The Plant Cell, Vol. 24: 21–32, January 2012, www.plantcell.org

Pericentromeric Effects Shape the Patterns of Divergence, Retention, and Expression of Duplicated Genes in the Paleopolyploid Soybean [C][W]

Jianchang Du[a,b], Zhixi Tian[a,1], Yi Sui[a], Meixia Zhao[a,c], Qijian Song[d], Steven B. Cannon[e], Perry Cregan[d], and Jianxin Ma[a,2]

[a] Department of Agronomy, Purdue University, West Lafayette, Indiana 47907
[b] Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China
[c] Institute of Oil Crops, Chinese Academy of Agricultural Sciences, Wuhan 430062, China
[d] U.S. Department of Agriculture, Agricultural Research Service, Soybean Genomics and Improvement Laboratory, Beltsville Agricultural Research Center-West, Beltsville, Maryland 20705
[e] U.S. Department of Agriculture, Agricultural Research Service, Corn Insect and Crop Genetics Research Unit, Ames, Iowa 50011

The evolutionary forces that govern the divergence and retention of duplicated genes in polyploids are poorly understood. In this study, we first investigated the rates of nonsynonymous substitution (Ka) and the rates of synonymous substitution (Ks) for a nearly complete set of genes in the paleopolyploid soybean (*Glycine max*) by comparing the orthologs between soybean and its progenitor species *Glycine soja* and then compared the patterns of gene divergence and expression between pericentromeric regions and chromosomal arms in different gene categories. Our results reveal strong associations between duplication status and Ka and gene expression levels and overall low Ks and low levels of gene expression in pericentromeric regions. It is theorized that deleterious mutations can easily accumulate in recombination-suppressed regions, because of Hill-Robertson effects. Intriguingly, the genes in pericentromeric regions—the cold spots for meiotic recombination in soybean—showed significantly lower Ka and higher levels of expression than their homoeologs in chromosomal arms. This asymmetric evolution of two members of individual whole genome duplication (WGD)-derived gene pairs, echoing the biased accumulation of singletons in pericentromeric regions, suggests that distinct genomic features between the two distinct chromatin types are important determinants shaping the patterns of divergence and retention of WGD-derived genes.

INTRODUCTION

Genomic duplications on various scales, such as amplification of individual genes, segmental duplication, and whole genome duplication (WGD) by polyploidy, have been recognized as important contributors to evolutionary innovation in eukaryotes (Ohno, 1970). Among these duplication events, WGD is particularly common among flowering plants and is often recurrent (Soltis and Soltis, 1999; Otto and Whitton, 2000). It is believed that all flowering plants have undergone at least one round of WGD (Jiao et al., 2011)—30 to 80% are currently polyploids, and others are paleopolyploids (Otto and Whitton, 2000; Wendel, 2000; Wolfe, 2001; Soltis and Soltis, 2009). Most paleopolyploids have undergone extensive genomic rearrangements, including elimination of a large fraction of duplications (Lynch and Conery, 2003; De Bodt et al., 2005) and accumulation of mutations that may contribute to functional divergence of duplicated genes (Blanc and Wolfe, 2004; Gu et al., 2005; Sémon and Wolfe, 2007; Ha et al., 2009).

The patterns of elimination/retention of duplicated genes after WGD have been investigated in *Arabidopsis thaliana* and maize (*Zea mays*). It was shown that, in both species, genes were preferentially removed from one of the two homologs or homoeologs derived from WGD, an evolutionary process termed "biased fractionation" (Thomas et al., 2006; Woodhouse et al., 2010). Further comparative sequence analysis of the maize, sorghum (*Sorghum bicolor*), and rice (*Oryza sativa*) genomes revealed that the fractionation of the two homoeologs in maize was predominately achieved by elimination of single genes, perhaps through illegitimate recombination (Woodhouse et al., 2010), a process that accumulates small deletions without the requirement for the participation of a recA protein or large (>50 bp) stretches of sequence homology (Bennetzen et al., 2005). The mechanisms that underlie biased fractionation are not clear yet, although rapid functional divergence and biased expression of duplicated genes seem to be major factors promoting their

[1] Current address: Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China.
[2] Address correspondence to maj@purdue.edu.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Jianxin Ma (maj@purdue.edu).
[C]Some figures in this article are displayed in color online but in black and white in the print edition.
[W]Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.111.092759

retention in the genome (Force et al., 1999; Lynch and Force, 2000; He and Zhang, 2005; Sémon and Wolfe, 2007; Schnable et al., 2011).

Theoretically, evolutionary rates of duplicated genes should increase as the two copies evolve toward functional divergence (Ohno, 1970). Several studies have revealed a tendency of rate acceleration immediately after gene duplication (Lynch and Conery, 2000; Kondrashov et al., 2002). By contrast, recent studies showed that duplicated genes evolve slower than singletons in eukaryotes (Yang et al., 2003; Davis and Petrov, 2004; Jordan et al., 2004; Yang and Gaut, 2011). These two sets of observations may not be inconsistent, but reflect variation in the evolutionary modes and dynamics of duplicated genes within different time frames.

Recombination has been recognized as one of the key factors that play pivotal roles in shaping genomic features, such as the pattern of nucleotide variation and distribution of genes and transposable elements (TEs) in eukaryotes (Gaut et al., 2007; Tian et al., 2009). It is apparent that variations of recombination rates and genomic features are both dynamic evolutionary processes. However, recombination rates (centimorgans per megabase [cM/Mb]) along chromosomes within a genome have routinely been estimated by integrating physical and genetic maps, with the latter generally constructed using a mapping population derived from two varieties of a species or two recently diverged subspecies that are capable of intercrossing (Gaut et al., 2007). By contrast, genomic features, such as the evolutionary rates of genes, are often estimated by comparison of two species that diverged tens of millions of years ago (Yang and Gaut, 2011). Thus, it is often difficult or impossible to compare recombination rates and genomic parameters on a similar time scale.

Nevertheless, a recent analysis of evolutionary rates for orthologous genes between *A. thaliana* and *Arabidopsis lyrata* that diverged from a common ancestor ~13 million years ago (mya) (Beilstein et al., 2010), demonstrated that the synonymous substitution rates (Ks) vary along chromosomes, perhaps as a function of recombination, and the nonsynonymous substitution rates (Ka) largely correlate with the expression patterns and duplication status (i.e., whether a gene was duplicated and retained after the most recent WGD ~47 to 65 mya) (Beilstein et al., 2010). However, few studies have performed large-scale or genome-wide comparisons between paralogs or homoeologs regarding their rates of evolution and surrounding chromatin environment, such as recombination, mainly because of the lack of sequences from closely related genomes. As a result, little is known regarding the evolutionary forces acting to shape the patterns of divergence and retention of duplicated genes in any paleopolyploid organism.

Soybean (*Glycine max*), which is one of the most economically important leguminous seed crops, was domesticated from its annual wild relative, *Glycine soja*, in China ~5000 years ago (Carter et al., 2004). It is documented that the *Glycine* lineage has undergone two rounds of WGD within the last 60 million years, with the latter (occurring between 5 and 13 mya [Doyle and Egan, 2010; Schlueter et al., 2004; Schmutz et al., 2010]) perhaps being an allotetraploidy event, as proposed by analysis of centromere satellite repeats (Gill et al., 2009). In the sequence of the 1.1-gigabase soybean (cv Williams 82) genome, there are 46,430 predicted high-confidence genes, of which 31,264 (i.e., 15,632 gene pairs) exist as "recent" paralogs. These paralogs are believed to have been duplicated and retained after the 13-mya tetraploidy event, and 15,166 have reverted to singletons (Schmutz et al., 2010). One striking feature of the soybean genome is that ~57% of the genomic sequence occurs in recombination-suppressed heterochromatic regions surrounding centromeres (referred to as pericentromeric regions), where 10,029 high-confidence genes are harbored. Another unusual observation is that the proportion of long terminal repeat retrotransposons in nonpericentromeric regions (referred to as chromosomal arms) of soybean is unusually low (8.7%)—even lower than in chromosomal arms of rice (17%) (Tian et al. 2009), which possesses a much smaller (<400 Mb) genome.

