 (0)	0 (0)	1 (0.022)	1 (0.022)	0 (0)	0 (0)	0 (0)	0 (0)
Os04_Os08_Sb06_Sb07	43	38	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Os08_Os09_Sb07_Sb02	194	185	7 (0.038)	3 (0.016)	3 (0.016)	1 (0.005)	0 (0)	0 (0)	0 (0)	0 (0)
Os11_Os12_Sb05_Sb08	143	131	80 (0.611)	73 (0.557)	70 (0.534)	59 (0.450)	0 (0)	0 (0)	0 (0)	0 (0)
Summary	1811	1721	163 (0.088)	94 (0.055)	113 (0.066)	71 (0.041)	311 (0.181)	150 (0.087)	237 (0.138)	139 (0.081)

[a]All detected quartets are included.
[b]Quartets successfully aligned are included.
[c]Conversion rates are in parentheses.
[d]B.P. denotes bootstrapping percentage.

frequently affected by gene conversion (Table 3). In rice, affected genes have an average distance of 6.1 Mb to termini (*P*-value = 0.03), with wholly converted being 3 Mb (*P*-value = 4.7×10^{-41}), as compared with 6.6 Mb of total rice genes in quartets. In sorghum, affected genes have an average distance of 7.6 Mb to termini (*P*-value = 2.1×10^{-4}), with wholly converted being 5.4 Mb (4.3×10^{-13}), as compared with 8.6 Mb. In rice, >50% of wholly converted genes are in the initial 2 Mb regions on the chromosomal termini, in which ~40% of the duplicated genes have been converted. In sorghum, 48.6% of wholly converted genes are in the initial 2 Mb regions on the chromosomal termini, in which ~34.5% of the duplicated genes have been converted.

Conversion and GC content

We found no correlation between gene conversion and GC content. In both rice and sorghum, converted paralogs usually have similar GC content to nonconverted paralogs (in rice: 0.58 vs. 0.58; in sorghum: 0.58 vs. 0.59) (Supplemental Table 1). This indicates that conversion is not the cause of GC elevation, as further discussed in the Supplemental text.

Conversion and gene function

Converted genes tend to have more similar expression patterns than nonconverted duplicates. We obtained expression measures from microarrays for 917 rice duplicated pairs, finding 24.8% of them significantly correlated in expression, as compared to random samples. Comparatively, 38.5% of the wholly converted duplicates are significantly correlated in expression, a much higher percentage than all the duplicates (*P*-value < 2.2×10^{-16}, being the smallest *P*-value that R language can output), while the correlation pattern between partially converted genes is similar to that of other duplicates. By checking the expressed sequence tags (ESTs) of sorghum unigenes, we obtained similar findings as to rice. Both the wholly and partially converted genes often have similar numbers of ESTs (Pearson coefficient rho = 0.57/0.66 and *P*-value = $1.4 \times 10^{-4}/7.3 \times 10^{-11}$, respectively), much higher than the nonconverted genes (rho = 0.293).

We found weak evidence that genes with specific functions have been preferentially converted. By checking the Pfam domains in the converted and nonconverted duplicated genes, we found that a small fraction (5.1% and 5.7%, respectively) of rice and sorghum duplicates have been likely preferentially converted at significance level 0.01. The most affected domains are LysM domains (PF01476), citrate transporter domains (PF03600), and EF hand domains (PF00036) in rice, and multicopper oxidase domains (PF07732 and PF00394) and pollen allergen domains (PF01357) in sorghum. However, after Bonferroni correction, none is significantly enriched in converted genes.

Discussion

Polyploidy and conversion

Recently, genome-scale analysis indicates that gene conversion may be quite common in divergent species such as yeast (Gao and Innan 2004) and mammals (Ezawa et al. 2006), and was supposed to explain low sequence divergence between duplicated genes in plants after large-scale genome doubling (Chapman et al. 2006).

