The fate of duplicated genes in a polyploid plant genome

Anne Roulin[1,2], Paul L. Auer[3,4,†], Marc Libault[5,6,†], Jessica Schlueter[1,7,†], Andrew Farmer[8], Greg May[8], Gary Stacey[5], Rebecca W. Doerge[3] and Scott A. Jackson[1,*]

[1]*Institute for Plant Breeding, Genetics and Genomics, University of Georgia, 111 Riverbend Road, Athens, GA, 30602, USA,*

[2]*Zoologisches Institut, Universität Basel, Vesalgasse 1, CH-4051, Basel, Switzerland,*

[3]*Department of Statistics, Purdue University, West Lafayette, IN, 47907, USA,*

[4]*Fred Hutchinson Cancer Research Center, Seattle, WA, USA,*

[5]*Divisions of Plant Science and Biochemistry, University of Missouri, Columbia, MO, 65211, USA,*

[6]*Department of Botany and Microbiology, University of Oklahoma, Norman, OK, 73019, USA,*

[7]*College of Computing and Informatics, University of North Carolina Charlotte, Charlotte, NC, 28223, USA, and*

[8]*National Center for Genome Resources, Santa Fe, NM, USA*

SUMMARY

Polyploidy is generally not tolerated in animals, but is widespread in plant genomes and may result in extensive genetic redundancy. The fate of duplicated genes is poorly understood, both functionally and evolutionarily. Soybean (*Glycine max* L.) has undergone two separate polyploidy events (13 and 59 million years ago) that have resulted in 75% of its genes being present in multiple copies. It therefore constitutes a good model to study the impact of whole-genome duplication on gene expression. Using RNA-seq, we tested the functional fate of a set of approximately 18 000 duplicated genes. Across seven tissues tested, approximately 50% of paralogs were differentially expressed and thus had undergone expression sub-functionalization. Based on gene ontology and expression data, our analysis also revealed that only a small proportion of the duplicated genes have been neo-functionalized or non-functionalized. In addition, duplicated genes were often found in collinear blocks, and several blocks of duplicated genes were co-regulated, suggesting some type of epigenetic or positional regulation. We also found that transcription factors and ribosomal protein genes were differentially expressed in many tissues, suggesting that the main consequence of polyploidy in soybean may be at the regulatory level.

Keywords: polyploidy, duplicated gene expression, sub-functionalization, *Glycine max*, RNA-seq, genome evolution.

INTRODUCTION

Angiosperms represent the largest group of plants, with 350 000 known taxa (Van de Peer *et al.*, 2009). They underwent diversification in the mid-Cretaceous period (i.e. 100 million years ago, MYA), and, in contrast to pteridophytes and gymnosperms, maintained a high radiation rate over a long period of time (Lidgard and Crane, 1988; Crane and Lidgard, 1989; Crepet and Niklas, 2009). Defined by Darwin as an 'abominable mystery', this prominence of flowering plants on earth has been extensively studied. Recent theories suggest that carpel evolution, double fertilization and flower development, as well as additional innovations such as reduced cost of seed production and short generation time, contributed to the explosive success of angiosperms (Stuessy, 2004; Lord and Westoby, 2011). Because many

genes involved in reproduction and flower development were duplicated before the monocot/dicot radiation (Jiao *et al.*, 2011), whole-genome duplications (WGDs) are believed to be at the origin of angiosperm radiation (De Bodt *et al.*, 2005). Polyploidy, or WGD, is a process that recurrently shaped eukaryotic genomes. Although, in animals, this process is mainly restricted to amphibians and fish (Otto and Whitton, 2000), polyploidy has played a major evolutionary role in plants. Complete genome analyses strongly support the conclusion that, in addition to lineage-specific WGDs, a triplication (γ) and two WGD (ρ and σ), respectively, occurred in eudicots and monocots (Vision, 2000; Jaillon *et al.*, 2007; Lyons *et al.*, 2008; Tang *et al.*, 2010). Recent work also demonstrated that two

rounds of WGD occurred 319 and 192 MYA, respectively, shortly before seed plant and angiosperm radiation (Jiao *et al.*, 2011).

Models have been proposed in which duplicated genes are pseudogenized (loss of regulatory sub-function; Moore and Purugganan, 2005), sub-functionalized (partitioning of the function between daughter copies; Cusack and Wolfe, 2007) and/or neo-functionalized (functional diversification; Blanc and Wolfe, 2004). These provide testable hypotheses suggesting that neo-functionalized gene copies undergo positive selection ($K_a/K_s > 1$), whereas, sub-functionalized gene copies, showing transcriptional divergence across tissues or cell types, are expected to undergo purifying selection ($K_a/K_s < 1$). Additionally, because redundancy allows one of the copies to accumulate mutations without an immediate effect on the fitness of the organism, polyploidy may give rise to new allelic variants (Feldman and Levy, 2009), gene family expansion (Zahn *et al.*, 2005; Veron *et al.*, 2007) and changes in gene expression (Levy and Feldman, 2004; Rapp *et al.*, 2009; Wang *et al.*, 2012). Moreover, polyploidy has been shown to be associated with methylation changes (Yu *et al.*, 2010; Hegarty *et al.*, 2011; Kenan-Eichler *et al.*, 2011; Zhao *et al.*, 2011) and potentially activation of transposable elements (Kenan-Eichler *et al.*, 2011). These genetic and epigenetic modifications undoubtedly lead to evolution of new traits and thus increased adaptability. This may explain why some newly established polyploids have a competitive advantage compared to their diploid parents. As an example, the recent allopolyploid *Spartina anglica*, which occurred in the late 19th century, rapidly invaded the British and French coasts and has gradually replaced its diploid parents (Baumel *et al.*, 2001, 2002). Fawcett *et al.* (2009) also argued that this capacity to adapt and colonize new niches may have been responsible for angiosperm radiation and their survival during the Cretaceous–Tertiary (KT) crisis. Nevertheless, after duplication, many polyploids gradually return to diploid-like chromosome pairing through accumulation of rearrangements and sequence deletions (Blanc and Wolfe, 2004; Levy and Feldman, 2004; Feldman and Levy, 2009; Hufton and Panopoulou, 2009; Tate *et al.*, 2009; Buggs *et al.*, 2010; Xiong *et al.*, 2011), a process referred as diploidization. However, paleopolyploid genomes may retain several copies of the same gene (Tate *et al.*, 2009), and thus provide an opportunity to study the fate of duplicated genes.

Soybean, *Glycine max*, is part of the legume family, which is one of the largest families in flowering plants, with more than 20 000 species (Doyle and Luckow, 2003). Due to its agricultural prominence, the *Glycine max* genome has been fully sequenced (Schmutz *et al.*, 2010), and provides an opportunity to explore the molecular effects of genome duplication, as it has experienced at least two polypoid events, the most ancient being 59 MYA.