Considering the extremely contrasting genomic features, such as the rates of recombination between the pericentromeric regions and chromosomal arms, and the existence of a vast number of genes in the pericentromeric regions, we wondered whether these distinct regions have different effects in shaping the patterns of divergence, retention, and expression of duplicated genes in soybean. Recently, 14 *G. max* and 17 *G. soja* accessions were resequenced (Lam et al., 2010), providing an unprecedented opportunity to investigate genome-wide variation of evolutionary rates within a recent evolutionary time frame in the context of chromatin environment and duplication status. In this study, we first calculated the rates of substitutions at synonymous and nonsynonymous sites between the *G. soja* and *G. max* populations for a nearly complete set of soybean genes. We then compared evolutionary rates and gene expression patterns between pericentromeric regions and chromosomal arms in several gene categories: i) WGD-derived genes (or just "WGD genes") with both copies in chromosomal arms, ii) WGD genes with both copies in pericentromeric regions, iii) WGD genes with one copy in pericentromeric regions and the other in chromosomal arms, iv) singletons in chromosomal arms, and v) singletons in pericentromeric regions. Through these analyses, we were able to draw a comprehensive picture illustrating how and to what extent recombination-suppressed pericentromeric regions influence the divergence, retention, and expression of duplicated genes in a complex paleopolyploid genome.

RESULTS

Contrasting Genomic Features between Pericentromeric Regions and Chromosomal Arms—An Overview

The pericentromeric regions of the soybean genome were defined on the basis of recombination-suppressed genomic blocks with transitions in gene density and TE density along chromosomes (Schmutz et al., 2010; Du et al., 2010). These regions range from 15.3 to 45.3 Mb, with an average of 27.2 Mb, comprising 57% of the soybean genome but accounting for only 6.9% of recombination (i.e., genetic distance) (Table 1; see Supplemental Figure 1 online). Of the 20 chromosomes, 16 contain recombination-suppressed pericentromeric regions that are larger than the corresponding chromosomal arms (see

Table 1. Comparison between the Pericentromeric Regions and Chromosomal Arms of 20 Individual Chromosomes of the Soybean Genome

Features[a]	Pericentromeric Regions[b]	Chromosomal Arms[b]	P Values[c]
Ratios of DNA (%)	56.57 ± 11.08	43.43 ± 11.08	0.0006
Proportion of LTR RT DNA (%)	47.24 ± 5.60	8.70 ± 1.91	<0.0001
Proportion of DNA TE DNA (%)	21.54 ± 3.40	8.88 ± 1.21	<0.0001
Gene densities (genes/Mb)	19.1 ± 5.5	89.6 ± 6.5	<0.0001
Ratios of recombination (%)	7.1 ± 4.3	92.9 ± 4.3	<0.0001
Recombination rates (cM/Mb)	0.29 ± 0.15	5.19 ± 0.76	<0.0001
Ka (×1000)	0.1387 ± 0.0408	0.1337 ± 0.0161	0.6090
Ks (×1000)	0.3231 ± 0.0869	0.3942 ± 0.0628	0.0052
ω (Ka/Ks)	0.4126 ± 0.0643	0.3612 ± 0.0290	<0.0001
Expression levels	15.99 ± 66.19	22.87 ± 78.32	<0.0001

LTR RT, long terminal repeat retrotransposons.
[a]Ka and Ks were calculated by pairwise comparison between *G. max* and *G. soja* accessions.
[b]Mean ± SD.
[c]Student's *t* test.

Supplemental Table 1 and Supplemental Figure 1 online). As shown in Table 1, the pericentromeric regions contain significantly higher proportions of TEs and lower densities of genes than do the chromosomal arms. The average rates of genetic recombination (GR) are 0.29 cM/Mb in pericentromeric regions and 5.19 cM/Mb in chromosomal arms (Table 1; see Supplemental Figure 1 online), corresponding to an ~18-fold reduction of GR rate in the former regions.

Lower Evolutionary Rates for Genes in Pericentromeric Regions Than in Chromosomal Arms—a Ks Scenario

We began with the complete set of genes (46,430) annotated in the soybean reference genome and then removed low-confidence homoeologous gene pairs and genes with low quality or missing resequencing data from the 14 *G. max* and 17 *G. soja* genomes (Lam et al., 2010) (see Methods). This left 14,577 WGD gene pairs and 12,994 singletons as the final data set for analysis of evolutionary rates. Each of these genes in the reference genome was aligned with its orthologous gene sequences from the 31 resequenced genomes, and then the Ka, Ks, and Ka/Ks (ω) for each gene between the *G. max* and *G. soja* populations were calculated by pairwise comparison (see Supplemental Data Set 1 and Supplemental Figure 2 online). The results showed that the average Ks for pericentromeric regions was significantly lower than for the chromosomal arms, indicating slower evolution of genes in pericentromeric regions (Table 1). By contrast, no significant difference in Ka between the two regions was detected (Table 1), suggesting that overall, genes in both regions have undergone similar levels of selective constraints—although the average ω for genes in pericentromeric regions was significantly higher than in chromosomal arms.

Higher Level of Preservation of WGD Genes in Chromosomal Arms Than Pericentromeric Regions

To determine whether the retention of the WGD genes was biased for or against pericentromeric regions, which exhibited overall slower evolutionary rates, we compared the distribution of the 14,577 WGD gene pairs and 12,994 singletons between the pericentromeric regions and chromosomal arms of the 20 soybean chromosomes (Table 2). Of the 14,577 WGD gene pairs, 11,285 (77.4%) are in chromosomal arms, and 853 (5.9%) are in pericentromeric regions. These two categories of WGD genes were referred to as WGD-I genes. The remaining 2439 (16.7%) gene pairs each have one copy in a chromosomal arm and the other in a pericentromeric region and were referred to as WGD-II genes (Figure 1). By contrast, 8389 (64.5%) of the 12,994 singletons are located in chromosomal arms, and the remaining 4605 (35.4%) singletons are in pericentromeric regions. When we excluded the 2439 WGD-II gene pairs, the ratios of duplicated gene pairs to singletons were 1:5.40 in pericentromeric regions versus 1.34:1 in chromosomal arms (Table 2), corresponding to a 15.6% retention rate of WGD genes in the former versus 57.5% in the latter.

Lower Evolutionary Rates for WGD Genes Than Singletons in the Whole Genome—a Ka Scenario

To understand the evolutionary dynamics of the WGD genes versus singletons in the soybean genome, we compared the evolutionary rates for the 12,138 WGD-I gene pairs and the 12,994 singletons using the same data set as described above (Table 2). These WGD gene pairs were predicted based not only on the Ks analysis of homologous genes but also on the identification of 149 homoeologous genomic blocks that contain these genes (Schmutz et al., 2010). Our analysis demonstrates that the average Ka for the WGD genes is significantly lower than that for the singletons, whereas there is no significant difference in Ks between the WGD genes and the singletons (Table 3). Overall, the average ω for the WGD genes is significantly lower than for the singletons (Table 3), indicating that the former have experienced an overall higher level of purifying selection than the latter. These results, consistent with previous observations from several other eukaryotes (Davis and Petrov, 2004; Jordan et al., 2004; Yang and Gaut, 2011), suggest that the WGD genes have undergone a stronger level of functional constraint than the singletons.

