Divergence in Enzymatic Activities in the Soybean GST Supergene Family Provides New Insight into the Evolutionary Dynamics of Whole-Genome Duplicates

Hai-Jing Liu,[†,1,2] Zhen-Xin Tang,[†,3] Xue-Min Han,[3] Zhi-Ling Yang,[3] Fu-Min Zhang,[1] Hai-Ling Yang,[3] Yan-Jing Liu,[*,1] and Qing-Yin Zeng[*1,2]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing, China

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: yanjing.liu@ibcas.ac.cn; qingyin.zeng@ibcas.ac.cn.

Associate editor: Michael Purugganan

Abstract

Whole-genome duplication (WGD), or polyploidy, is a major force in plant genome evolution. A duplicate of all genes is present in the genome immediately following a WGD event. However, the evolutionary mechanisms responsible for the loss of, or retention and subsequent functional divergence of polyploidy-derived duplicates remain largely unknown. In this study we reconstructed the evolutionary history of the glutathione S-transferase (GST) gene family from the soybean genome, and identified 72 GST duplicated gene pairs formed by a recent *Glycine*-specific WGD event occurring approximately 13 Ma. We found that 72% of duplicated GST gene pairs experienced gene losses or pseudogenization, whereas 28% of GST gene pairs have been retained in the soybean genome. The GST pseudogenes were under relaxed selective constraints, whereas functional GSTs were subject to strong purifying selection. Plant GST genes play important roles in stress tolerance and detoxification metabolism. By examining the gene expression responses to abiotic stresses and enzymatic properties of the ancestral and current proteins, we found that polyploidy-derived GST duplicates show the divergence in enzymatic activities. Through site-directed mutagenesis of ancestral proteins, this study revealed that nonsynonymous substitutions of key amino acid sites play an important role in the divergence of enzymatic functions of polyploidy-derived GST duplicates. These findings provide new insights into the evolutionary and functional dynamics of polyploidy-derived duplicate genes.

Key words: **gene and genome duplication, gene family, glutathione S-transferase, enzyme activity, functional divergence.**

Article

Introduction

Whole-genome duplication (WGD), or polyploidy, is now recognized for providing tremendous evolutionary potential and adaptive capabilities in eukaryotes (Soltis et al. 2014). WGD events are especially widespread in plants, and all angiosperms share at least two WGD events in their common evolutionary history (Jiao et al. 2011). Immediately following a WGD event, a newly formed polyploid contains a duplicate copy of each gene. Some duplicate genes subsequently become pseudogenes by accumulating deleterious mutants, whereas others persist and evolve diverse functions. Why some duplicate genes can be retained for such a long time post-WGD is a pivotal question. Neofunctionalization (acquisition of a novel function for one copy), subfunctionalization (partitioning of the functions of the ancestral gene between the two copies), relative dosage constraint (also known as the dosage balance hypothesis), and absolute dosage constraint are all plausible candidate models to explain the longer retention of some duplicates (Force et al. 1999; Birchler and Veitia 2007; Freeling 2009; Bekaert et al. 2011; Conant et al. 2014). Although previous theoretical and experimental studies have advanced our understanding of the possible retention mechanisms of polyploidy-derived duplicated genes, large comparative biochemical data sets are needed to reconstruct the evolutionary history that resulted in the functional diversification of the retained duplicate genes. Evolutionary mechanisms responsible for the retention and functional divergence of duplicate genes formed by WGD remain largely unknown.

The soybean (*Glycine max*) is an attractive system for studying the above questions because the soybean genome has undergone two WGD events, occurring approximately 59 and 13 Ma (Schlueter et al. 2004; Schmutz et al. 2010; Vanneste et al. 2014). The recent WGD (13 Ma) was probably an allotetraploidy event, as proposed by analysis of centromere satellite repeats (Gill et al. 2009). Based on the chromosome numbers, phylogenetic analysis of gene families in legumes, and comparative genomics analysis, the recent WGD event was found only in the genus *Glycine* (Egan and Doyle 2010; Schmutz et al. 2010; Cannon et al. 2015), and has been designated the *Glycine*-specific WGD (Schmutz et al. 2010). The soybean genome contains 46,430 high-confidence

protein-coding genes, of which 31,264 (15,632 gene pairs) exist as "recent" paralogs, and 15,166 have reverted to singletons (Schmutz et al. 2010). In addition, RNA-seq data showed approximately 50% of paralogs were differentially expressed (Roulin et al. 2013). Abundant polyploidy-derived duplicated gene pairs in the soybean genome make it ideal for studying the evolutionary and functional dynamics of duplicate genes following polyploidy.

Glutathione S-transferases (GSTs, EC 2.5.1.18) are multifunctional proteins encoded by a highly divergent ancient gene family. As a phase II detoxification enzyme, GSTs are involved in the detoxification of xenobiotic and endobiotic compounds by conjugating glutathione (GSH) to various hydrophobic and electrophilic substrates (Frova 2003). Plant GSTs form a large gene family of over 55 members in the *Arabidopsis*, poplar (*Populus trichocarpa*), and rice (*Oryza sativa*) genomes (Lan et al. 2009; Dixon and Edwards 2010; Jain et al. 2010). Based on amino acid identity, gene structure, and substrate specificity, plant GSTs have been divided into eight classes: Tau, phi, lambda, theta, zeta, dehydroascorbate reductase (DHAR), elongation factor 1 gamma (EF1Bγ), and tetrachlorohydroquinone dehalogenase (TCHQD) (Lan et al. 2009). We recently identified two new GST classes (hemerythrin and iota) in nonvascular plants (Liu et al. 2013). Tau, phi, lambda, DHAR, hemerythrin, and iota class GSTs are plant-specific (Edwards and Dixon 2005; Liu et al. 2013). Tau and phi GSTs are the most abundant in vascular plants, and have broad substrate specificities (Dixon et al. 2009; Lan et al. 2009; Yang et al. 2014). Because they are enzymatic proteins, comprehensive studies that combine genomic structure, gene expression, and enzymatic analyses of GSTs can elucidate the functional mechanisms responsible for the retention and functional divergence of duplicate genes.

In this study, to examine the evolutionary fates of polyploidy-derived duplicate genes at the genome level for gene expression and protein function, we conducted genome-wide annotation of the GST supergene family in the soybean genome, and reconstructed the evolutionary history of this large gene family. Seventy-two GST duplicate gene pairs created by a recent *Glycine*-specific WGD event were identified. Functional divergences of these duplicate genes were characterized by examining the gene expression responses to abiotic stresses and enzymatic properties of the ancestral, current, and mutant proteins. This study provides new insights into the evolutionary and functional dynamics of duplicate genes formed by WGD.

Results

The GST Gene Family in the Soybean

One hundred and one gene loci encoding putative GST proteins were identified in the *Glycine max* var. Williams 82 genome (supplementary table S1, Supplementary Material online). Based on the presence of frame shifts disrupting the coding region or stop codons occurring prematurely resulting in a truncated protein, 24 of the 101 putative GST genes were considered putative pseudogenes. After revising the frame shifts by deleting one or two nucleotides or removing the stop codons, these sequences were included in the phylogenetic and gene expression analyses. Domain analysis using the National Center for Biotechnology Information (NCBI) conserved domain search indicated that all predicted proteins encoded by the 101 genes contain typical GST N- and C-terminal domains, suggesting that all 101 genes are members of the GST family. The predicted proteins encoded by these 101 genes were initially classified based on the NCBI conserved domain search. These 101 full-length soybean GSTs were divided into eight classes. The tau and phi class GSTs were the most numerous, represented by 63 and 14 members, respectively. The zeta, theta, and TCHQD class GSTs were each represented by three members, both the DHAR and EF1Bγ classes by four members, and the lambda class by seven members.