The more recent event was probably an allopolyploid (Gill *et al.*, 2009), and is the result of a merger of two genomes that diverged approximately 13 MYA and reunited approximately 5–10 MYA when the genus was formed, as this polyploidy event is found only in the genus *Glycine* (Doyle *et al.*, 2003; Straub *et al.*, 2006; Innes *et al.*, 2008; Stefanović *et al.*, 2009). These duplication events result in a genome with approximately 46 430 'high-confidence' genes, of which 75% are present as more than one copy. We took advantage of the availability of these genomic resources to study the evolution of duplicated genes. It is likely that differential expression of duplicates among tissues varies and contributes to phenotypic variation in polyploids (Buggs *et al.*, 2010). However, relatively few studies have investigated this process (Adams *et al.*, 2003; Flagel *et al.*, 2008; Chaudhary *et al.*, 2009; Buggs *et al.*, 2010; Flagel and Wendel, 2010). Using RNA-seq, we tested for transcriptional divergence of 17 500 gene pairs using expression data from seven tissues. Focusing on genes present in only two copies (8910 gene pairs derived from the latest WGD), we estimated the number of neo-functionalized and sub-functionalized genes and compared these data with selection pressures (K_a/K_s ratio) and functional annotation. Our study provides a comprehensive view of gene evolution following a relatively recent duplication event.

RESULTS

Differential expression of all duplicated genes

We were able to map a large number of the total transcripts (82–90%) to the soybean genome, with a smaller percentage (34–64%) that mapped uniquely due to short read length and the duplicated nature of the genome. The green pod library is aberrant in the low number of reads aligning uniquely; this is due to a large number of ribosomal (26S) sequences. Considering all the duplicated genes in *Glycine max* (genes with 2–6 copies), the expression data show that 34–54% of the paralog gene pairs harbor evidence of transcriptional divergence in any one of the seven tissues (Table 1). Therefore, on average, approximately 50% of the paralog pairs were transcriptionally diverged, or, conversely, 50% showed no evidence of divergence of transcriptional levels between the duplicates. Interestingly, 1210 gene pairs transcriptionally diverged in the same direction across all seven tissues. We computationally identified 720 homeologous segments around the soybean genome (Figure 1a). Maintaining the false discovery rate at 0.05, co-expression of duplicated genes within blocks (gene pairs in a given homeolog that are differentially expressed in the same direction) was observed in 3.5–7.7% of cases of homeologous regions (Table 2 and Figure S1).

Also of interest were differences in homeologous expression across tissues. As an 'all by all' tissue comparison would produce too complex a picture to analyze, we

Table 1 Percentage of differentially expressed paralogs by tissue sample for the whole gene set (2–6 copies) and for strictly duplicated genes

Tissue	Whole gene set	Strictly duplicated genes
Apical meristem	53.6	45.4
Flower	51.6	43.7
Green pods	33.5	25.2
Leaves	48.5	39.8
Nodule	50.4	40.9
Root	52.8	44.4
Root tip	48.8	38.5

used only two pairs of tissues that we considered the most similar (i.e. root versus nodule tissues and flower versus leaf tissues) to test for differential expression for each gene set. As an example, when comparing gene expression between root and nodule samples, we found 16 539 genes out of 32 250 that were differentially expressed. Both tissues had approximately the same number of over- and under-expressed genes (Figure 2). However, when placing genes into the context of homeologous blocks of duplicated genes, we found 22 blocks that were over-expressed in roots versus nodules, but none that were over-expressed in nodules versus roots. A different picture was obtained when comparing flower versus leaf tissues. In that case, 10 753 genes of 32 388 were differentially expressed. In Figure 2, more dots are present above the line at zero than are under it, suggesting an overall over-expression of genes in flower. This was further corroborated by the fact that we found four homeologous blocks that were over-expressed in flowers versus leaves and only one that was over-expressed in leaves versus flowers.

Analysis of recently duplicated genes

In order to understand the role of neo- or sub-functionalization on the short-term evolution of duplicated genes in soybean, we focused our further analyses on genes harboring only two copies in the genome (strictly duplicated genes from the recent WGD, Figure 1b). Expression data showed that a large proportion of strictly duplicated genes were differentially expressed (from 25% in green pods to 45% in apical meristems, Tables 1 and 3). When performing pairwise comparisons of tissues, we found that most of the genes were over- or under-expressed in the same direction. However, overall, only 611 pairs of strictly duplicated genes were over- or under-expressed in the same direction across all seven tissues (Figure 1c).

The K_s distribution for all duplicated genes showed two obvious peaks, centered around 0.1 and 0.6 (Figure 3). Strictly duplicated genes had a $K_s < 0.4$, and 92% of those genes had a K_a/K_s ratio <0.5. Only 29 of the 8910 gene pairs have K_a/K_s ratios > 1 (Figure 4). A logistic regression model showed that ln (K_a/K_s) values had a statistically significant effect on the odds of differential expression of duplicate gene pairs in all the tissue samples (Table 4). As the sign of the logistic regression coefficients for ln (K_a/K_s) is negative, the odds of differential expression of duplicate gene pairs decreases with increasing ln (K_a/K_s) values. β values were similar across tissues, suggesting that the ln (K_a/K_s) value has a similar effect on differential gene expression across all tissues.

Gene ontologies (GO terms) were used to classify genes according to their molecular function and the pathway in which they are involved. Overall, more than 70% of the genes were associated with the GO term 'molecular function'. We found that 96% of strictly duplicated genes maintained the same function, but 4% (351 pairs of 8910)
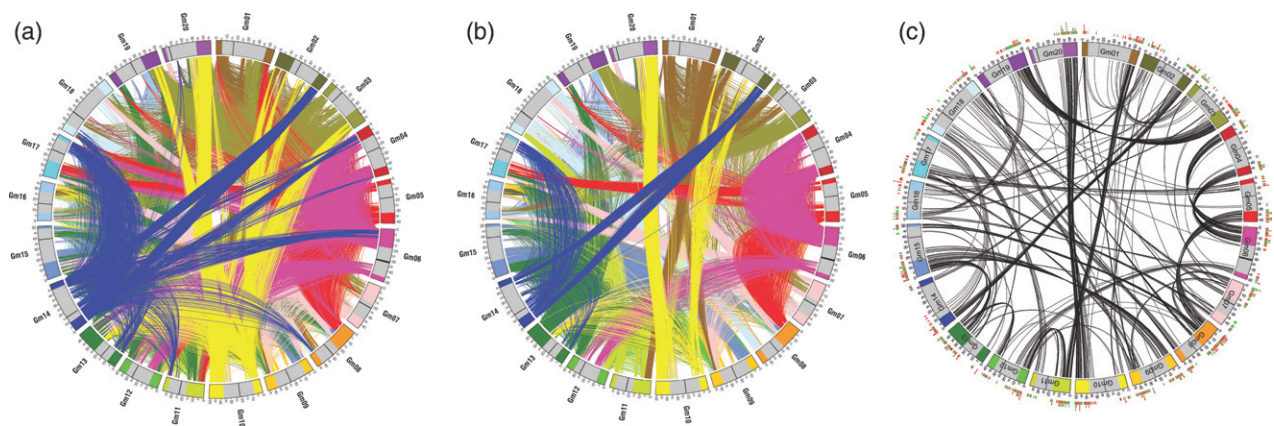

Figure 1. Homeologous relationships between the 20 soybean chromosomes. Homeologous relationships based on
(a) the 17 547 gene pairs, (b) the 8910 strictly duplicate gene pairs, and
(c) the 611 gene pairs over- or under-expressed in the same direction in the seven tissues. Red and green lines in front of genes (outside the circle) indicate whether the gene is over or under-expressed, respectively.