Table 2. Distribution of the WGD Genes and Singletons in the Pericentromeric Regions and Chromosomal Arms of Soybean

Duplication Status	Pericentromeric Regions	Chromosomal Arms
WGD-I gene pairs[a]	853	11,285
Singletons	4,605	8,389
WGD-II genes[b]	2,439	2,439
Ratios of WGD-I gene pairs to singletons	1:5.40	1.34:1

[a]WGD gene pairs with both members in either pericentromeric regions or chromosomal arms.
[b]WGD genes in pericentromeric regions with corresponding homoeologs in chromosomal arms.

Higher Ks for the WGD Genes Than Singletons in Pericentromeric Regions versus Lower Ks for the WGD Genes Than Singletons in Chromosomal Arms

As described above, the evolutionary rates of soybean genes are generally associated with their chromosomal locations (pericentromeric regions versus chromosomal arms) and their duplication status (WGD genes versus singletons), and both the WGD genes and singletons exhibited biased distribution between chromosomal arms and pericentromeric regions. To understand how and to what extent the WGD genes and singletons evolve differently, it is crucial to compare the two categories of genes in a similar genomic background.

We thus divided the 14,577 WGD gene pairs and 12,994 singletons into two groups based on their chromosomal distribution—pericentromeric regions or chromosomal arms—and performed an intragroup comparison of evolutionary rates (Figures 2 and 3; Table 3). In pericentromeric regions, the average Ks for the WGD genes was found to be significantly or nearly significantly higher than that for singletons. By contrast, the average Ks for the WGD genes was significantly lower than that for singletons in chromosomal arms. Nevertheless, the average Ka for the WGD genes was lower than that for singletons, and the average ω for the WGD genes was lower than that for singletons in both pericentromeric regions and chromosomal arms, indicating a higher level of selective constraint on the WGD genes.

Higher Ks Values for the Singletons in Pericentromeric Regions Than in Chromosomal Arms, and Similar Ks Values for the WGD Genes between the Two Chromatin Environments

To assess further how genomic background influences the paces of gene evolution in the paleopolyploid soybean, we compared the pericentromeric regions and chromosomal arms with regard to the evolutionary rates for the WGD genes and for singletons separately. As shown in Table 4, no significant difference in either Ka or Ks for the WGD genes was detected between the pericentromeric regions and chromosomal arms, suggesting that the evolutionary rates for the WGD genes are not associated or strongly associated with chromatin environment. Singletons also showed similar Ka values between these two chromatin environments, but the average Ks for the singletons in pericentromeric regions is significantly lower than in chromosomal arms. The WGD genes show similar ω values between the two backgrounds, whereas singletons have significantly higher ω values in pericentromeric regions than in chromosomal arms.

Evolutionary Rates for the Two Copies of the WGD Gene Pairs—Asymmetric Evolution between Pericentromeric Regions and Chromosomal Arms

In an attempt to shed light on the evolutionary forces that drive the divergence of the WGD genes, and probably also the nonrandom retention of the WGD genes in distinct chromatin environments, we analyzed and compared the evolutionary rates for the two copies of each of the 2439 WGD gene pairs. Each gene pair is composed of one copy in a chromosomal arm and the other in a pericentromeric region (Table 5). As shown in Figure

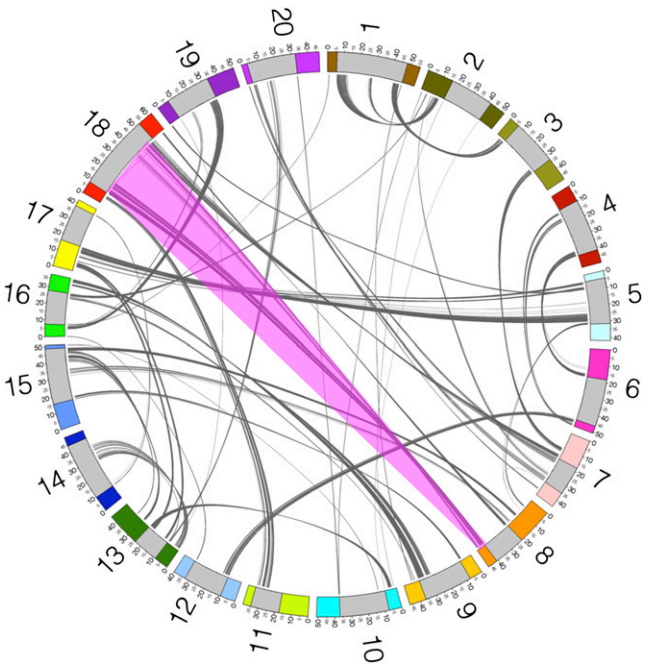

Figure 1. Homoeologous Gene Pairs between Pericentromeric Regions and Chromosomal Arms of Soybean.

The 20 chromosomes are shown in a circle with lines each connecting the two members of each of the 2439 homoeologous gene pairs between pericentromeric regions (coded gray) and chromosomal arms (coded colors). The pink background highlights 212 genes in the pericentromeric region of chromosome 18 versus their proposed homoeologs in the short arm of chromosome 8.

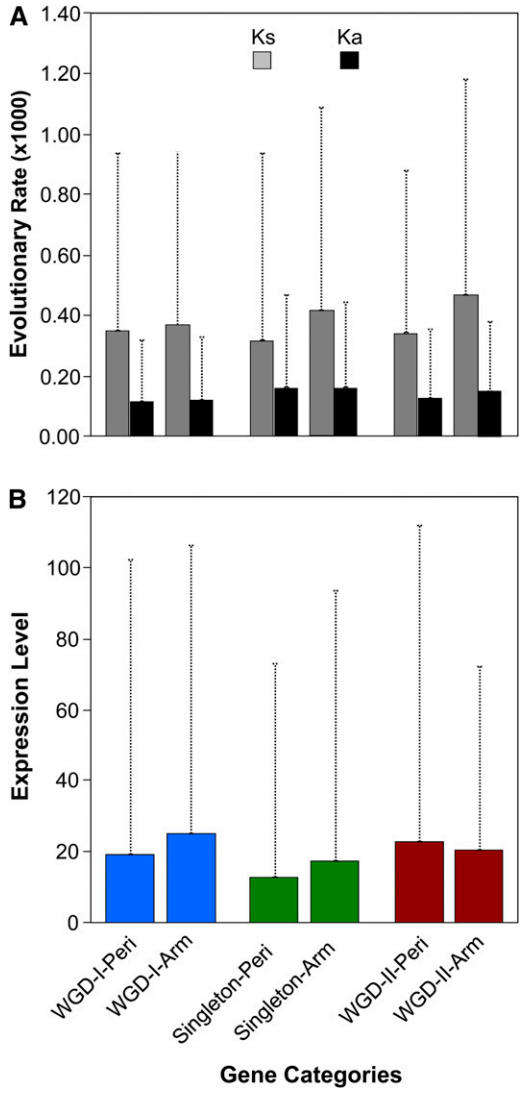

Figure 2. Comparisons of Evolutionary Rates and Expression Levels among Different Gene Categories between Pericentromeric Regions and Chromosomal Arms.

(A) The mean values of Ka and Ks.
(**B**) The average values of gene expression levels. The dotted lines on the top of the bars indicate the upper range of the Ka or Ks values.
[See online article for color version of this figure.]

1, many gene blocks and their predicted homoeologs have been shuffled after the 13-mya WGD event, and most of these corresponding homoeologous gene blocks have experienced complete or partial switches from euchromatic to heterochromatic status or vice versa. A typical example of such switches is reflected by the homoeologous gene pairs located in chromosomes 8 and 18. As illustrated in Figure 1, a cluster of genes (212) in the short arm of chromosome 8 correspond to their homoeologs scattered in the pericentromeric region of chromosome 18.