Table 2. Nucleotide substitution rates of quartets in rice and sorghum

	Rice paralogs			Sorghum paralogs			Rice–sorghum orthologs		
	p_N	p_S	p_N/p_S	p_N	p_S	p_N/p_S	p_N	p_S	p_N/p_S
Genes wholly converted[a]	0.055	0.146	0.343	0.082	0.249	0.311	0.147	0.424	0.345
Genes nonconverted[b]	0.199	0.486	0.436	0.197	0.499	0.417	0.132	0.396	0.339
P-value	7.58×10^{-28}	7.36×10^{-86}	3.70×10^{-3}	2.09×10^{-19}	2.76×10^{-57}	3.07×10^{-4}	0.078	0.041	0.594

[a]Both rice and sorghum paralogs (69 in total) were wholly converted.
[b]Neither rice nor sorghum paralogs (1021) were converted.

Table 3. Gene physical location and gene conversion

Distance to telomere	<2 Mbp	2–4 Mbp	4–6 Mbp	6–8 Mbp	8–10 Mbp	>10 Mbp
Rice						
Genes in quartets	510	620	649	451	354	712
Genes wholly converted (%)	0.316	0.131	0.032	0.062	0.037	0.020
All converted genes (%)	0.400	0.255	0.250	0.275	0.299	0.261
Sorghum						
Genes in quartets	391	409	552	471	366	1107
Gene wholly converted (%)	0.263	0.110	0.027	0.013	0.030	0.029
All converted genes (%)	0.345	0.242	0.190	0.163	0.232	0.167

Angiosperms have been recursively affected by WGDs (Tang et al. 2008a), which may often be followed by genome instability, characterized by massive DNA rearrangement, inversion, and DNA loss, often leading to reestablishment of diploid heredity. Soon after polyploidization, multiple homologous chromosomes or chromosomal segments may compete to pair and recombine with one another, forming multivalent structures during meiosis. DNA rearrangement may inhibit the chance of pairing between affected chromosomes or chromosomal segments. Gradually, structural and sequence divergence may establish neo-homologous chromosome pairs with bivalent structure reestablished during meiosis. Those chromosomes or chromosomal segments sharing ancestry, but less similar in structure and sequence, are then referred to as homoeologous. Though widespread and frequent recombination between homoeologous DNA segments may have been restricted, occasional and small-scale recombination may persist for a long time. The present analysis in rice and sorghum reveals extensive homoeologous recombination millions of years after genome duplication, which may have contributed to the evolution of these plants and their ancestors. By using GENECONV (Sawyer 1989), pairwise searching for conversion between *Arabidopsis* paralogs produced by whole-genome duplication found no evidence of conversion (Zhang et al. 2002), however, using the same method, an exploration for conversion in rice revealed 377 events in 626 multigene families (Xu et al. 2008). This controversy may result from the fact that *Arabidopsis* genes diverge faster than rice genes and may more frequently escape conversion; or massive genome fractionation after recurrent genome duplications may have greatly restricted conversion (Tang et al. 2008a).

As shown above, physical location is related to conversion rate in rice and sorghum with more conversion nearer to the telomere. Assuming that sequence similarity is the physical basis for recombination, this finding is reasonable for several reasons. First, gene sequences, often more abundant in regions away from centromeres, are more conservative than other sequences and, therefore, better preserve sequence similarity with their homoeolog. Gene colinearity is often found in gene-dense (enchromatic) regions, where legitimate recombination (in contrast to illegitimate recombination) is active but not in the gene-scarce (heterochromatid) regions (Bowers et al. 2005). Active recombination may preserve sequence similarity by removing deleterious mutations (Carvalho 2003). Second, repetitive elements are often enriched in pericentromeric regions, which reduce large-scale sequence similarity between homoeologous segments by inducing DNA rearrangements and mutations. In both rice and sorghum, long-terminal repeat (LTR) elements are substantially enriched in the pericentromeric regions, making up ~50% of all DNA in rice and ~80% in sorghum, as compared to only 20%–30% in the gene-dense regions (Yu et al. 2005; Paterson et al. 2009). In the initial 2 Mb DNA short arms of Os11 and Os12, where conversion is the highest, only ~15% of sequences are LTRs, as compared with an average of ~42% throughout the genome (Yu et al. 2002). The corresponding regions in sorghum show similarly low levels of LTRs.