Conserved gene structures were found within each class among the 101 full-length soybean GSTs (fig. 1*C*). With the exception of *GSTU54*, all 62 tau GST genes contained a single intron, whereas all 14 phi GST genes had a two-intron/three-exon structure. The zeta and lambda class GSTs contained nine introns, and the DHAR and theta GSTs contained five and six introns, respectively. Each of the soybean TCHQD genes contained only one intron. The EF1Bγ GSTs included both a GST domain and an EF1Bγ domain, with five introns observed in the GST domain. The class-specific gene structures further support the subfamily designations among the 101 full-length soybean GSTs.

In addition to full-length GST genes, 65 fragments containing partial GST domains were identified from the *Glycine max* var. Williams 82 genome (supplementary table S2, Supplementary Material online). The length of these GST fragments ranged from 21 amino acids in FR-F1 and FR-F3 to 205 amino acids in FR-L4. These GST fragments did not produce full-length functional GST proteins. Thus, in this study, these 65 GST fragments were considered as putative pseudogenes.

Duplicate Gene Pairs Formed by a Recent *Glycine*-Specific WGD Event

The distributions of 101 full-length GST genes and 65 GST fragments on soybean chromosomes were examined for this study. The GST genes are unevenly localized on the 20 soybean chromosomes (fig. 2). Twenty-nine GST gene clusters containing 76 full-length GST genes and 32 GST fragments were observed on 16 soybean chromosomes. The paralogous segments created by the recent *Glycine*-specific WGD are shown in figure 2. Except for nine GST fragments (FR-U5, FR-U16, FR-U19, FR-F2, FR-Z5, FR-Z6, FR-T1, FR-T2, and FR-DHAR2), all 101 full-length GSTs and 56 GST fragments were localized in these duplicate blocks.

We performed a comprehensive analysis to identify duplicate gene pairs formed by a *Glycine*-specific WGD event. First, GST gene pairs each located in a pair of paralogous blocks formed by *Glycine*-specific WGD were considered as candidate duplicate gene pairs (fig. 2 and supplementary fig. S1, Supplementary Material online). Second, phylogenetic analysis showed that the GST genes or fragments in candidate

Fig. 1. Phylogenetic relationships among the soybean GSTs, their expression patterns, and gene structures. In (*A*) numbers on the branches indicate the bootstrap percentage values calculated from 100 replicates, and only values greater than 50% are shown. GST genes designated as GSTU, F, T, Z, and L correspond to tau, phi, theta, zeta, and lambda class GSTs, respectively. GST genes belonging to different classes are indicated with different colors. Putative pseudogenes are indicated with asterisks. In (*B*), the blue box indicates positive detection of gene expression in the corresponding tissue under normal growth conditions (N) and following H_2O_2 (H), atrazine (A), CDNB (C), and salicylic acid (S) treatments. Symbols in the VE, V3, and R3 growth stages correspond to tissues shown in supplementary figure S4, Supplementary Material online. In (*C*), the GST N-terminal domain, C-terminal domain, and EF1Bγ domain are highlighted by the green, purple, and gray boxes, respectively, whereas introns are indicated as lines.

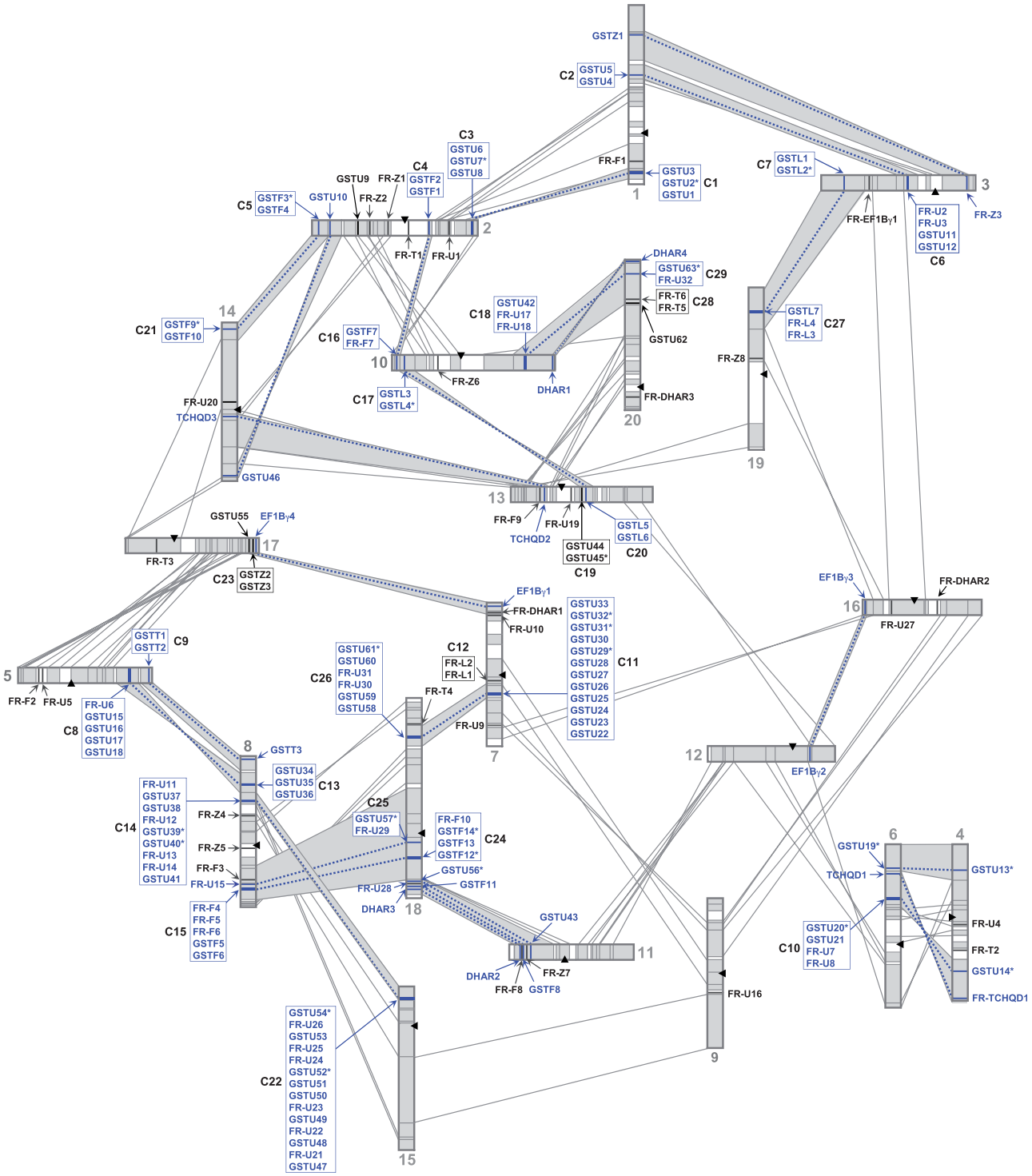


Fig. 2. Genomic localizations of soybean GST genes. Regions that are assumed to correspond to homologous genome blocks are shaded in gray and connected by lines. Paralogous GST genes and clusters are indicated by blue dashed lines within the gray-shaded trapezoids. FR indicates a GST fragment. GST genes designated as GSTU, F, T, Z, and L correspond to tau, phi, theta, zeta, and lambda class GSTs, respectively. Putative full-length pseudogenes are indicated with asterisks. The GST clusters are numbered with C1, C2, etc. The positions of centromeres are indicated by black triangles (McClean et al. 2010).