We found that the mean Ks for the copies of these WGD gene pairs in the chromosomal arms are significantly higher than the mean Ks for their homoeologous copies in the pericentromeric regions. It is particularly interesting that the copies of these WGD gene pairs in the chromosomal arms exhibited the highest levels of Ks in comparison with either singletons or the WGD genes having both copies in chromosomal arms. The copies of the WGD gene pairs in the chromosomal arms also showed a higher level of Ka than their homoeologs in pericentromeric regions and those duplicated genes with both copies in either pericentromeric regions or chromosomal arms.

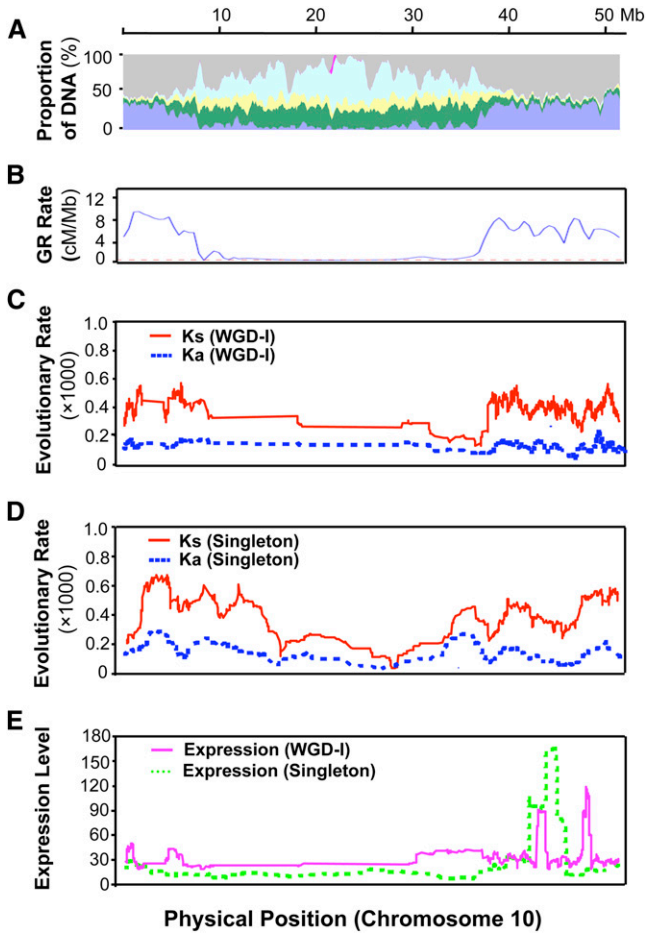

Figure 3. Genomic Features, Evolutionary Rates, and Expression of Genes along Chromosome 10 of Soybean.

(A) The proportions of major DNA components, including genes (blue), DNA transposons (green), Copia-like retrotransposons (yellow), Gypsy-like retrotransposons (cyan), and centromeric satellite repeats (pink).
(B) The variation of local GR rates.
(C) The variation of Ks and Ka of the WGD-I genes along the chromosome.
(D) The variation of Ks and Ka of the singletons along the chromosome. In both panels **(C)** and **(D)**, the Ka and Ks were visualized using 50 continuous genes as a window and one gene as a shift.
(E) The expression levels of the WGD-I genes and singletons along the chromosome.

Table 3. Comparison of the Evolutionary Rates and Expression Levels between WGD Genes and Singletons in the 20 Chromosomes of the Soybean Genome

	Pericentromeric Regions			Chromosomal Arms		
Features[a]	WGD-I Genes[b,c]	Singletons[c]	P Values[d]	WGD-I Genes[b,c]	Singletons[c]	P Values[d]
Ka (×1000)	0.1144 ± 0.2042	0.1598 ± 0.3094	<0.0001	0.1206 ± 0.2086	0.1628 ± 0.2778	<0.0001
Ks (×1000)	0.3500 ± 0.5865	0.3180 ± 0.6168	0.0643	0.3721 ± 0.5694	0.4161 ± 0.6707	<0.0001
ω (Ka/Ks)	0.3332 ± 0.4985	0.4664 ± 0.7660	<0.0001	0.3424 ± 0.8419	0.4124 ± 0.7194	<0.0001
Expression Levels	18.97 ± 83.39	12.63 ± 60.47	<0.0001	25.17 ± 81.22	17.42 ± 76.16	<0.0001

[a]Ka and Ks were calculated by pairwise comparison between *G. max* and *G. soja* accessions.
[b]WGD gene pairs with both members in either pericentromeric regions or chromosomal arms.
[c]Mean ± SD.
[d]Student's *t* test.

Evolutionary Rates for the Two Copies of the WGD Gene Pairs—Asymmetric Evolution between High and Low Recombination Regions of Chromosomal Arms

Theoretically, recombination allows evolution to progress more rapidly (Hartl and Clark, 2007); thus, the lower value of Ks for the copies of the WGD genes in pericentromeric regions versus their homoeologs in chromosomal arms may be explained by the suppression of recombination in the pericentromeric regions. In an attempt to address the potential influence of recombination in shaping the divergence patterns of the two copies of each of the duplicated genes located in chromosomal arms, we first estimated local recombination rates at the two copies of each gene pair (see Methods). To do this, we used MareyMap, an R-based tool for estimating recombination rates by comparison of genetic and physical maps (Rezvoy et al., 2007). The cM/Mb plots along each of the soybean chromosomes were obtained based on 1703 markers that were genetically mapped using a single F2 population of an elite cv Williams 82 and a *G. soja* accession PI479752 and were physically anchored to the soybean (cv Williams 82) reference genome (Schmutz et al., 2010) (see Supplemental Data Set 2 online). We then categorized the two copies of a WGD gene pair into two separate groups—the high recombination group (HRG) and the low recombination group (LRG)—based on the estimated recombination rates at the midpoints of individual genes and subsequently analyzed the Ka, Ks, and ω for the HRG genes and their LRG homoeologs. As shown in Table 5, the HRG genes showed overall higher levels of Ks and Ka than the LRG genes. In particular, significant differences between the Ka values for HRG and for LRG were detected. These results echo the comparative analysis of WGD genes in chromosomal arms and their homoeologs in pericentromeric regions and suggest that recombination is an evolutionary force driving the divergence of duplicated genes in the soybean genome.

Interplay among Evolutionary Rates, Duplication Status, and Gene Expression

In an attempt to shed light on the evolutionary forces for functional divergence of duplicated genes, we compared evolutionary rates and the levels of gene expression within and between pericentromeric regions and chromosomal arms under the same gene categories described above (Figures 2 and 3). The gene expression data from eight different soybean tissues and developmental time points were obtained from Libault et al., (2010). The expression level of a gene was estimated by the average value of all eight samples.

We obtained the following relationships among different gene categories: i) The expression level of genes in pericentromeric regions was significantly lower than in chromosomal arms (Table 1); ii) the expression level of the WGD genes was significantly higher than that of singletons in both pericentromeric regions and

Table 4. Comparison of the Evolutionary Rates and Expression Levels between Pericentromeric Regions and Chromosomal Arms of the 20 Chromosomes of the Soybean Genome

	WGD-I Genes[b]			Singletons		
Features[a]	Pericentromeric Regions[c]	Chromosomal Arms[c]	P Values[d]	Pericentromeric Regions[c]	Chromosomal Arms[c]	P Values[d]
Ka (×1000)	0.1144 ± 0.2042	0.1206 ± 0.2086	0.2427	0.1598 ± 0.3094	0.1628 ± 0.2778	0.5762
Ks (×1000)	0.3500 ± 0.5865	0.3721 ± 0.5694	0.1468	0.3180 ± 0.6168	0.4161 ± 0.6707	<0.0001
ω (Ka/Ks)	0.3332 ± 0.4985	0.3424 ± 0.8419	0.6299	0.4664 ± 0.7660	0.4124 ± 0.7194	0.0072
Expression Levels	18.97 ± 83.39	25.17 ± 81.22	0.0042	12.63 ± 60.47	17.42 ± 76.16	<0.0001

[a]Ka and Ks were calculated by pairwise comparison between *G. max* and *G. soja* accessions.
[b]WGD gene pairs with both members in either pericentromeric regions or chromosomal arms.
[c]Mean ± SD.
[d]Student's *t* test.