Table 2 Distribution of homoeologous blocks enriched for over- or under-expressed gene pairs

Tissue	Blocks enriched for differential expression	Percentage of total blocks
Apical meristem	46	6.4
Flower	33	4.6
Green pods	55	7.6
Leaves	31	4.3
Nodule	27	3.8
Root	31	4.3
Root tip	25	3.5

showed different functions between duplicates at the annotation level. K_a/K_s values for these 4% of neo-functionalized genes were statistically higher than those from genes with same function (0.36 versus 0.24, respectively, $P < 2.2e-16$). When analyzing expression data per class of annotated function, we found that the terms 'transcription factors', 'DNA binding', 'magnesium ion binding' and 'structural constituents of ribosomes' contained a significant number of differentially expressed genes compared to other functions even though they are not the most represented classes, excepting the regulation of transcription pathway (Figure 5).

DISCUSSION

Differential expression between duplicated genes is postulated to contribute to phenotypic variation (Buggs *et al.*, 2010), especially in polyploids in which copy number may increase rapidly. The prevalence of expression sub-functionalization after polyploidization (variation in relative expression of homeologs among tissues in the polyploids) has been assessed in only a few studies. In allopolyploid cotton, it was shown that, of 63 genes expressed in 24 tissues, 40% showed biased expression between homeologs (Chaudhary *et al.*, 2009). Similar patterns of biased expression have been observed in Arabidopsis (Blanc and Wolfe, 2004; Groszmann *et al.*, 2011), *Tragopogon mirus* (Buggs *et al.*, 2010), *Paramecium tetraurelia* (Arnaiz *et al.*, 2010), *Xenopus laevis* (Sémon and Wolfe, 2008) and humans (Gu *et al.*, 2002), but generally only for a limited number of genes. In addition to previous studies performed using whole-transcriptome analysis (Galbraith and Birnbaum, 2006; Higgins *et al.*, 2012), our work presents some new insights into duplicated gene expression in plants.

Soybean underwent two rounds of duplication, 59 and 13 MYA (Figure 3), and retained 17 547 genes pairs, of which 8910 are strictly duplicated genes (two copies only). We found that, on average, 50% of duplicated genes showed differential expression between duplicates, regardless of the age of the duplication (Table 1). Focusing on genes with only two copies in the genome (all duplicated 13 MYA), we

found that the vast majority have a K_a/K_s ratio <0.4 (Figure 4), indicative of a highly purifying selective pressure at the nucleotide level. In addition, transcriptional divergence and K_a/K_s ratios were negatively correlated (Table 4) in genes across the sampled tissues, suggesting that sub-functionalization across tissues has increased evolutionary pressures to maintain gene function. This is in agreement with previous results showing that, in some Arabidopsis ecotypes and in rice, in contrast to what was predicted, the Single Nucleotide Polymorphisms (SNPs) result in less radical amino acid changes in genes for which a duplicated copy exists in the genome (Chapman *et al.*, 2006). This reinforces theories of the retention of duplicates through sub-functionalization, where the partitioning of genes across tissues implies that both copies must remain functional (Force *et al.*, 1999; Lynch and Force, 2000). It also demonstrates that, if expression sub-functionalization is established rapidly after polyploidization in soybean, as in cotton (Chaudhary *et al.*, 2009), this process has been maintained over time.

Although the soybean genome has undergone diploidization/fractionation (Schlueter *et al.*, 2006; Schmutz *et al.*, 2010), there are still large regions of homology that remain from both duplication events (Figure 1). Thus, we were able to determine whether evolutionary pressures to retain gene activity, or to allow transcriptional divergence, occurred at the level of individual genes, or at larger chromosome-level regions. Overall, 720 homeologous blocks of duplication were computationally identified. A small proportion of the total homeologous blocks, between 3.5 and 7.7%, showed evidence of biased, or differential, expression across the entire block of homeologs/paralogs (Table 2). Significant clusters of co-expressed genes were also observed across tissues. As an example, although both tissues harbor approximately the same number of over- and under-expressed genes (Figure 2), 22 blocks (3% of the total homeologous blocks) were over-expressed in roots but under-expressed in nodules, and none of the blocks were over-expressed in nodules but under-expressed in roots. In contrast, more genes were over-expressed in flowers than in leaves (Figure 2). However, only four blocks were over-expressed in flowers versus leaves and only one was over-expressed in leaves versus flowers. It is possible that our analyses under-estimate the number of blocks showing evidence of coordinated gene expression, as, in some instances, the size of block may be big enough to mask the sub-structure, i.e. groups of genes within the larger block that show coordinated expression. Although the experimental design included three biological replicates, they were pooled and sequenced as a single library rather than bar-coded by replicate. This design precluded a statistical analysis that could leverage the replicated nature of the experiment. As noted by Auer and Doerge (2010), when RNA-seq data are analyzed for

Figure 2. Differential expression of duplicates across tissues.