duplicate gene pairs were grouped together (fig. 3). Third, collinearity analysis showed that the regions flanking the candidate duplicated gene pair contained at least ten paralogous gene pairs formed by *Glycine*-specific WGD (supplementary fig. S2, Supplementary Material online). Finally, for each candidate duplicated gene pair located in GST clusters, the most parsimonious scenario for gene duplication, loss, and rearrangement was reconstructed based on the gene tree and the positions of genes within clusters (fig. 3). The resulting 72 GST gene pairs formed by the recent *Glycine*-specific WGD were identified in this study (table 1). Among the 72 gene pairs, tau and phi class GST pairs were the most numerous, represented

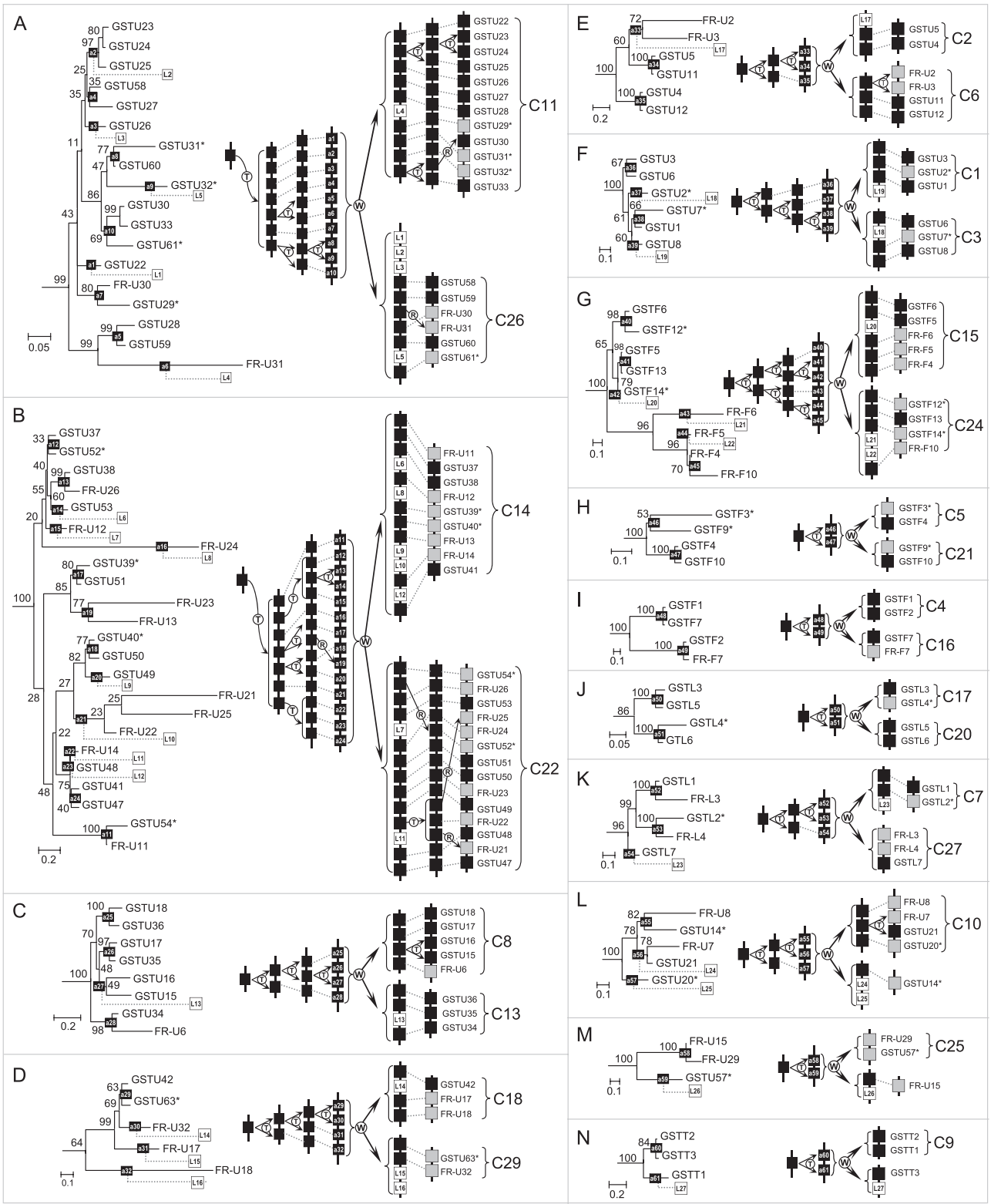


Fig. 3. Phylogenetic trees and hypothetical evolutionary histories of the soybean GST clusters. (*A*) to (*N*) correspond to different cluster pairs. Numbers on the branches indicate the bootstrap percentage values calculated from 100 replicates. The letters T, W, and R in the schematic diagram showing the hypothetical origins of GST genes indicate putative tandem duplication, WGD, and rearrangements, respectively. Putative gene loss events are indicated with dashed lines in phylogenetic trees. Gray, white, and black boxes represent pseudogenes, lost genes, and functional GST genes, respectively. Full-length pseudogenes are indicated by asterisks. FR indicates a GST fragment. GST genes designated as GSTU, F, T, and L correspond to tau, phi, theta, and lambda class GSTs, respectively. The GST clusters numbered with C1, C2, etc. are shown in figure 2. Ancestral GSTs that existed in the soybean genome before the *Glycine*-specific WGD are designated as a1, a2, etc.

Table 1. Paralogous Gene Pairs Formed by *Glycine*-Specific WGD Event.