Table 5. Comparison of the Evolutionary Rates and Expression Levels between Two Members of Individual WGD Genes

	Two Members of Individual WGD-II Gene Pairs[b]			Two Members of Individual WGD-I Gene Pairs[c]		
Features[a]	One Member in Pericentromeric Regions[d]	The Other Member in Chromosomal Arms[d]	P Values[e]	LRG Members[d]	HRG Members[d]	P Values[e]
Ka (×1000)	0.1241 ± 0.2297	0.1442 ± 0.2228	0.0011	0.1191 ± 0.2107	0.1247 ± 0.2100	0.0342
Ks (×1000)	0.3420 ± 0.5352	0.4682 ± 0.7146	<0.0001	0.3679 ± 0.5611	0.3799 ± 0.5789	0.1059
ω (Ka/Ks)	0.4050 ± 0.7615	0.3742 ± 0.6317	0.3367	0.3434 ± 0.9814	0.3469 ± 0.6692	0.8401
Expression Levels	22.93 ± 88.63	20.35 ± 51.64	0.1102	24.36 ± 68.79	26.00 ± 94.02	0.0105

[a]Ka and Ks were calculated by pairwise comparison between *G. max* and *G. soja* accessions.
[b]Individual WGD gene pairs with one member in a pericentromeric region and the other in a chromosomal arm.
[c]Individual WGD gene pairs in chromosomal arms with one member categorized as LRG and the other as HRG.
[d]Mean ± SD.
[e]Student's paired *t* test.

chromosomal arms (Figure 2; Table 3); iii) the expression level of the WGD genes with both copies of each homoeologous gene pair in pericentromeric regions was significantly lower than that of the WGD genes with both copies of each homoeologous gene pair in chromosomal arms (Figure 2; Table 4); and iv) the expression level of singletons in pericentromeric regions was significantly lower than in chromosomal arms (Figure 2; Table 4). These observations together with the analyses of evolutionary rates described earlier (Figure 2; Tables 1, 3, and 4) suggest that the overall expression levels of genes are associated with both duplication status and recombination rates but are not associated with Ka. These observations seem to echo a recent study that suggests that the biased fractionation of the WGD genes in maize is associated with the bias in gene expression (Schnable et al. 2011).

When the two copies of a WGD gene pair were compared, we obtained seemingly inconsistent results between two categories of WGD genes (Figure 2; Table 5). On average, the expression level of gene copies in pericentromeric regions was even higher than that of their homoeologs in chromosomal arms, although the difference was not statistically significant (P = 0.1102) (Table 5). This higher level of gene expression seems to be associated with the lower Ka (Table 5). By contrast, the expression level of gene copies in the LRG of chromosomal arms was significantly lower than that of their homoeologs in the HRG of chromosomal arms. Thus, the high level of expression versus low Ka is likely to be a unique feature of a subset of genes that occur in pericentromeric regions in contrast with their homoeologs in chromosomal arms.

We further analyzed potential correlations among local GR rates, evolutionary rates, and the levels of gene expression in chromosomal arms using the WGD-I genes and singletons as two independent data sets. The local GR rate of individual genes was estimated based on a total of 1703 markers generated from a single mapping population (see Supplemental Data Set 2 online) (see Methods). As shown in Table 6, for both WGD genes and singletons, Ka was positively correlated with the levels of gene expression, whereas no significant correlations of local GR rates to evolutionary rates and gene expression were detected in chromosomal arms. This suggests that the short divergence time between *G. max* and *G. soja* may limit our capability to detect this correlation. Likewise, our data and approach do not allow accurate estimation of local GR rates.

Patterns of Divergence, Distribution, and Expression of Transcription Factors—Difference and Consistence

Previous studies indicated that the genes retained as duplicated pairs after WGD events tend to belong to specific classes, such as transcription factors and members of large multiprotein complexes (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005). To understand whether and/or to what extent these specific genes are consistent or inconsistent with the general patterns of soybean gene evolution revealed by analysis of the 14,577 WGD genes and 12,994 singletons, we examined the divergence, retention, and expression of the retained transcription factor gene pairs versus transcription factor singletons in the soybean genome.

Based on the putative soybean transcription factors database SoyDB, which was constructed by Wang et al. (2010), 1742 of the 14,577 WGD gene pairs and 1172 of the 12,944 singletons are annotated as putative transcription factors. When the WGD-II transcription factors were excluded, the ratios of duplicated transcription factor pairs to singletons were 1:4.8 in pericentromeric regions versus 1.85:1 in chromosomal arms (see Supplemental

Table 6. Correlation among Ka, Ks, GR Rates, and Expression Levels of Genes in Chromosomal Arms

	WGD-I Genes[b,c]		Singletons[c]	
Comparison[a]	*r*	P Value	*r*	P Value
Ka versus Ks	0.2157	<0.0001	0.2806	<0.0001
Ka versus ω	0.5052	<0.0001	0.5284	<0.0001
Ks versus ω	−0.1993	<0.0001	−0.2145	<0.0001
Ka versus expression	−0.0708	<0.0001	0.0464	0.0001
Ks versus expression	−0.0013	0.8522	−0.0040	0.7410
ω versus expression	−0.0520	<0.0001	−0.0568	0.0005
GR rates versus Ka	−0.0056	0.4266	0.0096	0.4289
GR rates versus Ks	−0.0116	0.0961	0.0199	0.1011
GR rates versus ω	0.0048	0.6067	−0.0006	0.9734
GR rates versus expression	0.0038	0.5898	0.0192	0.1146

[a]Ka and Ks were calculated by pairwise comparison between *G. max* and *G. soja* accessions.
[b]WGD gene pairs with both members in chromosomal arms.
[c]Pearson's correlation test.

Table 2 online), corresponding to approximately nine times higher accumulation rates of transcription factor singletons in the former than in the latter regions. This distribution pattern of transcription factor genes is similar to that observed for the complete set of the soybean genes indicated in Table 2. Although the differences of Ka and Ks and expression levels between the WGD-I transcription factor gene pairs and singletons in either pericentromeric regions or chromosomal arms were found to be less significant (see Supplemental Table 3 online) than detected between all WGD-I genes and all singletons in the soybean genome (Table 3), the patterns revealed by both data sets are consistent. In addition, the patterns of divergence and levels of expression between two members of individual WGD-II transcription gene pairs (see Supplemental Table 4 online) are overall consistent as revealed by analysis of all WGD-II genes in the soybean genome (Table 5).

DISCUSSION

Duplication Status, Rather Than Recombination, as the Primary Determinant of Ka

Previous studies revealed slower evolution of duplicates than singletons in eukaryotes (Yang et al., 2003; Davis and Petrov, 2004; Jordan et al., 2004; Yang and Gaut, 2011). However, few studies investigated the effects of local recombination rates on evolutionary rates or compared these two parameters on a similar time scale. As a result, potential interplay among recombination, duplication status, and evolutionary rates was largely unclear. By comparing the same and different categories of genes within and between pericentromeric regions and chromosomal arms—two extreme samples of the cold spots and hotspots for recombination—we found that WGD genes consistently showed significant lower Ka than singletons in both regions (Table 3) and that neither the WGD genes nor singletons exhibited significant difference of Ka between these two regions (Table 4). Because *G. soja* and *G. max* diverged quite recently (Kim et al. 2010), it would be reasonable to speculate that these two species maintain similar genomic landscapes, including the boundaries of recombination-suppressed pericentromeric regions of orthologous chromosomes in these genomes. This speculation is supported by the general consensus of soybean genetic maps independently developed using different mapping populations derived from crosses between *G. max* and *G. max* and between *G. max* and *G. soja* (Hyten et al., 2010). Thus, the observations that we described above would be interpreted as the most convincing evidence that has been garnered from any eukaryotic organisms in support of the conclusion that duplication status, rather than recombination, is the primary determinant of Ka.