The log fold change of the normalized gene expression is plotted on the *y* axis, and the mean log expression is plotted on the *x* axis. Blue dots represent genes that were significantly differentially expressed; gray dots represent genes that were equally expressed. The red horizontal line is at zero, providing a visual check for symmetry. Top: root versus nodule. The plot appears symmetrical, suggesting that, overall, both tissues have approximately the same number of over- and under-expressed genes. Bottom: leaf versus flower. This graph is non-symmetrical (more blue dots above zero), suggesting that, overall, genes in flower tissue are over-expressed compared to leaf tissue.
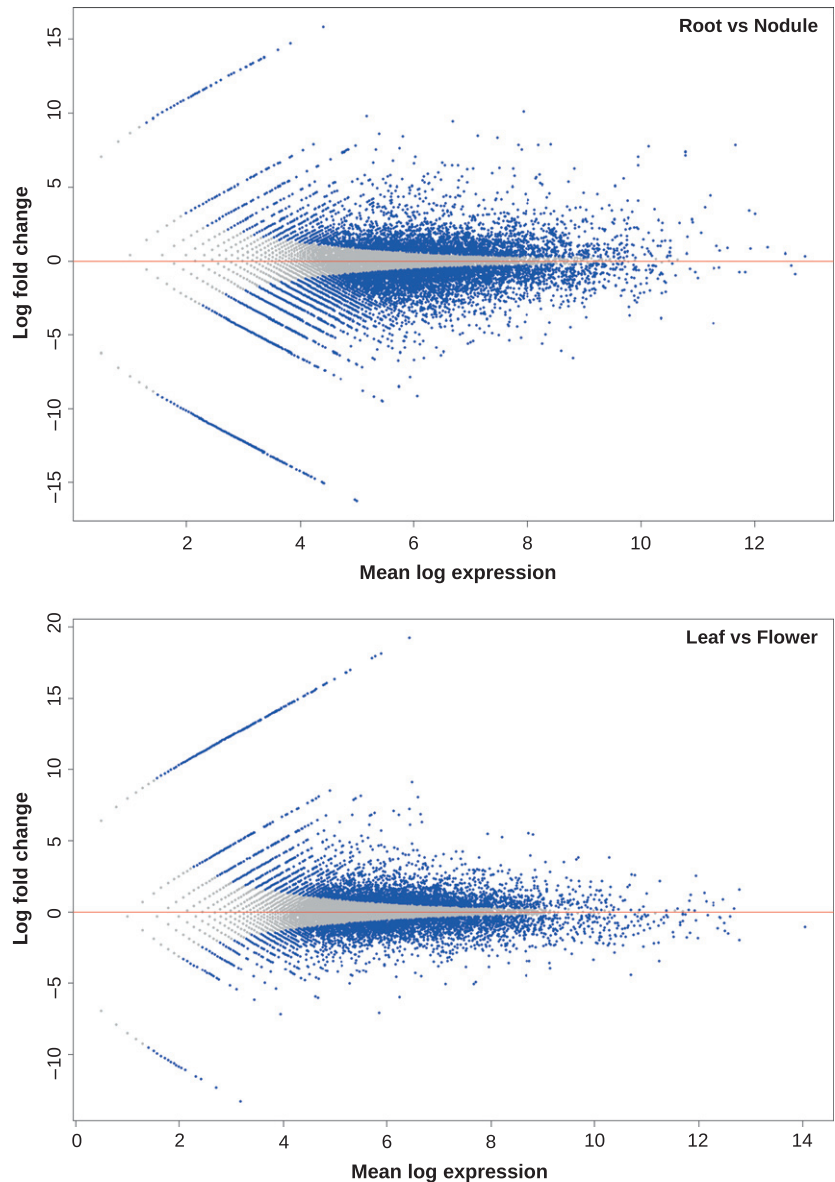
differential expression in lieu of biological replication, true false-discovery rates (FDRs) are often higher than the nominal (i.e. reported) FDRs. Accordingly, the lack of replication should be noted as a limitation of this study due to the possibility that true FDRs are inflated above the reported levels. However, our results do suggest that coordinated regulation of genes within homeologous blocks does occur in soybean, but at a low occurrence, and is variable between tissues. Previous studies documented epigenetic re-patterning after polyploidization as shown in synthetic wheat (Qi *et al.*, 2010; Zhao *et al.*, 2011), in the recent polyploid *Spartina anglica* (Salmon *et al.*, 2005) and in *Senecio* (Hegarty *et al.*, 2011). However, epigenetic mutations have been shown to be reversible within a few generations in plants (Becker *et al.*, 2011). We do not yet know to what extent epigenetic patterning effects differential gene

expression of duplicated genes in a polyploid such as *Glycine max*. However, the presence of biased differential expression across entire blocks of homeologs/paralogs provides candidate regions to investigate to role of epigenetic mutation in differential expression.

When comparing the gene expression of strictly duplicated genes between two tissues, the majority of genes were expressed in the same direction, i.e. over- or under-expressed in both tissues (Table 3). In fact, 1210 of the whole set of duplicated genes (17 547 gene pairs) and 611 gene pairs of the 8910 strictly duplicated genes showed evidence of transcriptional divergence in the same direction across all seven tissues (Figure 1c). These may be candidates for genes that are being non-functionalized, although further tissue sampling may reveal tissues in which the other copy is more highly expressed.

Table 3 Matrix of differential expression of strictly duplicated genes (8910 pairs)

	Apical meristem (4045 pairs)	Flower (3890 pairs)	Green pods (2242 pairs)	Leaves (3546 pairs)	Nodule (3640 pairs)	Root (3959 pairs)	Root tip (3434 genes)
Apical meristem		2391	1691	2207	1913	2229	1946
Flower	186		1510	2386	1834	2087	1666
Green pods	85	116		1409	1238	1506	1281
Leaves	199	137	120		1711	1987	1562
Nodule	356	370	178	340		2129	1613
Root	256	265	124	234	247		2127
Root tip	263	331	132	300	355	167	

The upper diagonal of the matrix gives the number of gene pairs differentially expressed in same manner in the compared tissues (over- or under-expressed in both tissues). The lower diagonal gives the number of gene pairs differentially expressed in the compared tissues (over-expressed in one tissue but under-expressed in the other tissue). Numbers in parentheses indicate the number of gene pairs that are differentially expressed per tissue.
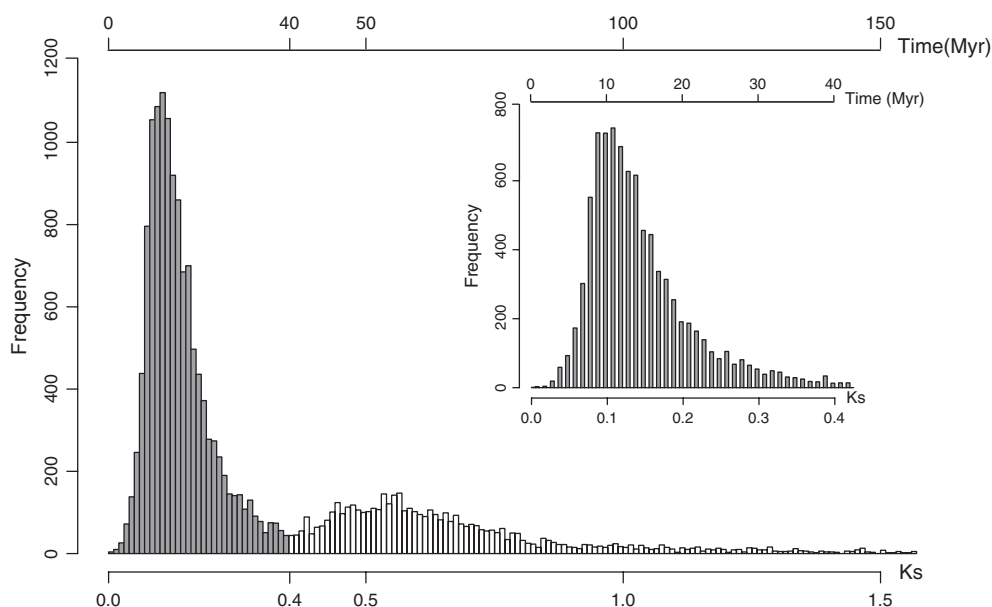
Figure 3. K_s distributions.
Histogram showing pairwise K_s values, converted into million years (My), for gene families harboring 2–6 copies in the soybean genome. K_s values between 0 and 0.39 (gray) correspond to the duplication 13 MYA; K_s values >0.4 (white) correspond to the duplication 59 MYA. The inset histogram shows pairwise K_s values calculated for strictly duplicated genes only, converted into million years (My).