Ancestral Copy	Fate[a]	Gene1	Gene2	d_S	d_N	d_N/d_S	Gene Expression[b]	Substrate Specificities[c]
a1	RL	*GSTU22*	L1					
a2	RL	*GSTU23*	L2					
		GSTU24						
		GSTU25						
a3	RL	*GSTU26*	L3					
a4	RR	*GSTU27*	*GSTU58*	0.319	0.065	0.204	AA	SS
a5	RR	*GSTU28*	*GSTU59*	0.166	0.051	0.307	AS	SS
a6	LL	FR-U31	L4					
a7	LL	*GSTU29**	FR-U30					
a8	RL	*GSTU60*	*GSTU31**	0.179	0.098	0.547	AA	
a9	LL	*GSTU32**	L5					
a10	RL	*GSTU30*	*GSTU61**				AS	
		GSTU33						
a11	LL	*GSTU54**	FR-U11					
a12	RL	*GSTU37*	*GSTU52**	0.347	0.054	0.156	AS	
a13	RL	*GSTU38*	FR-U26					
a14	RL	*GSTU53*	L6					
a15	LL	FR-U12	L7					
a16	LL	FR-U24	L8					
a17	RL	*GSTU51*	*GSTU39**	0.192	0.056	0.292	AS	
a18	RL	*GSTU50*	*GSTU40**	0.274	0.060	0.219	AA	
a19	LL	FR-U13	FR-U23					
a20	RL	*GSTU49*	L9					
a21	LL	FR-U21	L10					
		FR-U22						
		FR-U25						
a22	LL	FR-U14	L11					
a23	RL	*GSTU48*	L12					
a24	RR	*GSTU41*	*GSTU47*	0.196	0.056	0.286	AA	SS
a25	RR	*GSTU36*	*GSTU18*	0.203	0.066	0.325	AS	PS
a26	RR	*GSTU17*	*GSTU35*	0.198	0.042	0.212	AA	SS
a27	RL	*GSTU15*	L13					
		GSTU16						
a28	RL	*GSTU34*	FR-U6					
a29	RL	*GSTU42*	*GSTU63**	0.090	0.053	0.589	AS	
a30	LL	FR-U32	L14					
a31	LL	FR-U17	L15					
a32	LL	FR-U18	L16					
a33	LL	FR-U2	L17					
		FR-U3						
a34	RR	*GSTU5*	*GSTU11*	0.100	0.021	0.210	SS	SS
a35	RR	*GSTU4*	*GSTU12*	0.192	0.018	0.094	AA	PS
a36	RR	*GSTU3*	*GSTU6*	0.181	0.085	0.470	AA	SS
a37	LL	*GSTU2**	L18					
a38	RL	*GSTU1*	*GSTU7**	0.299	0.154	0.515	AS	
a39	RL	*GSTU8*	L19					
a40	RL	*GSTF6*	*GSTF12**	0.183	0.107	0.585	SN	
a41	RR	*GSTF5*	*GSTF13*	0.191	0.012	0.063	AA	PS
a42	LL	*GSTF14**	L20					
a43	LL	FR-F6	L21					
a44	LL	FR-F5	L22					
a45	LL	FR-F4	FR-F10					
a46	LL	*GSTF3**	*GSTF9**	0.379	0.179	0.472	SN	
a47	RR	*GSTF4*	*GSTF10*	0.143	0.016	0.112	AA	PS

(continued)

Table 1. Continued

Ancestral Copy	Fate[a]	Gene1	Gene2	d_S	d_N	d_N/d_S	Gene Expression[b]	Substrate Specificities[c]
a48	RR	*GSTF1*	*GSTF7*	0.165	0.028	0.170	SS	
a49	RL	*GSTF2*	FR-F7					
a50	RR	*GSTL3*	*GSTL5*	0.097	0.023	0.237	AA	
a51	RL	*GSTL6*	*GSTL4**	0.083	0.045	0.542	AN	
a52	RL	*GSTL1*	FR-L3					
a53	LL	*GSTL2**	FR-L4					
a54	RL	*GSTL7*	L23					
a55	LL	*GSTU14**	FR-U8					
a56	RL	*GSTU21*	L24					
		FR-U7						
a57	LL	*GSTU20**	L25					
a58	LL	FR-U15	FR-U29					
a59	LL	*GSTU57**	L26					
a60	RR	*GSTT2*	*GSTT3*	0.109	0.020	0.183	AA	
a61	RL	*GSTT1*	L27					
a62	RL	*GSTZ1*	FR-Z3					
a63	LL	*GSTU19**	*GSTU13**	0.158	0.112	0.709	AS	
a64	RL	*TCHQD1*	FR-TCHQD1					
a65	RR	*GSTU10*	*GSTU46*	0.158	0.006	0.038	SS	SS
a66	RR	*EF1BY1*	*EF1BY4*	0.080	0.019	0.238	AA	
a67	RR	*DHAR1*	*DHAR4*	0.127	0.021	0.165	AA	SS
a68	RR	*DHAR2*	*DHAR3*	0.066	0.022	0.333	AA	SS
a69	RR	*GSTF11*	*GSTF8*	0.163	0.026	0.160	AA	PS
a70	RL	*GSTU43*	*GSTU56**	0.194	0.054	0.278	AN	
a71	RR	*EF1BY2*	*EF1BY3*	0.183	0.019	0.104	AA	
a72	RR	*TCHQD2*	*TCHQD3*	0.149	0.033	0.221	AS	

NOTE.—Synonymous (d_S) and nonsynonymous substitution (d_N) rates are presented for each pair. Full-length pseudogenes are indicated with asterisks. FR indicates a GST fragment. Predicted lost genes were designated as L1, L2, etc.

[a]The fates of duplicate gene pairs were categorized as follows: RR, both duplicate genes were retained; RL, one duplicate was retained, whereas the other became a pseudogene or was lost; LL, two duplicates were degenerated into pseudogenes.

[b]Observed gene expression patterns were categorized into five classes: AA, both duplicates were expressed in all tissues under all growth conditions; AN, one duplicate was expressed, whereas the other was not detected in any tissues; AS, one duplicate was expressed in all tissues under all growth conditions, whereas the other was selectively expressed in response to a specific treatment and/or in a specific tissue; SS, both duplicates were selectively expressed in response to a specific treatment and/or in a specific tissue; SN, one duplicate showed a selective expression pattern, whereas the other was not detectable in any tissues examined.

[c]The encoded enzyme activity patterns were categorized as follows: SS, both duplicates showed a similar substrate spectrum; PS, the two duplicates showed a partially overlapping substrate spectrum.

by 47 and 11 pairs, respectively. Lambda class GSTs contained five gene pairs, and each of the DHAR, EF1Bγ, TCHQD and theta class GSTs contained two gene pairs. In contrast, the zeta class GSTs had one duplicate gene pair.

Evolutionary fates of these 72 polyploidy-derived duplicate pairs were sorted into three patterns: 1) Both duplicate genes were retained (RR model); 2) one duplicate was retained, whereas the other became a pseudogene or was lost (RL model); and 3) two duplicates degenerated into pseudogenes (LL model). Among the 72 GST duplicate gene pairs, 20, 27 and 25 pairs belonged to RR, RL and LL models (table 1), respectively. Thus, 28% of the GST gene pairs were retained in the soybean genome (RR model), whereas 72% of the gene pairs experienced gene losses or pseudogenization (RL and LL models).

Differentiation of Selective Pressure between Functional GSTs and Pseudogenes

To investigate the differentiation of selective pressure between functional GSTs and pseudogenes, we identified two types of duplicate gene pairs formed by the *Glycine*-specific WGD. One type was designated the FF gene pair, in which two duplicates were functional. In this study, the duplicate genes that had expression patterns in soybean tissues and did not contain a premature stop codon or frameshift mutations in the coding regions were defined as functional duplicates. Twenty duplicate gene pairs (also belonging to the RR model in table 1) were of this type. Another type was designated the FP gene pair. In this type, the ancestral copy of two duplicates was functional, but only one duplicate of its two descendant genes remains functional, whereas the other one becomes a full-length GST pseudogene. Ten duplicate gene pairs (a8, a10, a12, a17, a18, a29, a38, a40, a51 and a70, shown in table 1) were of this type. The synonymous substitutions (d_S values) between the FP and FF gene pairs did not show a significant difference (Kolmogorov–Smirnov test $P = 0.236$, fig. 4). However, the d_N/d_S values of the FP and FF gene pairs were significantly different (Kolmogorov–Smirnov test $P = 0.035$, fig. 4). The average d_N/d_S values of the FP and FF gene pairs were 0.40 and 0.21, respectively. Higher d_N/d_S values in the FP gene pairs were due to a greater accumulation of