The slower evolution of WGD genes than singletons observed in soybean and several other eukaryotes is supportive of the gene balance theory, which predicts that maintaining proper balance in the concentrations of protein subunits in a macromolecular complex and members of regulatory networks and highly connected portions of signaling networks is vital to maintain normal function and that an imbalance may lead to either decreased fitness or lethality (Birchler and Veitia, 2007; Freeling, 2008; Veitia et al., 2008; Edger and Pires, 2009). However, it seems to contradict a classical model, which predicts that one or the other copy of a duplicated gene pair with redundant functions can accumulate deleterious mutations and eventually be lost without effect on the fitness of an individual (Walsh, 1995). Nevertheless, the duplicated genes may show distinct paces and patterns of evolution within different timeframes after WGD events (Jordan et al. 2004).

Although the mean Ka values of neither the WGD genes nor the singletons were statistically different between pericentromeric regions and chromosomal arms, the potential influence of recombination on Ka should not be fully disregarded. Indeed, we observed lower Ka for both WGD genes and singletons in the former than in the latter regions. The lack of statistically detected difference may simply reflect the limits of statistical power in detecting potential correlation between recombination rates and Ka, in particular, given such a short divergence time of *G. max* and *G. soja*. A recent study in three *Oryza* species compared 13 genes in the functional centromeric region of rice chromosome 8 (Cen8) and 1515 genes dispersed on the short arms of chromosome 3 (Chr3S), and revealed that the mean values of Ka and Ks for Cen8 were significantly lower than those for genes in Chr3S (Fan et al., 2011). However, because the duplication status of these genes was not investigated, whether the detected difference in evolutionary rates (e.g., Ka) was caused by recombination remains obscure. It is also unclear whether those two sets of genes are representative of the whole genome. Of course, other centromeric and pericentromeric attributes, such as the overall low density of genes, high density of TEs (Tian et al., 2009), high levels of DNA methylation (Warburton, 2004; Lister et al., 2008) and histone modification (Cokus et al., 2008; Stimpson and Sullivan, 2011), and low levels of gene expression (Libault et al., 2010) may also be contributing factors to variation in the rates of evolution (Yang and Gaut, 2011).

Does Recombination Facilitate Single Nucleotide Mutation?

The relationships between recombination and mutation rates have been investigated in several organisms, but whether recombination facilitates the generation of single nucleotide mutation remains equivocal (Gaut et al., 2007). This may be partially because of inaccurate estimation of local recombination rates (Gaut et al., 2007) or inappropriate comparison of recombination rates and evolutionary rates (e.g., the two parameters were evaluated on different time scales) (Clément et al., 2006).

If one assumes that selection does not act on synonymous sites of a gene based on the neutral theory of molecular evolution, then Ks should be equal to the mutation rate of the gene (Kimura, 1968; Hurst, 2002; Chamary et al., 2006; Duret, 2009). Under this assumption, the overall smaller value of the mean Ks of genes in pericentromeric regions than in chromosomal arms would be interpreted as a lower mutation rate in the former than in the latter (Table 1), perhaps as a function of recombination rates (Begun and Aquadro, 1992; Gaut et al., 2007; Yang and Gaut, 2011). Indeed, levels of nucleotide variability, which may be partially explained by mutation rates, have often been found to be positively correlated with recombination rates, because of genetic hitchhiking of neutral (e.g., synonymous) sites with linked selected nonsynonymous sites (Lercher and Hurst, 2002;

Spencer et al., 2006). Intriguingly, the WGD genes and singletons exhibited different patterns of Ks variation within and between the two chromatin environments (Tables 3 and 4), suggesting that, beyond recombination, the duplicated status of genes may also be a factor influencing the rates of synonymous mutations.

Pericentromeric Effects on the Divergence of Homoeologous Gene Pairs

One of the most remarkable observations is the asymmetric evolution of homoeologous genes between pericentromeric regions and chromosomal arms (Figure 2; Table 5). Although neither the WGD genes nor singletons showed significant difference of Ka between these two distinct regions (Table 4), the Ka value of 2439 genes in pericentromeric regions was found to be significantly lower than their homoeologous copies in chromosomal arms, as was Ks (Table 5). An evolutionary model predicts that the recombination-suppressed pericentromeric regions should allow more rapid accumulation of deleterious mutations than chromosomal arms, because of Hill-Robertson effects (Hill and Robertson, 1966); however, recombination can also facilitate single nucleotide mutations (Lercher and Hurst, 2002; Jelesko et al., 2004; Schuermann et al., 2005). Therefore, the slow evolution of gene copies in pericentromeric regions versus their homoeologous copies in chromosomal arms may be explained by strong purifying selection and reduced mutation rates in the pericentromeric regions. This interpretation echoes a recent study that demonstrated that the 13 genes in a centromeric region were under strong purifying selection compared with genes located on chromosomal arms of three *Oryza* species, although the duplication status of these genes was not investigated (Fan et al., 2011). Variation in mutation rates may also be influenced by other factors, such as gene and transposon densities, chromatin structure, and the pattern of gene expression (Ellegren et al., 2003; Wolfe and Li, 2003; Baer et al., 2007; Lin et al., 2010; Yang and Gaut, 2011).

Biased Expression of Homologous Gene Pairs

Transcriptome atlases of plant genomes have revealed generally lower levels of gene expression in pericentromeric regions than in chromosomal arms (Zeller et al., 2009; Zhang et al., 2010; Libault et al., 2010). As expected, both WGD genes and singletons show lower levels of gene expression in pericentromeric regions than in chromosomal arms of the soybean genome. These reduced levels of gene expression may be associated with the biased distribution of methylated DNA, which is largely associated with transposable elements (Hollister and Gaut, 2009) and with histone modification and chromatin structure (Karlić et al., 2010) in pericentromeric regions.

Given the reduction of expression levels of the WGD genes and singletons, it was extremely intriguing that the expression level of the 2439 genes in pericentromeric regions was higher than that of their homoeologous copies in chromosomal arms, although a significant difference was not statistically detected. Nevertheless, the asymmetric expression of these homoeologous genes was consistent with their asymmetric divergence, indicating different levels of functional constraints between genes in pericentromeric regions and their homologous copies in chromosomal arms. It is also notable that the 2439 genes in pericentromeric regions showed a higher expression level than any other categories of genes in the same regions (Figure 2; Tables 3 to 5).

Pericentromeric Effects on Biased Fractionation of Homoeologous Genes

The high level of expression and low level of divergence of the 2439 genes in pericentromeric regions versus their homoeologs in chromosomal arms represents a striking observation. Because the member of a gene pair that is expressed at a lower level and evolves at a faster pace tends to be deleted more easily than the other member of the pair, members of individual homologous gene pairs in pericentromeric regions would have survived longer than their homoeologs in chromosomal arms during host genome evolution. This deduction seems to be supported by the observation that the ratios of singletons to WGD genes in the former are ~7.3 times higher than in the latter regions (Table 2). It has been suggested that the most recent WGD of soybean and maize occurred at a similar time (Schlueter et al., 2004; Swigonová et al. 2004), followed by biased fractionation in both genomes (Schlueter et al. 2006; Thomas et al. 2006; Woodhouse et al., 2010). If biased retention of WGD genes in soybean is predominantly caused by single gene deletions instead of relocations, as observed in maize (Woodhouse et al., 2010), then the high proportion of singletons in pericentromeric regions of soybean is likely to be an outcome of biased deletion of their homoeologs in chromosomal arms. We should note that the biased accumulation of singletons in pericentromeric regions may also be, at least partially, caused by preferential insertions of genes in pericentromeric regions, as observed in *A. thaliana* (Freeling et al. 2008).