Gene ontology (GO) analysis performed on the 8910 strictly duplicated gene pairs showed that nearly all gene pairs retained function (96.6%), at least at the level of GO annotation. Only 4% of the strictly duplicated genes have putatively neo-functionalized over the past 13 million years, even though their K_a/K_s values, while slightly higher than those for gene pairs that retained function (0.36 versus 0.24), were below 1. Overall, only 29 gene pairs, 0.3% of the total, had a K_a/K_s values > 1. However, these genes showed the same function for both copies and therefore constitute candidates to study the process of neo-functionalization. Neo-functionalization has been shown to be responsible for important evolutionary features. For instance, in Arabidopsis, *Toc75*, a gene whose product is involved in the import of nucleus-encoded proteins into chloroplasts, originated from a duplication in an ancient eukaryotic organism more than 1.2 billion years ago and evolved through neo-functionalization (Töpel *et al.*, 2012). Regulatory neo-functionalization (i.e. gain of a new expression pattern) has played a role in pollen evolution in Arabidopsis (Liu *et al.*, 2011). Evolution through neo-functionalization has also been demonstrated in animals. In *Drosophila melanogaster*, functional investigations showed that *CG11700*, a gene specifically expressed in males, was neo-functionalized from a polyubiquitin gene, and its product has evolved as a factor that is responsible for the trade-off between male fecundity and lifespan (Zhan *et al.*, 2012). Genes involved in sex determination in vertebrates also evolved through
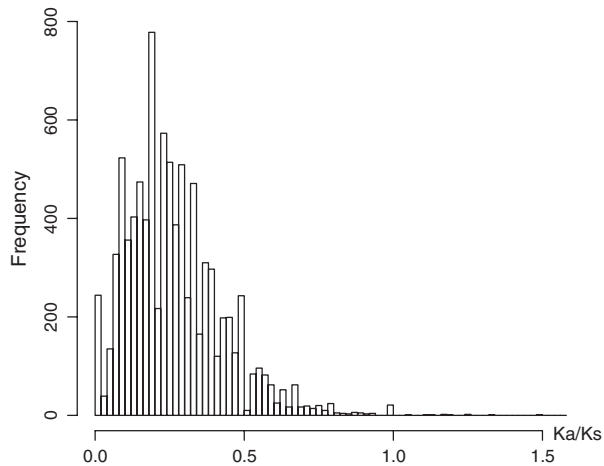
Figure 4. K_a/K_s values for the 8910 strictly duplicated genes.

Table 4 Summary of the seven model fits with ln (K_a/K_s) as the predictor variable

Tissue sample	Decrease in the odds of differential expression for unit increase in ln (K_a/K_s)	β value	P value
Apical meristem	1.036863	−0.0362	0.00087
Flower	1.032105	−0.0316	0.0045
Green pods	1.053586	−0.0522	9.20E−006
Leaves	1.035102	−0.0345	0.0021
Nodule	1.062899	−0.061	1.10E−008
Root	1.038108	−0.0374	0.00057
Root tip	1.056118	−0.0546	4.20E−007

The table shows the decrease in odds of differential expression of a duplicate gene pair for a unit increase in the ln (K_a/K_s) value, the β value, which represents the effect of ln (K_a/K_s) on the odds of differential expression of duplicate gene pairs, and the P value for the logistic regression hypothesis test for the relationship between the ln (K_a/K_s) value and the odds of differential expression. A unit increase in ln (K_a/K_s) is equivalent to an approximately 2.72 times increase in K_a/K_s value.

neo-functionalization (Mawaribuchi *et al.*, 2011), as did genes involved in venom production in marine cone snails (Chang and Duda, 2012). However, neo-functionalization remains a relative rare event when compared to the number of duplicated genes. As this process involves selection of beneficial mutations (Freeling, 2008), neo-functionalization is a slow evolutionary mechanism (Freeling, 2009). This may explain why, in the case of recent polyploids, only few case of neo-functionalization have been reported, while in ancient paleopolyploids, such as Arabidopsis, more duplicates have been shown to have acquired new functions (Blanc and Wolfe, 2004).

A large proportion of strictly duplicated genes are involved in the regulation of transcription (Figure 5) at the functional or pathway level. Interestingly, only four categories (transcription factors, DNA binding, structural constituents of ribosomes and magnesium ion binding), were significantly differentially expressed in at least one tissue (Figure 5). Following the last tetraploidization event in Arabidopsis, transcription factors and ribosomal protein genes were shown to be enriched in the genome (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere *et al.*, 2005). Ribosomal protein genes are also enriched in yeast (Papp *et al.*, 2003), while transcription factors became enriched in poplar (Rodgers-Melnick *et al.*, 2011) and after the pre-grass polyploidy in rice (Tian *et al.*, 2005). Here, we show that transcription factors and ribosomal protein genes, as in other species, were retained after the polyploidization 13 MYA, and that these genes are more likely to be differentially expressed. This is in agreement with the theory of Ohno, who first suggested that the main consequence of polyploidization is to increase the complexity of regulatory networks (Ohno, 1970).

CONCLUSION

Among the many models that attempt to explain how/why duplicated genes are retained after polyploidy (for review, see Freeling, 2009), sub-functionalization is the most popular hypothesis even though it remains controversial (Freeling, 2008). Here, we demonstrated that, after two rounds of polyploidization, a large proportion of duplicated genes exhibit differential expression, most of which show evidence of sub-functionalization at the expression level. As this is a relatively rapid process compared to the classical model of sub-functionalization through mutations (Doyle *et al.*, 2008), it may be one of the major evolutionary effects of polyploidization. This is especially true in soybean, where transcription factors and ribosomal protein genes are the two main gene classes that are differentially expressed in several tissues. Therefore, polyploidization events provide raw material to enhance network complexity and organismal adaptability to new environmental conditions. Sub-functionalization, by increasing the pressure to keep both copies of the same gene, may constitute a transitional step to neo-functionalization for some genes, as supported by the small number of putatively neo-functionalized genes in soybean.