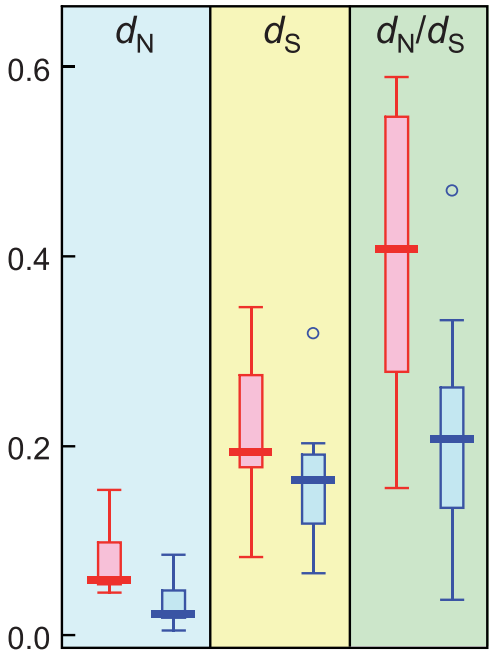


FIG. 4. Selective pressures of polyploidy-derived GST duplicated gene pairs. Red and blue boxes represent FP gene pairs and FF gene pairs, respectively. FP gene pair, the ancestral gene of two duplicates was functional, but only one duplicate of its two descendant genes remains functional, whereas the other one became a full-length GST pseudogene. FF gene pair, two duplicates were functional.

nonsynonymous substitutions (higher d_N values, fig. 4). These data indicated that pseudogenes were under more relaxed selective constraints than functional GSTs.

Analysis of the soybean full-length GST pseudogenes found that these pseudogenes were formed by premature stop codons or/and frameshift mutations in the protein-coding regions. As the majority of full-length GST pseudogenes shows expression patterns in the soybean, the coding sequence before the first premature stop codon or the first frameshift sites can hypothetically translate into a partial GST domain, whereas the regions beyond it cannot. This raises an interesting question as to whether the difference in selective pressure exists between the regions before and after the first premature stop codon or the first frameshift sites. To assess this, we calculated the d_N/d_S values of the two parts of the protein-coding regions. The d_N/d_S values of the two parts of protein-coding regions did not show significant difference (paired-samples t-test, $P > 0.35$, supplementary fig. S3, Supplementary Material online). The average d_N/d_S value of the two parts of the protein-coding regions was 0.46. These data indicated that all coding regions of the pseudogenes have undergone similar relaxed selective constraints.

Expression Patterns of GST Genes in the Soybean

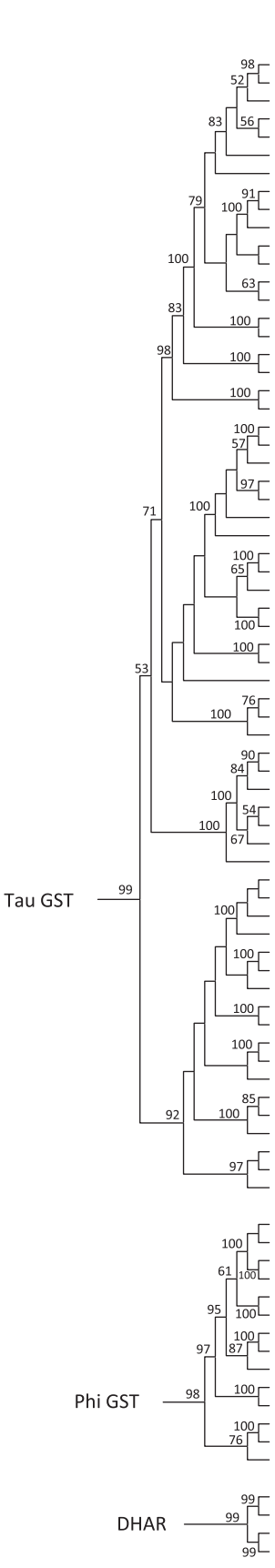
The expression patterns of all 101 full-length GSTs were examined by polymerase chain reaction (PCR) under normal growth conditions among different tissues throughout the soybean (*Glycine max* var. Williams 82) growth stages (V5, VE, V3, and R3 stages, supplementary fig. S4, Supplementary Material online). To investigate the response of soybean GSTs to stresses, we examined the expression patterns of all 101 full-length GSTs under four stress treatments (H_2O_2, atrazine, 1-chloro-2,4-dinitrobenzene, and salicylic acid applications) for the five different tissues of the V5 growth stage (supplementary fig. S4, Supplementary Material online).

Of the 101 full-length soybean GST genes, 66 were expressed in all tissues examined under all growth conditions (fig. 1*B*). Only seven putative pseudogenes (*GSTU32*, *57*, *56*, *54*, *GSTF9*, *12*, and *GSTL4*) were not expressed in any tissue or in response to any treatment applied in this study. These seven genes might be expressed at subdetectable levels, or they might be only induced in response to treatments and/or in tissues not examined in this study. The remaining 28 genes were selectively expressed in response to a specific treatment and/or in a specific tissue. Variation in expression patterns was found among tau, phi, lambda, and TCHQD class GSTs, whereas all of the DHAR, EF1Bγ, theta, and zeta GSTs were expressed in all tissues examined under all growth conditions (fig. 1*B*).

Although this study identified 72 polyploidy-derived GST duplicate pairs, only 32 contained full-length GSTs in each duplicate. Among the 32 full-length GST duplicate pairs, 16 showed the same tissue-specific expression pattern between the duplicates because both of the duplicate genes were expressed in all tissues examined (AA model in table 1). On the contrary, the rest of the 16 duplicate pairs showed expression divergence between the duplicates. Four divergent patterns of gene expression were observed in these 16 duplicate pairs. The first pattern contained two gene pairs, and one copy of each duplicate pair was expressed in all tissues examined, whereas the other was not detected in any tissue type under any of the growth conditions (AN model in table 1). In the second pattern, found in nine gene pairs, one duplicate copy was expressed in all tissues under all growth conditions, whereas the other was selectively expressed in response to a specific treatment and/or in a specific tissue (AS model in table 1). The third pattern contained three gene pairs, and both duplicates of each pair were selectively expressed in response to a specific treatment and/or in a specific tissue (SS model in table 1). In the last pattern, observed in only two gene pairs, one copy showed a selective expression pattern and the other was not detectable in all tissues examined in this study (SN model in table 1).

Substrate Specificity of the Soybean GSTs

To investigate the catalytic characteristics of soybean GST proteins, which may be related to their biological functions, we selected tau, phi, and DHAR class GSTs for protein expression and purification. Except for 22 pseudogenes, a total of 59 soybean GSTs, including 45 tau, 10 phi and four DHAR GSTs, were subcloned into *Escherichia coli* for protein expression. All of these GSTs were expressed as soluble proteins in *E. coli*, except for three tau GSTs (GSTU16, 44, and 49), which were expressed as inclusion bodies. In addition, three purified GST proteins (GSTU15, GSTF1, and GSTF6) were not stable and easily precipitated in the assay buffer. Thus, 53 purified GST proteins, including 41 tau, 8 phi, and 4 DHAR GSTs, were examined for activity assays in this study (fig. 5).