Analysis of the duplicated regions in the soybean genome has revealed extensive genomic rearrangements (Schmutz et al., 2010; Severin et al., 2011). It is obvious that those rearrangements have substantially reshaped the landscape of the soybean genome in the past 13 million years, leading to dramatic differentiation of genomic features between duplicated regions over evolutionary time (Schmutz et al., 2010), including transitions from euchromatin (e.g., chromosomal arms) to heterochromatin (e.g., pericentromeric regions) or vice versa (Figure 1). Asymmetric evolution of duplicated genes was also observed between ~1-Mb homoeologous regions from two chromosomal arms of soybean, and such asymmetry seems to be associated with diverged genomic features, such as local genomic rearrangement and content of transposable elements. Genomic reshuffling and restructuring occurs over evolutionary time; thus, it is certain that the divergence between two members of individual WGD gene pairs and preferential retention of WGD genes both are dynamic evolutionary processes.

METHODS

Characterization of Homoeologous Gene Pairs and Singletons

Of the 46,430 high-confidence genes identified in the soybean (*Glycine max*) genome, 31,264 were predicted to exist as paralogs (Schmutz et al., 2010). To increase the accuracy of the prediction of WGD gene pairs, we

included only gene pairs and singletons located in the 149 large duplicated genomic segments. These gene pairs were supported not only by sequence homology, but also by a syntenic relationship and were deemed to be formed by the allopolyploidy event that occurred ~13 mya. The orthologs of these genes in the 31 resequenced *G. max* and *Glycine soja* genomes were extracted from the sequences previously described by Lam et al. (2010). The gene sequences from the reference genome and the resequenced genomes were aligned using the MUSCLE program (Edgar, 2004). Gene sequences from the resequenced genome with missing portions greater than 50% of their full lengths (estimated based on the reference genome) and gene sequences that contained frame shift mutations or stop codons were excluded from further analysis. Ambiguous sites were coded as "N" to minimize possible sequencing errors and/or inaccurate assembly. The final data set in this study includes 14,577 WGD gene pairs and 12,994 singletons. These genes were classified into several categories: i) WGD gene pairs with both duplicated copies located in pericentromeric regions; ii) WGD gene pairs with both duplicated copies located in chromosomal arms; iii) WGD gene pairs with one copy in a pericentromeric region and the other in a chromosomal arm, as illustrated in Figure 1 using Circos (Krzywinski et al., 2009); iv) singletons in pericentromeric regions; and v) singletons in chromosomal arms.

Analysis of Sequence Divergence

The predicted coding sequences were used for gene divergence analysis. The Ks, Ka, and their ratio ω (i.e., Ka/Ks), were estimated using the yn00 module integrated in the PAML package (Yang, 2007). Ks and Ka were calculated by pairwise comparison of orthologous gene sequences between *G. max* and *G. soja*, which may reduce potential effects of unavoidable errors in mapping of short resequencing reads to the reference genome sequence on evaluation of sequence divergence between *G. soja* and *G. max* populations.

Estimation of Local GR Rates

The local GR rates were estimated using MareyMap (Rezvoy et al., 2007). A total of 1703 markers generated from a single mapping population (Williams 82 × G. *soja* PI 479752) (Hyten et al., 2010) were used for local GR rate estimation. The GR rate at the midpoint of each gene was used for analysis of potential correlation with the evolutionary rate and level of gene expression. The GR-suppressed pericentromeric regions were defined based on the comparison of soybean physical and genetic maps as previously described (Schmutz et al., 2010). The average GR rate for a particular region (e.g., pericentromeric region) was calculated based on the genetic distance and physical distance that the region spanned (see Supplemental Data Set 1 online).

Analysis of Gene Expression

The whole genome transcriptomic data from eight tissues of Williams 82 were generated by Libault et al. (2010) and were used to evaluate the expression level of the genes analyzed in our study. The expression level of a gene was calculated as the average value of the levels of expression in the eight samples. The expression level of a gene in each sample was measured by the number of Illumina/Solexa reads per million reads uniquely aligned to the gene as described by Libault et al. (2010).

Statistical Analyses—Correlation and Student's *t* Test

Comparison of genomic features, evolutionary rates, and expression levels between pericentromeric regions and chromosomal arms, between the WGD genes and singletons, and between two members of individual WGD gene pairs were conducted using Student's *t* test. The correlations of evolutionary rates with local GR rates and expression levels of genes were assessed using Pearson's correlation by 10,000 bootstrap resampling using the SAS software.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Physical and Genetic Lengths of Pericentromeric Regions and Arms of the 20 Chromosomes of Soybean.

Supplemental Figure 2. The Variation of Local GR Rates and Evolutionary Rates along the 20 Chromosomes of Soybean.

Supplemental Table 1. Comparison of Recombination and Physical Lengths of Pericentromeric Regions and Chromosomal Arms of the 20 Chromosomes.

Supplemental Table 2. Distribution of the WGD and Singleton Transcription Factors in the Pericentromeric Regions and Chromosomal Arms of Soybean.

Supplemental Table 3. Comparison of the Evolutionary Rates and Expression Levels between the WGD and Singleton Transcription Factors in the 20 Chromosomes of the Soybean Genome.

Supplemental Table 4. Comparison of the Evolutionary Rates and Expression Levels between Two Members of Individual WGD-II Transcription Factor Gene Pairs.

Supplemental Data Set 1. Genes, Location, Local GR Rates, Evolutionary Rates, and Duplication Status.

Supplemental Data Set 2. Genetic and Physical Locations of Molecular Markers for Estimation of Local Recombination Rates.

ACKNOWLEDGMENTS

We thank Hon-Ming Lam and Xin Liu for providing the soybean genome resequencing data, and Brandon Gaut and Michael Purugganan for their help interpreting some observations reported in this study. This work was partially supported by Indiana Soybean Alliance (J.M.), National Science Foundation Plant Genome Research Program (IOS-0822258) (J.M., P.C.), Purdue Agricultural Research Award (J.M.), and Jiangsu Academy of Agricultural Sciences Startup Funds (J.D.).

AUTHOR CONTRIBUTIONS

J.D. performed the research and analyzed the data; Z.T., Y.S., M.Z., Q.S., S.B.C., and P.C. contributed new analytic tools. J.M. designed the research, analyzed the data, and wrote the article.

Received October 14, 2011; revised December 7, 2011; accepted December 20, 2011; published January 6, 2012.

REFERENCES

Baer, C.F., Miyamoto, M.M., and Denver, D.R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev. Genet. **8:** 619–631.

Begun, D.J., and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356:** 519–520.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester,

S.R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana.* Proc. Natl. Acad. Sci. USA **107:** 18724–18728.

Bennetzen, J.L., Ma, J., and Devos, K.M. (2005). Mechanisms of recent genome size variation in flowering plants. Ann. Bot. (Lond.) **95:** 127–132.

Birchler, J.A., and Veitia, R.A. (2007). The gene balance hypothesis: From classical genetics to modern genomics. Plant Cell **19:** 395–402.

Blanc, G., and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667–1678.

Carter, T.E., Nelson, R.L., Sneller, C.H., and Cui, Z. (2004). Genetic diversity in soybean. In Soybeans: Improvement, Production, and Uses, 3rd ed, H.R. Boerma and J.E. Specht, eds. Agron. Monogr. No. 16. (Madison, WI: American Society of Agronomy), pp. 303–416.

Chamary, J.V., Parmley, J.L., and Hurst, L.D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. **7:** 98–108.

Clément, Y., Tavares, R., and Marais, G.A. (2006). Does lack of recombination enhance asymmetric evolution among duplicate genes? Insights from the Drosophila melanogaster genome. Gene **385:** 89–95.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature **452:** 215–219.