EXPERIMENTAL PROCEDURES

Plant culture

All tissues described below were isolated from soybean *Glycine max* (L.) Merr. cultivar Williams 82. For each tissue, three independent biological replicates were performed for different set of plants to ensure reproducibility of the plant tissues analyzed (i.e. seeds were sown three times on different days, and tissues were harvested as described below). Soybean seeds were surface-sterilized and germinated for 3 days between moist Whatman filter paper (Whatman, www.whatman.com). Root tips were harvested
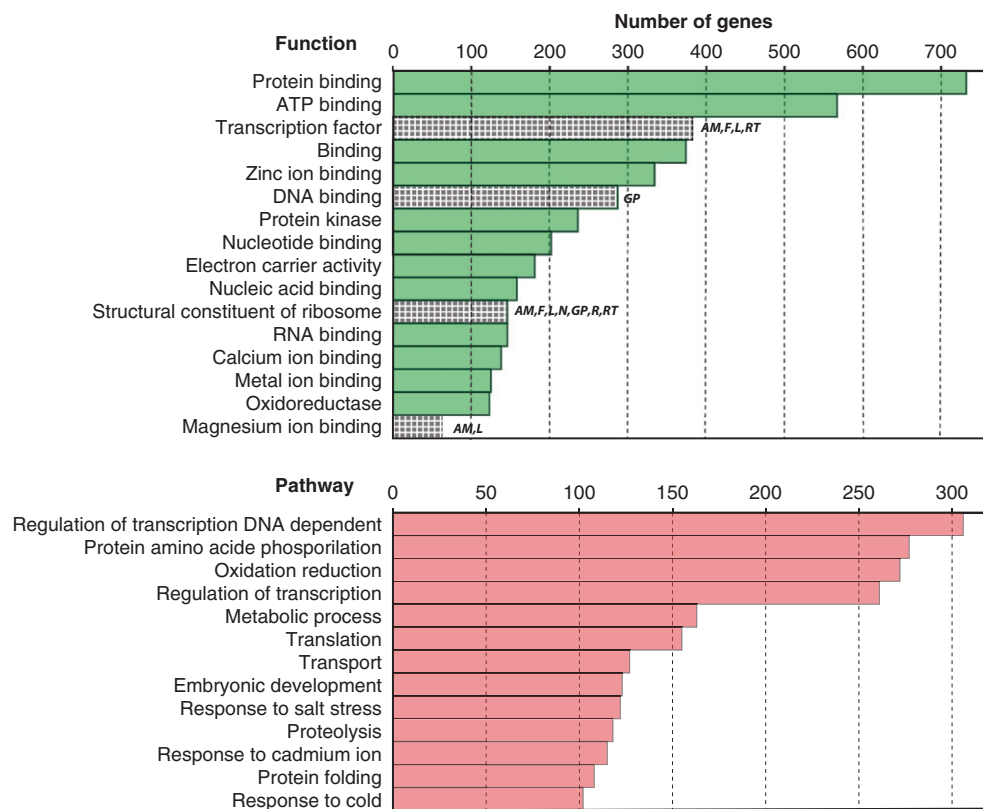
Figure 5. Gene ontology annotation.
The histogram at the top represents the number of genes associated with a function, for functions represented by more than 100 gene sequences or that contain a significant number of differentially expressed duplicate gene pairs. Dotted bars represent functions that contain a significant number of differentially expressed duplicate gene pairs in at least one of the following tissues: AM, apical meristem; F, flower; L, leaf; N, nodule; GP, green pod; R, root; RT, root tip. The histogram at the bottom represents the number of genes associated with a pathway, for pathways represented by more than 100 gene sequences. Pathway denominations are identical to the ones in the GO annotation output, thus regulation of transcription appears in two categories: DNA-dependent and DNA-independent.

from 3-day-old seedlings. To produce other tissues, germinated seedlings were transferred to the greenhouse under long-day conditions (16 h light/8 h dark) at 27°C on Promix Bx soil (Premier Horticulture, http://www.pthorticulture.com/). Fourteen-day-old shoot apical meristems (V2 stage), 18-day-old trifoliate leaves and roots (V2 stage), flowers (R2 stage) and pods (R6 stage) were harvested (http://extension.agron.iastate.edu/soybean/production_growthstages.html). Nodules were harvested 32 days after inoculation of 1 ml of *B. japonicum* suspension (OD$_{600}$ = 0.1) to 3-day-old-seedlings.

RNA extraction and sequencing

Total RNA was isolated as described by Libault *et al.* (2010). For each tissue, equal amounts of RNA isolated from the three independent biological replicates were pooled and sequenced using the Solexa platform (Illumina, Inc., http://www.illumina.com) after first- and second-strand cDNA synthesis. Between 4.18 and 6.84 million reads of approximately 36 bp were generated for each tissue.

Duplicate gene analysis for differential expression within tissue samples

Sequence filtering and alignment were processed as described by Libault *et al.* (2010). Read counts used in expression analyses were based on the subset of uniquely aligned reads that also over-

lapped the genomic spans of the Glyma1 (www.phytozome.org) gene predictions. Read counts for a given sample were normalized by using values for a gene's uniquely aligned read counts per million reads uniquely aligning within that sample. For a given gene, the measured expression, i.e. the level at which it was transcribed, was proportional to its length. For each tissue sample, differential expression was tested for each pair of duplicate genes as defined by Schmutz *et al.* (2010), using the exact conditional test (Gu *et al.*, 2008). For each pair of genes, *P* values were computed from the exact conditional test and adjusted to maintain the false discovery rate (FDR) at 0.05 across gene pairs using the Benjamini –Hochberg method (Benjamini and Hochberg, 1995). Only gene pairs whose expression differed from 0 were included in the analysis. For further information about the statistical analyses, see Methods S1.

Homeologous block analysis within tissue samples

Identification of homoeologous blocks was performed using I-ADHoRE version 2.1 (Simillion *et al.*, 2008), and pairs of duplicate genes were structured into pre-defined blocks of duplicate genes (i.e., homeologs). For each tissue, we analyzed whether the gene pair homeologs present on a given block were statistically differentially expressed in the same direction as described in Methods S1. *P* values were calculated for every defined duplicate block, and adjusted to maintain the FDR across duplicate blocks at 0.1.