GST (analysis)	Specific activity (μmol/min per mg) to each substrate						
	CDNB	NBD-Cl	DCNB	NBC	Fluorodifen	Cum-OOH	DHA
GSTU30 (CA)	5.07 ± 0.20	6.94 ± 0.96	0.04 ± 0.01	0.26 ± 0.02	0.55 ± 0.14	0.08 ± 0.01	n.d.
GSTU33 (CA)	7.14 ± 0.20	6.99 ± 0.07	0.06 ± 0.01	0.28 ± 0.03	n.d.	0.04 ± 0.01	n.d.
GSTU61*(C-)							
GSTU31*(C-)							
GSTU60 (CA)	9.70 ± 0.17	23.02 ± 2.05	0.09 ± 0.01	0.37 ± 0.01	0.32 ± 0.01	0.22 ± 0.02	n.d.
GSTU32*(--)							
GSTU22 (CA)	5.55 ± 0.17	7.77 ± 0.19	0.05 ± 0.03	1.18 ± 0.14	2.38 ± 0.02	0.19 ± 0.01	n.d.
GSTU23 (CA)	4.75 ± 0.10	8.55 ± 0.20	0.02 ± 0.01	0.49 ± 0.13	3.17 ± 0.32	0.11 ± 0.01	n.d.
GSTU24 (CA)	13.00 ± 0.37	10.74 ± 1.68	0.07 ± 0.01	0.79 ± 0.02	6.98 ± 0.23	0.11 ± 0.01	n.d.
GSTU25 (CA)	4.39 ± 0.14	4.12 ± 0.18	0.07 ± 0.01	0.37 ± 0.06	15.23 ± 0.98	0.12 ± 0.01	n.d.
GSTU27 (CA)	19.72 ± 0.23	12.61 ± 0.61	0.09 ± 0.01	0.52 ± 0.01	1.31 ± 0.23	0.42 ± 0.02	n.d.
GSTU58 (CA)	9.01 ± 0.15	5.48 ± 0.20	0.14 ± 0.01	1.39 ± 0.15	1.31 ± 0.04	0.08 ± 0.01	n.d.
GSTU26 (CA)	0.70 ± 0.03	4.34 ± 0.56	n.d.	0.18± 0.04	n.d.	0.06± 0.01	n.d.
GSTU29*(C-)							
GSTU28 (CA)	n.d.	0.15 ± 0.02	n.d.	n.d.	n.d.	3.67 ± 0.04	n.d.
GSTU59 (CA)	n.d.	0.13 ± 0.01	n.d.	n.d.	n.d.	4.91 ± 0.35	n.d.
GSTU5 (CA)	10.63 ± 0.27	8.96 ± 1.05	n.d.	0.58 ± 0.04	0.97 ± 0.01	0.03 ± 0.01	n.d.
GSTU11 (CA)	14.05 ± 0.69	12.62 ± 1.18	n.d.	0.80 ± 0.04	0.30 ± 0.02	0.04 ± 0.01	n.d.
GSTU4 (CA)	1.53 ± 0.01	1.23 ± 0.04	0.17 ± 0.02	n.d.	n.d.	0.06 ± 0.01	n.d.
GSTU12 (CA)	1.78 ± 0.05	1.54 ± 0.09	n.d.	n.d.	0.02 ± 0.01	0.05 ± 0.01	n.d.
GSTU17 (CA)	0.02 ± 0.01	0.04 ± 0.01	n.d.	0.09 ± 0.01	n.d.	n.d.	n.d.
GSTU35 (CA)	0.03 ± 0.01	0.08 ± 0.01	n.d.	0.35 ± 0.13	n.d.	n.d.	n.d.
GSTU16 (CI)							
GSTU18 (CA)	1.12 ± 0.04	6.78 ± 0.14	n.d.	1.11 ± 0.02	n.d.	0.02 ± 0.01	n.d.
GSTU36 (CA)	0.06 ± 0.01	0.11 ± 0.02	n.d.	0.19 ± 0.02	n.d.	n.d.	n.d.
GSTU15 (CS)							
GSTU34 (CA)	0.03 ± 0.01	0.25 ± 0.01	n.d.	n.d.	0.11 ± 0.08	0.05 ± 0.01	n.d.
GSTU42 (CA)	0.10 ± 0.01	0.08 ± 0.01	n.d.	n.d.	n.d.	0.05 ± 0.01	n.d.
GSTU63*(C-)							
GSTU62 (CA)	n.d.	n.d.	n.d.	0.39 ± 0.04	n.d.	0.03 ± 0.01	n.d.
GSTU44 (CI)							
GSTU45*(C-)							
GSTU13*(C-)							
GSTU19*(C-)							
GSTU9 (CA)	0.68 ± 0.02	0.35 ± 0.02	n.d.	0.07 ± 0.01	n.d.	0.24 ± 0.02	n.d.
GSTU14*(C-)							
GSTU21 (CA)	0.76 ± 0.03	0.14 ± 0.01	n.d.	n.d.	1.67 ± 0.29	0.06 ± 0.01	n.d.
GSTU20*(C-)							
GSTU3 (CA)	13.63 ± 0.43	6.75 ± 0.30	0.63 ± 0.02	0.40 ± 0.08	15.50 ± 1.39	0.31 ± 0.01	n.d.
GSTU6 (CA)	21.85 ± 0.05	22.48 ± 0.46	0.69 ± 0.01	0.28 ± 0.04	32.73 ± 0.32	0.57 ± 0.06	n.d.
GSTU8 (CA)	7.03 ± 0.19	6.48 ± 0.48	0.13 ± 0.01	0.22 ± 0.03	14.16 ± 0.26	0.24 ± 0.01	n.d.
GSTU2* (C-)							
GSTU7* (C-)							
GSTU1 (CA)	1.89 ± 0.04	1.92 ± 0.36	0.12 ± 0.01	0.21 ± 0.02	0.43 ± 0.05	1.57 ± 0.03	n.d.
GSTU57*(--)							
GSTU52*(C-)							
GSTU53 (CA)	7.15 ± 0.22	3.53 ± 0.10	0.47 ± 0.01	0.07 ± 0.02	0.04 ± 0.01	0.19 ± 0.01	n.d.
GSTU37 (CA)	2.76 ± 0.07	3.46 ± 0.08	0.21 ± 0.01	n.d.	n.d.	0.13 ± 0.01	n.d.
GSTU38 (CA)	0.19 ± 0.01	1.68 ± 0.03	n.d.	n.d.	n.d.	0.09 ± 0.01	n.d.
GSTU10 (CA)	9.69 ± 0.22	15.85 ± 0.93	0.74 ± 0.04	0.20 ± 0.01	n.d.	0.12 ± 0.01	n.d.
GSTU46 (CA)	17.53 ± 1.08	5.55 ± 0.36	0.82 ± 0.05	0.20 ± 0.02	n.d.	0.17 ± 0.02	n.d.
GSTU55 (CA)	16.08 ± 0.39	8.54 ± 0.49	0.15 ± 0.01	0.49 ± 0.03	0.27 ± 0.06	0.28 ± 0.02	n.d.
GSTU39*(C-)							
GSTU51 (CA)	0.37 ± 0.02	0.48 ± 0.01	0.05 ± 0.01	n.d.	n.d.	0.48 ± 0.01	n.d.
GSTU43 (CA)	0.04 ± 0.01	0.02 ± 0.01	n.d.	n.d.	n.d.	0.20 ± 0.01	n.d.
GSTU56*(--)							
GSTU54*(--)							
GSTU41 (CA)	5.70 ± 0.15	3.05 ± 0.14	0.65 ± 0.19	0.09 ± 0.02	0.93 ± 0.02	0.11 ± 0.01	n.d.
GSTU47 (CA)	2.79 ± 0.11	7.48 ± 0.28	0.30 ± 0.02	0.15 ± 0.02	1.69 ± 0.18	0.39 ± 0.03	n.d.
GSTU48 (CA)	3.13 ± 0.07	2.91 ± 0.15	0.41 ± 0.03	0.14 ± 0.02	2.41 ± 0.12	0.12 ± 0.01	n.d.
GSTU40*(C-)							
GSTU49 (CI)							
GSTU50 (CA)	1.26 ± 0.01	0.48 ± 0.02	n.d.	0.15 ± 0.01	0.29 ± 0.09	0.03 ± 0.01	n.d.
GSTF3* (C-)							
GSTF9* (--)							
GSTF4 (CA)	0.06 ± 0.01	n.d.	0.05 ± 0.01	0.16 ± 0.04	0.21 ± 0.01	n.d.	n.d.
GSTF10 (CA)	0.06 ± 0.01	n.d.	0.03 ± 0.01	0.14 ± 0.02	n.d.	n.d.	n.d.
GSTF6 (CS)							
GSTF12* (--)							
GSTF5 (CA)	3.97 ± 0.74	n.d.	0.07 ± 0.01	0.07 ± 0.03	n.d.	0.16 ± 0.01	n.d.
GSTF13 (CA)	4.19 ± 0.28	n.d.	0.32 ± 0.02	n.d.	0.09 ± 0.01	0.44 ± 0.02	n.d.
GSTF14* (C-)							
GSTF8 (CA)	0.03 ± 0.01	n.d.	n.d.	1.07 ± 0.38	n.d.	n.d.	n.d.
GSTF11 (CA)	0.01 ± 0.01	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
GSTF1 (CS)							
GSTF7 (CA)	0.07 ± 0.01	0.51 ± 0.03	0.04 ± 0.01	0.11 ± 0.01	0.12 ± 0.03	0.09 ± 0.01	n.d.
GSTF2 (CA)	1.42 ± 0.07	1.16 ± 0.47	n.d.	3.62 ± 0.57	0.09 ± 0.02	0.25 ± 0.01	n.d.
DHAR1 (CA)	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	29.81 ± 0.88
DHAR4 (CA)	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	27.41 ± 0.40
DHAR2 (CA)	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	16.35 ± 0.44
DHAR3 (CA)	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	17.73 ± 0.44