The sizes of duplicated blocks of genes may be positively correlated with conversion rate. The smallest blocks, such as the ones between Os03 and Os12, and Os04 and Os08, have the lowest conversion (Table 1), as do their orthologous sorghum segments. When small duplicated blocks are buried in chromosomes that otherwise share little or no homoeology, they may have little chance to pair. This may be particularly true when other regions of the chromosome do have large-scale homoeology with other chromosomes, leaving the small duplicated regions at a disadvantage in forming homoeologous duplexes.

DNA inversion may have directly contributed to recombination restriction between homoeologous regions. Though Os01 and Os05 share large-scale homoeology characterized by ~600 homoeologous genes and 476 quartets, the conversion rate between them is among the lowest, as is also the case with the orthologous sorghum chromosomes. A possible explanation is that the homoeologous genes are in two divided groups near each end of the chromosomes, and that a large inversion before the rice–sorghum divergence in the short arm (Paterson et al. 2009) may have reduced competence to form homoeologous duplexes.

Recombination might have been restricted in a non-synchronized manner. We found that conversion rates differ among duplicated blocks produced by GD. We infer that recombination suppression among homoeologous blocks might not have occurred at the same time, i.e., some may be restricted earlier than others, in that antirecombination factors, such as chromosomal rearrangements, might have occurred in a stochastic way. Though there is clear evidence that the paralogous segment on the termini of the short arms of Os11 and Os12, and the corresponding regions on S05 and S08, was produced before species divergence, illegitimate recombination has continued for millions of years.

Interestingly, the rice and sorghum orthologous chromosomes/chromosomal segments show similar patterns of gene conversion. This might be explained in that the divergence levels between the ancestral paralogous chromosomes in the cereal common ancestor influenced the recombination pattern in the offspring. Unlike the other rice chromosomes, Os02, Os04, and Os06, and their respective orthologs Sb04, Sb06, and Sb10, have higher partial, rather than whole, gene conversions. However, the direct cause needs further exploration.

Homoeologous recombination may occur at very different rates among different duplicated blocks and have been on-going

between specific chromosomal segments. This is supported by the fact that some homoeologous genes have very little nucleotide difference. Clear evidence is from the homoeologous genes on the initial 3 Mb of the short arms' termini of Os11 and Os12, where p_S is as low as zero, as described in our previous report (Wang et al. 2007). Previously, the segment was reported to be recently duplicated (5–7 Mya) (The Rice Chromosomes 11 and 12 Sequencing Consortia 2005; Wang et al. 2005), but a comparison with the sorghum genome shows that it existed before sorghum and rice diverged about 50 Mya (Paterson et al. 2009), showing its ancient nature.

Conversion and evolution

Gene conversion has been widely used to explain the evolution of large gene families, such as histone genes (Ohta 1984) and rRNA genes (Brown et al. 1972). It was reported that gene conversion may make these genes evolve at quite a slow pace. Our above analysis indicates that converted genes may evolve faster than those not converted. Though the converted genes have relatively small sequence divergence between them, they only appear young. The fast evolution of converted genes can be explained by classical evolutionary theory, which anticipates that gene redundancy may lead to relatively fast mutation accumulation in at least one of the genes, supported by both comparative sequence analysis and genetic theory (Lynch and Conery 2000). Therefore, we propose that conversion acts as an occasional, sometimes frequent, interruption to gene evolution, after which the homogenized gene copies restart their respective evolutionary journeys; and as an accelerating force contributing abrupt changes to affected genes. Our research supports the previously proposed linkage between gene conversion and highly conserved gene clusters. However, the conservative nature of these genes leads to the occurrence of gene conversion between homologs, this conversion may not actually contribute to gene conservation.

Methods

Inference of homologous quartets

Rice and sorghum (version 1.0) sequences were downloaded from the RAP2 database (http://rgp.dna.affrc.go.jp/E/index.html) and Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/). We performed all-against-all BLASTP between rice and sorghum predicted proteins to search for potential anchors (E_value $< 1 \times 10^{-5}$, top five matches) between every possible pair of chromosomes in multiple genomes. The homologous pairs are used as the input for MCscan (Tang et al. 2008b). A built-in scoring scheme for MCscan is min($-\log_{10}E$_value, 50) for every matching gene pair and -1 for each 10 Kb distance between anchors and blocks that have scores >300 were kept. The resulting syntenic chains are evaluated using a procedure by ColinearScan (Wang et al. 2006).