Davis, J.C., and Petrov, D.A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. **2:** E55.

De Bodt, S., Maere, S., and Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. Trends Ecol. Evol. (Amst.) **20:** 591–597.

Doyle, J.J., and Egan, A.N. (2010). Dating the origins of polyploidy events. New Phytol. **186:** 73–85.

Du, J., Grant, D., Tian, Z., Nelson, R.T., Zhu, L., Shoemaker, R.C., and Ma, J. (2010). SoyTEdb: A comprehensive database of transposable elements in the soybean genome. BMC Genomics **11:** 113.

Duret, L. (2009). Mutation patterns in the human genome: More variable than expected. PLoS Biol. **7:** e1000028.

Edgar, R.C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:** 113.

Edger, P.P., and Pires, J.C. (2009). Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. **17:** 699–717.

Ellegren, H., Smith, N.G., and Webster, M.T. (2003). Mutation rate variation in the mammalian genome. Curr. Opin. Genet. Dev. **13:** 562–568.

Fan, C., Walling, J.G., Zhang, J., Hirsch, C.D., Jiang, J., and Wing, R.A. (2011). Conservation and purifying selection of transcribed genes located in a rice centromere. Plant Cell **23:** 2821–2830.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.

Freeling, M. (2008). The evolutionary position of subfunctionalization, downgraded. Genome Dyn. **4:** 25–40.

Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., and Lisch, D. (2008). Many or most genes in *Arabidopsis* transposed after the origin of the order *Brassicales.* Genome Res. **18:** 1924–1937.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J., and Anderson, L.K. (2007). Recombination: An underappreciated factor in the evolution of plant genomes. Nat. Rev. Genet. **8:** 77–84.

Gill, N., Findley, S., Walling, J.G., Hans, C., Ma, J., Doyle, J., Stacey, G., and Jackson, S.A. (2009). Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. **151:** 1167–1174.

Gu, X., Zhang, Z., and Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc. Natl. Acad. Sci. USA **102:** 707–712.

Ha, M., Kim, E.D., and Chen, Z.J. (2009). Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc. Natl. Acad. Sci. USA **106:** 2295–2300.

Hartl, D.L., and Clark, A.G. (2007). Principles of Population Genetics, 4th ed. (Sunderland, MA: Sinauer Associates).

He, X., and Zhang, J. (2005). Gene complexity and gene duplicability. Curr. Biol. **15:** 1016–1021.

Hill, W.G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Hollister, J., and Gaut, B.S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. **19:** 1419–1428.

Hurst, L.D. (2002). The Ka/Ks ratio: Diagnosing the form of sequence evolution. Trends Genet. **18:** 486.

Hyten, D.L., Cannon, S.B., Song, Q., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D., and Cregan, P.B. (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics **11:** 38.

Jelesko, J.G., Carter, K., Thompson, W., Kinoshita, Y., and Gruissem, W. (2004). Meiotic recombination between paralogous *RBCSB* genes on sister chromatids of *Arabidopsis thaliana.* Genetics **166:** 947–957.

Jiao, Y., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100.

Jordan, I.K., Wolf, Y.I., and Koonin, E.V. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol. Biol. **4:** 22.

Karlić, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci. USA **107:** 2926–2931.

Kim, M.Y., et al. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. USA **107:** 22032–22037.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature **217:** 624–626.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002). Selection in the evolution of gene duplications. Genome Biol. **3:** research0008.1–research0008.9.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. **19:** 1639–1645.

Lam, H.M., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. **42:** 1053–1059.

Lercher, M.J., and Hurst, L.D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. **18:** 337–340.

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R.J., Franklin, L.D., He, J., Xu, D., May, G., and Stacey, G. (2010). An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. Plant J. **63:** 86–99.

Lin, J.Y., Stupar, R.M., Hans, C., Hyten, D.L., and Jackson, S.A. (2010). Structural and functional divergence of a 1-Mb duplicated region in the soybean (*Glycine max*) genome and comparison to an orthologous region from *Phaseolus vulgaris.* Plant Cell **22:** 2545–2561.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry,

C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis.* Cell **133:** 523–536.

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

Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science **302:** 1401–1404.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA **102:** 5454–5459.

Ohno, S. (1970). Evolution by Gene Duplication. (New York: Springer-Verlag), p. 160.

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. Annu. Rev. Genet. **34:** 401–437.

Rezvoy, C., Charif, D., Guéguen, L., and Marais, G.A. (2007). MareyMap: An R-based tool with graphical interface for estimating recombination rates. Bioinformatics **23:** 2188–2189.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., and Shoemaker, R.C. (2004). Mining EST databases to resolve evolutionary events in major crop species. Genome **47:** 868–876.

Schlueter, J.A., Scheffler, B.E., Schlueter, S.D., and Shoemaker, R.C. (2006). Sequence conservation of homeologous BACs and expression of homeologous genes in soybean (*Glycine max* L Merr). Genetics **174:** 1017–1028.

Schmutz, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463:** 178–183.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108:** 4069–4074.

Schuermann, D., Molinier, J., Fritsch, O., and Hohn, B. (2005). The dual nature of homologous recombination in plants. Trends Genet. **21:** 172–181.

Sémon, M., and Wolfe, K.H. (2007). Consequences of genome duplication. Curr. Opin. Genet. Dev. **17:** 505–512.

Seoighe, C., and Gehring, C. (2004). Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet. **20:** 461–464.

Severin, A.J., Cannon, S.B., Graham, M.M., Grant, D., and Shoemaker, R.C. (2011). Changes in twelve homoeologous genomic regions in soybean following three rounds of polyploidy. Plant Cell **23:** 3129–3136.

Soltis, D.E., and Soltis, P.S. (1999). Polyploidy: Recurrent formation and genome evolution. Trends Ecol. Evol. (Amst.) **14:** 348–352.

Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. Annu. Rev. Plant Biol. **60:** 561–588.

Spencer, C.C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The influence of recombination on human genetic diversity. PLoS Genet. **2:** e148.

Stimpson, K.M., and Sullivan, B.A. (2011). Histone H3K4 methylation keeps centromeres open for business. EMBO J. **30:** 233–234.

Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. Genome Res. **14:** 1916–1923.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. **16:** 934–946.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S., and Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. **19:** 2221–2230.

Veitia, R.A., Bottani, S., and Birchler, J.A. (2008). Cellular reactions to gene dosage imbalance: Genomic, transcriptomic and proteomic effects. Trends Genet. **24:** 390–397.

Walsh, J.B. (1995). How often do duplicated genes evolve new functions? Genetics **139:** 421–428.

Wang, Z., Libault, M., Joshi, T., Valliyodan, B., Nguyen, H.T., Xu, D., Stacey, G., and Cheng, J. (2010). SoyDB: A knowledge database of soybean transcription factors. BMC Plant Biol. **10:** 14.

Warburton, P.E. (2004). Centromeric heterochromatin comes clean with DNA methylation. Nature Methods **1:** 14–15.

Wendel, J.F. (2000). Genome evolution in polyploids. Plant Mol. Biol. **42:** 225–249.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2:** 333–341.

Wolfe, K.H., and Li, W.H. (2003). Molecular evolution meets the genomics revolution. Nat. Genet. **33**(Suppl): 255–265.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S., and Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol. **8:** e1000409.

Yang, J., Gu, Z., and Li, W.H. (2003). Rate of protein evolution versus fitness effect of gene deletion. Mol. Biol. Evol. **20:** 772–774.

Yang, L., and Gaut, B.S. (2011). Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol. Biol. Evol. **28:** 2359–2369.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586–1591.

Zeller, G., Henz, S.R., Widmer, C.K., Sachsenberg, T., Rätsch, G., Weigel, D., and Laubinger, S. (2009). Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole-genome tiling arrays. Plant J. **58:** 1068–1082.

Zhang, G., et al. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res. **20:** 646–654.