Fig. 5. Phylogenetic tree and enzyme activities of soybean GST proteins. Full-length pseudogenes are indicated by asterisks. Values shown are mean ± SD, as calculated from three replicates. n.d., no activity detected. Each GST name is suffixed with a key describing the associated analysis: C, successfully cloned; A, purified GST protein assayed; I, recombinant protein totally insoluble; S, purified recombinant protein instable in buffer; dash, analysis not performed.

The substrate specificities of the purified soybean GSTs were investigated using seven substrates: 1-chloro-2,4-dinitrobenzene (CDNB), 7-chloro-4-nitrobenzo-2-oxa-1,3-diazole (NBD-Cl), 1,2-dichloro-4-nitrobenzene (DCNB), 4-nitrobenzyl chloride (NBC), fluorodifen, cumene hydroperoxide (Cum-OOH), and dehydroascorbic acid (DHA). No tau and phi GSTs showed any enzymatic activity toward substrate DHA, whereas four DHAR proteins showed enzymatic activity only toward DHA. Among the 41 purified tau GSTs, 17 had enzymatic activity toward the 6 substrates, 6 toward 5 substrates, 9 toward 4 substrates, 6 toward 3 substrates, and 3 toward only 2 substrates. Of the eight purified phi GSTs, all showed activity toward CDNB, two toward NBD-Cl, five toward DCNB, six toward NBC, four toward fluorodifen, and four toward Cum-OOH.

Large variations in specific activities were observed toward different substrates among the tandem-arrayed GSTs in the clusters. For example, among nine purified GSTs in the cluster C11, six (GSTU22, 23, 24, 25, 27, and 30) showed enzymatic activity toward six substrates (CDNB, NBD-Cl, DCNB, NBC, fluorodifen, and Cum-OOH), GSTU33 showed activity toward five substrates, GSTU26 showed activity toward four substrates, and GSTU28 showed activity toward only two substrates. Although GSTU22, 23, 24, 25, 27, and 30 showed similar substrate spectra, their enzymatic activities toward each substrate varied from 3- to 28-fold. This pattern was also observed in other clusters. Thus, diversification in enzyme specificity and activity toward different substrates has apparently evolved among GSTs in the tandem arrays.

This study only found 14 polyploidy-derived duplicate pairs of which each duplicate has enzymatic activity. For these 14 gene pairs, two patterns of differentiation in enzyme specificity were observed (PS and SS patterns). The PS pattern contained five duplicate gene pairs (GSTU4/12, 18/36, and GSTF4/10, 5/13, 8/11), and the two duplicates of each pair showed a partially overlapping substrate spectrum. For the SS pattern, found in nine duplicate gene pairs (GSTU3/6, 5/11, 10/46, 17/35, 27/58, 28/59, 41/47, and DHAR1/4, 2/3), the two duplicates of each pair showed a similar substrate spectrum, but with a difference in their enzymatic activity toward each substrate.

Functional Divergence of Duplicate Gene Pairs Formed by the Recent WGD Event

To understand the evolutionary changes in the duplicate gene pairs, we reconstructed the most recent common ancestral protein of each duplicate gene pair. These ancestral proteins were expressed and purified, and the purified ancestral proteins were examined for activity assays. Among the seven substrates listed in figure 5, a majority of the soybean GST proteins showed enzymatic activity toward CDNB and NBD-Cl. Thus, we examined the enzymatic activities of the ancestral proteins using CDNB and NBD-Cl as substrates. In this study, we only found 14 polyploidy-derived duplicate pairs of which each duplicate has enzymatic activity. Among these 14 duplicate pairs, 4 pairs (GSTU28/59, GSTF4/10, DHAR1/4, and 2/3) did not show the divergence in enzymatic activities toward substrates CDNB and NBD-Cl. Thus, the rest ten duplicate pairs (GSTU3/6, 4/12, 5/11, 10/46, 17/35, 18/36, 27/58, 41/47, and GSTF5/13, 8/11) were selected to reconstruct the most recent common ancestral protein of each duplicate pair, and examine the enzymatic activities of the ancestral proteins of each duplicate pairs.

Complex patterns of divergence in enzyme activity were observed by comparing the enzymatic activities of the ancestral protein and its descendants (fig. 6). This study revealed six patterns of differentiation in enzyme activity. In the first category, two duplicates showed higher enzymatic activities toward CDNB and NBD-Cl than the ancestral protein. The duplicate gene pair GSTU3/6 fit into this category. In the second category, observed in only one duplicate pair (GSTU18/36), one duplicate protein (GSTU18) showed higher enzymatic activities toward CDNB and NBD-Cl than the ancestral protein, but the other (GSTU36) showed lower enzymatic activities toward two substrates than the ancestral protein. The third category contained one gene pair (GSTU4/12). One duplicate (GSTU12) showed higher enzymatic activities toward CDNB and NBD-Cl than the ancestral protein, but the other (GSTU4) showed similar enzymatic activity to the ancestral protein. In the fourth category, found in only two gene pairs (GSTU41/47, 27/58), one duplicate protein had higher enzymatic activity toward two substrates than the ancestral protein, whereas the other duplicate protein showed higher enzymatic activity toward one substrate, and lower enzymatic activity toward another substrate. The fifth category contained one gene pair (GSTU10/46). Two duplicate proteins showed greater enzymatic activity toward CDNB, and lower enzymatic activity toward substrate NBD-Cl compared with the ancestral protein. In the last category, which occurred in four gene pairs (GSTU5/11, 17/35, and GSTF5/13, 8/11), two duplicate proteins showed lower enzymatic activities with CDNB and NBD-Cl than the ancestral proteins.