Differential expression of genes and homeolog blocks across tissues

Differential expression between similar tissues, i.e. root versus nodule tissues and flower versus leaf tissues, was tested for every gene using Fisher's exact test, with a small *P* value corresponding to a statistically significant association between tissue type and expression of the gene. *P* values were subsequently adjusted to maintain the FDR at 0.05 across genes. For each pre-defined block of duplicate genes, we tested whether the genes were statistically differentially expressed in the same direction between roots and nodules as well as between flower and leaf tissues. *P* values were calculated following a similar protocol to that described in the previous section (details in Methods S1). *P* values were subsequently adjusted to maintain the FDR at 0.05 across genes.

K_a and K_s calculation and functional annotation

The number of non-synonymous (K_a) and synonymous (K_s) mutations and K_a/K_s ratio values were calculated using PAML software (Yang, 2007), based on alignment of both nucleotide and protein sequences of the gene and its duplicate. K_s values were calculated for the whole set of duplicated genes (genes harboring two copies or more, i.e. 17 547 pairs), and used to estimate time using a value of 0.056 synonymous transversions per site (Schmutz *et al.*, 2010). The K_a/K_s ratio was calculated for the 8910 pairs of strictly duplicated genes (genes harboring two copies only), and, for each copy, the function and the pathway in which the gene is involved were annotated based on similarity searches using the Blast2GO Gene Ontology (GO) tool (Conesa *et al.*, 2005).

Relationship between the K_a/K_s value and differential expression of duplicate gene pairs

A natural logarithm transformation was applied to K_a/K_s values, with a boundary correction of 1E−10 for a null K_a/K_s value. The relationship between the ln (K_a/K_s) value and the odds of differential expression of duplicate gene pairs for the seven tissue samples was modeled using the logistic regression equation:

$$\log\left(\frac{P_g}{1-P_g}\right) = \beta_0 + \beta_1 \left(\ln\left[\frac{K_a}{K_s}\right]\right)_g$$

where P_g is the probability of differential expression of duplicate gene pair 'g', β_0 is the intercept of the logistic regression model, and β_1 is the coefficient of the K_a/K_s value for duplicate gene pair 'g', i.e. the effect of ln (K_a/K_s) on the odds of differential expression of duplicate gene pair 'g'.

Relationship between functional annotation and differential expression of strictly duplicated gene pairs

For each of the 1201 GO functional classes identified and each tissue sample, the relationship between differential expression of duplicate genes and their biological function (as defined in the Blast2GO analysis) was tested using a hyper-geometric test (Fisher's exact test) (Fisher, 1966). *P* values were subsequently adjusted to maintain the FDR at 0.05. Duplicate gene pairs were assigned to a GO functional class if at least one of the two genes was involved in that function. Duplicate gene pairs were permitted to be present in multiple functional classes if the duplicate gene pair was involved in multiple functions.

Circos

Coding sequence information (start and end points as well as chromosome locations) was retrieved from the phytozome database (www.phytozome.net/soybean). In order to visualize duplicated regions in the soybean genome, lines were drawn between matching genes using Circos (Krzywinski *et al.*, 2009), first for the whole set of genes and then on the 8910 strictly duplicated gene pairs only. An additional circular layout was drawn for the strictly duplicated pairs of genes that were over/under-expressed in the same direction in the seven tissues. The results are displayed in Figure 1, with over-expressed copies in the seven tissues shown in red and under-expressed copies shown in green.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.
Figure S1. Co-expression of duplicated genes within blocks.
Methods S1. Detailed statistical analyses.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

Adams, K.L., Cronn, R., Percifield, R. and Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl Acad. Sci. USA*, **100**, 4649–4654.

Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E. and Sperling, L. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia*. *BMC Genomics*, **11**, 547.

Auer, P.L. and Doerge, R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

Baumel, A., Ainouche, M.L. and Levasseur, J.E. (2001) Molecular investigations in populations of *Spartina anglica* C.E. Hubbard (Poaceae) invading coastal Brittany (France). *Mol. Ecol.* **10**, 1689–1701.

Baumel, A., Ainouche, M., Kalendar, R. and Schulman, A.H. (2002) Retrotransposons and genomic stability in populations of the young allopolyploid species *Spartina anglica* C.E. Hubbard (Poaceae). *Mol. Biol. Evol.* **19**, 1218–1227.

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K. and Weigel, D. (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, **480**, 245–249.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300.

Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell*, **16**, 1679–1691.

Buggs, R.J., Elliott, N.M., Zhang, L., Koh, J., Viccini, L.F., Soltis, D.E. and Soltis, P.S. (2010) Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytol.* **186**, 175–183.

Chang, D. and Duda, T.F. (2012) Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Mol. Biol. Evol.* **29**, 2019–2029.

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. USA*, **103**, 2730–2735.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N., Springer, N.M. and Wendel, J.F. (2009) Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). *Genetics*, **182**, 503–517.

Conesa, A., Götz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Crane, P.R. and Lidgard, S. (1989) Angiosperm diversification and paleolatitudinal gradients in cretaceous floristic diversity. *Science*, **246**, 675–678.

Crepet, W.L. and Niklas, K.J. (2009) Darwin's second 'abominable mystery': why are there so many angiosperm species? *Am. J. Bot.* **96**, 366–381.

Cusack, B.P. and Wolfe, K.H. (2007) When gene marriages don't work out: divorce by subfunctionalization. *Trends Genet.* **23**, 270–272.

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

Doyle, J.J. and Luckow, M.A. (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiol.* **131**, 900–910.

Doyle, J.J., Doyle, J.L. and Harbison, C. (2003) Chloroplast-expressed glutamine synthetase in *Glycine* and related Leguminosae: phylogeny, gene duplication and ancient polyploidy. *Syst. Bot.* **28**, 567–577.

Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S. and Wendel, J.F. (2008) Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461.

Fawcett, J.A., Maere, S. and Van de Peer, Y. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl Acad. Sci. USA*, **106**, 5737–5742.

Feldman, M. and Levy, A.A. (2009) Genome evolution in allopolyploid wheat – a revolutionary reprogramming followed by gradual changes. *J. Genet. Genomics*, **36**, 511–518.

Fisher, R.A. (1966) *The Design of Experiments*, 8th edn. Edinburgh: Oliver and Boyd.

Flagel, L.E. and Wendel, J.F. (2010) Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol.* **186**, 184–193.

Flagel, L., Udall, J., Nettleton, D. and Wendel, J. (2008) Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biol.* **6**, 16.

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. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60**, 433–453.

Galbraith, D.W. and Birnbaum, K. (2006) Global studies of cell type-specific gene expression in plants. *Annu. Rev. Plant Biol.* **57**, 451–475.

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

Groszmann, M., Paicu, T., Alvarez, J.P., Swain, S.M. and Smyth, D.R. (2011) *SPATULA* and *ALCATRAZ* are partially redundant, functionally diverging bHLH genes required for Arabidopsis gynoecium and fruit development. *Plant J.* **68**, 816–829.