Detecting whole gene conversion

The above quartets were aligned with ClustalW (Thompson et al. 1994), and alignments for which gaps accounted for >50% of the alignment length or amino acid identity <40% were removed from further analysis. For paralogous quartets defined (Fig. 1A), since paralogous genes were produced prior to the species divergence, we anticipate that the orthologs S1 and R1, and S2 and R2 were more similar to one another than to their respective paralogs (Fig. 1B), if there had been no gene conversion. However, aberrant gene tree topologies reflect independent or concurrent conversion events (Fig. 1C–E). To detect possible whole gene conversion, we used phylogenetic analysis to identify possible topology changes in the homologous quartets. To reflect the gene tree topology, we characterized sequence similarity between homologs in quartets. A bootstrap test was performed to evaluate the significance of putative gene conversions with 1000 repetitive samplings to produce a bootstrap frequency indicating the confidence level of the supposed conversion. To detect possible partial gene conversion, we integrated a tree topology search and a dynamic programming algorithm to search the partially affected regions ≥10 nucleotides in length, as previously reported (Wang et al. 2007).

Pfam analysis

Genes in homologous quartets were linked to the PFAM domains (version 22) by running BLAST at E-value threshold 1×10^{-5}.

Expression analysis

Rice gene expression data were downloaded from NCBI Gene Expression Omnibus (GSE6893) (Barrett et al. 2009), containing 45 Affymetrix microarray slides and for 15 samples (each having three replicates), which was generated with various tissues and organs, including seedling root, young, and mature leaves, and at different stages of reproductive development, such as panicle and seed (Jain et al. 2007). Quality of arrays was assessed using affyPLM in Bioconductor (Gentleman et al. 2004), and one slide (GSM159192) was discarded for being a possible artifact, suggested by a broken image. We measured expression level by using the robust multiarray average (RMA) method (Irizarry et al. 2003). Presence calls for all probe sets in each slide were made by using the MAS5 algorithm (Affymetrix). A probe set was taken to be present in the sample if present calls were assigned for all replicates for that sample. Since we were analyzing the RAP2 gene models to reveal conversion, the loci defined by the Institute of Genome Research (TIGR) in the array analysis were linked to the RAP2 gene models. One RAP2 gene model may be related to multiple TIGR loci, and vice versa. We took only the genes with one-to-one correspondence between RAP2 models and TIGR loci in the present analysis. To measure the expression divergence between duplicated genes, the Pearson correlation coefficient was calculated for each pair with RAM measures.

Sorghum bicolor transcript assemblies (48932 unigenes assembled from 203575 ESTs) were downloaded from TIGR Plant Transcript Assemblies database (Childs et al. 2007). Each of the unigenes is composed of varying numbers of ESTs, which are used to approximate the number of times a particular gene model is sampled. To measure the expression pattern correlation between duplicated genes, the Pearson correlation coefficient between the numbers of ESTs of the duplicated genes was calculated.

Gene phylogeny

Example trees were constructed with MEGA (Tamura et al. 2007) based on protein sequence alignment, and bootstrap tests were performed to show stability of twigs in trees.

Acknowledgments

We appreciate financial support from the U.S. National Science Foundation (MCB-0450260 to A.H.P.). We thank Zhe Li at Beijing University for helpful discussions on rice gene expression.

References

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., et al. 2009. NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res.* **37:** D885–D890.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433–438.

Bowers, J.E., Arias, M.A., Asher, R., Avise, J.A., Ball, R.T., Brewer, G.A., Buss, R.W., Chen, A.H., Edwards, T.M., Estill, J.C., et al. 2005. Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc. Natl. Acad. Sci.* **102:** 13206–13211.

Brown, R.D., Mattoccia, E., and Tocchini-Valentini, G.P. 1972. On the role of RNA in gene amplification. *Acta Endocrinol. Suppl. (Copenh.)* **168:** 307–318.

Carvalho, A.B. 2003. The advantages of recombination. *Nat. Genet.* **34:** 128–129.

Chapman, B.A., Bowers, J.E., Feltus, F.A., and Paterson, A.H. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc. Natl. Acad. Sci.* **103:** 2730–2735.

Childs, K.L., Hamilton, J.P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P.D., Town, C.D., Buell, C.R., and Chan, A.P. 2007. The TIGR plant transcript assemblies database. *Nucleic Acids Res.* **35:** D846–D851.

Datta, A., Hendrix, M., Lipsitch, M., and Jinks-Robertson, S. 1997. Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. *Proc. Natl. Acad. Sci.* **94:** 9757–9762.

Dooner, H.K. and Martinez-Ferez, I.M. 1997. Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. *Plant Cell* **9:** 1633–1646.

Ezawa, K., OOta, S., and Saitou, N. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Mol. Biol. Evol.* **23:** 927–940.

Fu, H., Zheng, Z., and Dooner, H.K. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl. Acad. Sci.* **99:** 1082–1087.

Gao, L.Z. and Innan, H. 2004. Very low gene duplication rate in the yeast genome. *Science* **306:** 1367–1370.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5:** R80. doi: 10.1186/gb-2004-5-10-r80.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249–264.

Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A.K., and Khurana, J.P. 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiol.* **143:** 1467–1483.

Khakhlova, O. and Bock, R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J.* **46:** 85–94.

Li, L., Jean, M., and Belzile, F. 2006. The impact of sequence divergence and DNA mismatch repair on homeologous recombination in *Arabidopsis*. *Plant J.* **45:** 908–916.

Liharska, T., Wordragen, M., Kammen, A., Zabel, P., and Koornneef, M. 1996. Tomato chromosome 6: Effect of alien chromosomal segments on recombinant frequencies. *Genome* **39:** 485–491.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Ohta, T. 1984. Some models of gene conversion for treating the evolution of multigene families. *Genetics* **106:** 517–528.

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.

Paterson, A.H., Bowers, J.E., 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.

Puchta, H., Dujon, B., and Hohn, B. 1996. Two different but related mechanisms are used in plants for the repair of genomic double-strand breaks by homologous recombination. *Proc. Natl. Acad. Sci.* **93:** 5055–5060.

The Rice Chromosomes 11 and 12 Sequencing Consortia. 2005. The sequence of rice chromosomes 11 and 12, rice in disease resistance genes and recent gene duplications. *BMC Biology* **3:** 20. doi: 10.1186/1741-7007-3-20.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6:** 526–538.

Stambuk, S. and Radman, M. 1998. Mechanism and control of interspecies recombination in *Escherichia coli*. I. Mismatch repair, methylation, recombination and replication functions. *Genetics* **150:** 533–542.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24:** 1596–1599.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. 2008a. Synteny and collinearity in plant genomes. *Science* **320:** 486–488.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18:** 1944–1954.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. ClustalW: 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.

Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J. 2005. Duplication and DNA segmental loss in the rice genome: Implications for diploidization. *New Phytol.* **165:** 937–946.

Wang, X., Shi, X., Li, Z., Zhu, Q., Kong, L., Tang, W., Ge, S., and Luo, J. 2006. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics* **7:** 447. doi: 10.1186/1471-2105-7-447.

Wang, X., Tang, H., Bowers, J.E., Feltus, F.A., and Paterson, A.H. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177:** 1753–1763.

Xu, S., Clark, T., Zheng, H., Vang, S., Li, R., Wong, G.K., Wang, J., and Zheng, X. 2008. Gene conversion in the rice genome. *BMC Genomics* **9:** 93. doi: 10.1186/1471-2164-9-93.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79–92.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., et al. 2005. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3:** e38. doi: 10.1371/journal.pbio.0030038.

Zhang, L., Vision, T.J., and Gaut, B.S. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **19:** 1464–1473.

Received September 28, 2008; accepted in revised form February 24, 2009.

Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization

Xiyin Wang, Haibao Tang, John E. Bowers, et al.

Genome Res. 2009 19: 1026-1032 originally published online April 16, 2009
Access the most recent version at doi:10.1101/gr.087288.108