Gu, Z., Nicolae, D., Lu, H.H.–.S. and Li, W.H. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* **18**, 609–613.

Gu, K., Ng, H.K., Tang, M.L. and Schucany, W.R. (2008) Testing the ratio of two Poisson rates. *Biom. J.* **2**, 283–298.

Hegarty, M.J., Batstone, T., Barker, G.L., Edwards, K.J., Abbott, R.J. and Hiscock, S.J. (2011) Nonadditive changes to cytosine methylation as a consequence of hybridization and genome duplication in *Senecio* (Asteraceae). *Mol. Ecol.* **20**, 105–113.

Higgins, J.A., Magusin, A., Trick, M., Fraser, F. and Bancroft, I. (2012) Use of mRNA-Seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species *Brassica napus. BMC Genomics*, **13**, 247.

Hufton, A.L. and Panopoulou, G. (2009) Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev.* **19**, 600–606.

Innes, R.W., Ameline-Torregrosa, C., Ashfield, T. *et al.* (2008) Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiol.* **148**, 1740–1759.

Jaillon, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jiao, Y., Wickett, N.J., Ayyampalayam, S. *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.

Kenan-Eichler, M., Leshkowitz, D., Tal, L., Noor, E., Melamed-Bessudo, C., Feldman, M. and Levy, A.A. (2011) Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics*, **188**, 263–272.

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.

Levy, A.A. and Feldman, M. (2004) Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. *Biol. J. Linn. Soc.* **82**, 607–613.

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.

Lidgard, S. and Crane, P.R. (1988) Quantitative analyses of the early angiosperm radiation. *Nature*, **331**, 344–346.

Liu, S.–.L., Baute, G.J. and Adams, K.L. (2011) Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana. Genome Biol. Evol.* **3**, 1419–1436.

Lord, J.M. and Westoby, M. (2011) Accessory costs of seed production and the evolution of angiosperms. *Evolution*, **66**, 200–210.

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

Lyons, E., Pedersen, B., Kane, J. *et al.* (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781.

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.

Mawaribuchi, S., Yoshimoto, S., Ohashi, S., Takamatsu, N. and Ito, M. (2011) Molecular evolution of vertebrate sex-determining genes. *Chromosome Res.* **20**, 139–151.

Moore, R.C. and Purugganan, M.D. (2005) The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* **8**, 122–128.

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

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

Papp, C., Pal, C. and Hurst, L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194–197.

Qi, B., Zhong, X., Zhu, B., Zhao, N., Xu, L., Zhang, H., Yu, X. and Liu, B. (2010) Generality and characteristics of genetic and epigenetic changes in newly synthesized allotetraploid wheat lines. *J. Genet. Genomics* **37**, 737–748.

Rapp, R.A., Udall, J.A. and Wendel, J.F. (2009) Genomic expression dominance in allopolyploids. *BMC Biol.* **7**, 18.

Rodgers-Melnick, E., Mane, S.P., Dharmawardhana, P., Slavov, G.T., Crasta, O.R., Strauss, S.H., Brunner, A.M. and DiFazio, S.P. (2011) Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus. Genome Res.* **22**, 95–105.

Salmon, A., Ainouche, M.L. and Wendel, J.F. (2005) Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Mol. Ecol.* **14**, 1163–1175.

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

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

Sémon, M. and Wolfe, K.H. (2008) Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis. Proc. Natl Acad. Sci. USA*, **105**, 8333–8338.

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

Simillion, C., Janssens, K., Sterck, L. and Van de Peer, Y. (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, **24**, 127–128.

Stefanović, S., Pfeil, B.E., Palmer, J.D. and Doyle, J.J. (2009) Relationships among phaseoloid legumes based on sequences from eight chloroplast regions. *Syst. Bot.* **34**, 115–128.

Straub, S.C., Pfeil, B.E. and Doyle, J.J. (2006) Testing the polyploid past of soybean using a low-copy nuclear gene – is *Glycine* (Fabaceae: Papilionoideae) an auto- or allopolyploid? *Mol. Phylogenet. Evol.* **39**, 580–584.

Stuessy, T.F. (2004) A transitional combinational theory for the origin of angiosperms. *Taxon*, **53**, 3–16.

Tang, H., Bowers, J.E., Wang, X. and Paterson, A.H. (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl Acad. Sci. USA*, **107**, 472–477.

Tate, J.A., Joshi, P., Soltis, K.A., Soltis, P.S. and Soltis, D.E. (2009) On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biol.* **9**, 80.

Tian, C.G., Xiong, Y.Q., Liu, T.Y., Sun, S.H., Chen, L.B. and Chen, M.S. (2005) Evidence for an ancient whole-genome duplication event in rice and other cereals. *Yi Chuan Xue Bao* **32**, 519–527.

Töpel, M., Ling, Q. and Jarvis, P. (2012) Neofunctionalization within the Omp85 protein superfamily during chloroplast evolution. *Plant Signal. Behav.* **7**, 161–164.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. and Vandepoele, K. (2009) The flowering world: a tale of duplications. *Trends Plant Sci.* **14**, 680–688.

Veron, A.S., Kaufmann, K. and Bornberg-Bauer, E. (2007) Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Mol. Biol. Evol.* **24**, 670–678.

Vision, T.J. (2000) The origins of genomic duplications in Arabidopsis. *Science*, **290**, 2114–2117.

Wang, Y., Wang, X. and Paterson, A.H. (2012) Genome and gene duplications and gene expression divergence: a view from plants. *Ann. N. Y. Acad. Sci.* **1256**, 1–14.

Xiong, Z., Gaeta, R.T. and Pires, J.C. (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl Acad. Sci. USA*, **108**, 7908–7913.

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

Yu, Z., Haberer, G., Matthes, M., Rattei, T., Mayer, K.F.X., Gierl, A. and Torres-Ruiz, R.A. (2010) Impact of natural genetic variation on the transcriptome of autotetraploid *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **107**, 17809–17814.

Zahn, L.M., Kong, H., Leebens-Mack, J.H., Kim, S., Soltis, P.S., Landherr, L. L., Soltis, D.E., Depamphilis, C.W. and Ma, H. (2005) The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics*, **169**, 2209–2223.

Zhan, Z., Ding, Y., Zhao, R., Zhang, Y., Yu, H., Zhou, Q., Yang, S., Xiang, H. and Wang, W. (2012) Rapid functional divergence of a newly evolved polyubiquitin gene in Drosophila and its role in the trade-off between male fecundity and lifespan. *Mol. Biol. Evol.* **29**, 1407–1416.

Zhao, N., Zhu, B., Li, M., Wang, L., Xu, L., Zhang, H., Zheng, S., Qi, B., Han, F. and Liu, B. (2011) Extensive and heritable epigenetic remodeling and genetic stability accompany allohexaploidization of wheat. *Genetics*, **188**, 499–509.