Mutagenesis Analysis

We selected two ancestral proteins (AnGSTU3-6 and AnGSTU18-36) to investigate the roles of nonsynonymous substitutions accumulated post-WGD, to assess enzymatic functional changes. Based on three-dimensional structures of AnGSTU3-6 and AnGSTU18-36 modeled using the InsightII software package (Accelrys, Inc., San Diego, CA), six amino acid sites were selected to create mutant proteins using site-directed mutagenesis for the biochemical assays. Six amino acid sites in two ancestral proteins AnGSTU3-6 and AnGSTU18-36 were mutated to the corresponding amino acid sites present in the daughter genes (supplementary fig. S5, Supplementary Material online). Six mutant proteins were subsequently constructed to determine their enzymatic activity to substrates CDNB and NBD-Cl (fig. 6*K* and *L*).

For duplicate gene pair GSTU3/6, the ancestral protein AnGSTU3-6 showed much lower enzymatic activities toward substrates CDNB and NBD-Cl than GSTU3 and GSTU6 (fig. 6*A*). When Leu108 of AnGSTU3-6 was replaced

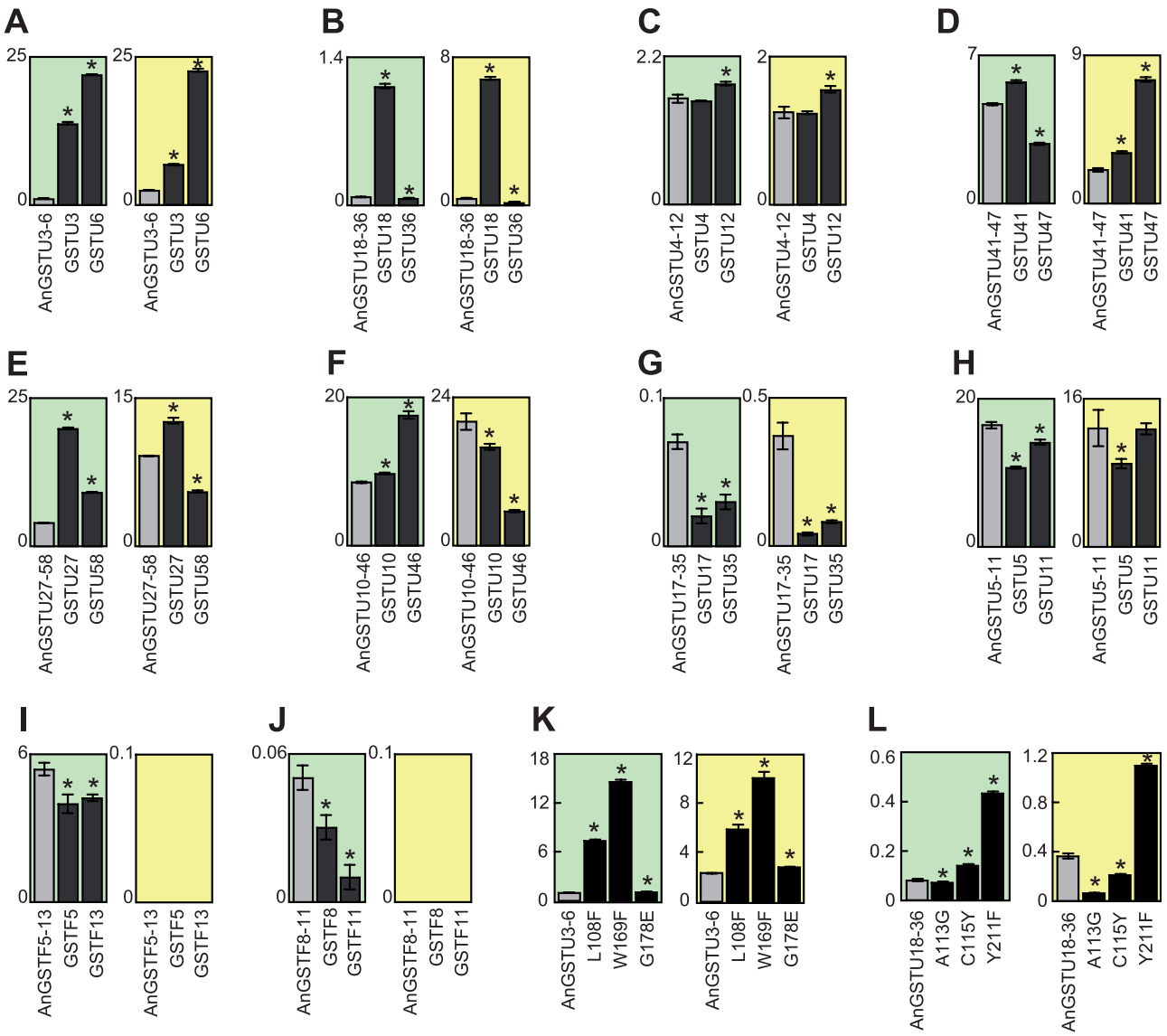


FIG. 6. Enzyme activities of the ancestral, current, and mutant soybean GST proteins. (*A*) to (*J*) correspond to different duplicate GST pairs and their ancestors. (*K*), ancestor of duplicate pair GSTU3/6 and its mutants. (*L*), ancestor of duplicate pair GSTU18/36 and its mutants. The ordinates denote the enzyme activities (μmol/min per mg). Green and yellow boxes represent enzyme activities of GST proteins toward substrates CDNB and NBD-Cl, respectively. Asterisk indicates significant difference ($P < 0.05$) in enzymatic activity between the ancestor and its descendant proteins, or between the ancestral protein and its mutants.

with Phe present in GSTU3, and Trp169 and Gly178 were replaced with the Phe and Glu present in GSTU6, respectively, the mutants L108F, W169F and G178E showed much higher enzymatic activities toward substrates CDNB and NBD-Cl than the wild-type protein AnGSTU3-6 ($P < 0.05$, Mann–Whitney *U* test, fig. 6*K*).

For the duplicate gene pair GSTU18/36, GSTU18 showed considerably higher enzymatic activity to substrates CDNB and NBD-Cl than that of ancestral protein AnGSTU18-36 (fig. 6*B*). Ala113, Cys115, and Tyr211 of AnGSTU18-36 were replaced with the Gly, Tyr, and Phe present in GSTU18, respectively. The mutant Y211F showed higher enzymatic activity to substrates CDNB and NBD-Cl than the wild-type protein AnGSTU18-36 ($P < 0.05$, Mann–Whitney *U*-test, fig. 6*L*). However, the mutants A113G and C115Y showed lower enzymatic activity toward substrate NBD-Cl than that of wild-type protein AnGSTU18-36 ($P < 0.05$, Mann–Whitney *U*-test).

Discussion

Pseudogenization of Duplicate Genes Post-WGD

WGD is an important force in plant genome evolution. Following a WGD event, all the genes that previously existed in the genome are present in duplicate. Studying the process and mechanism of polyploidy-derived duplicate gene loss/retention is particularly important for understanding the evolution of polyploidy. In this study, we identified 72 polyploidy-derived GST duplicated gene pairs in the soybean genome. Among these 72 gene pairs, 27 belonged to the RL model (one copy was retained, whereas another was changed into a pseudogene or was lost by deletion), and 25 belonged to the LL model (both copies degenerated into pseudogenes). We did

not absolutely exclude the possibility that a single loss of function event occurred prior to the *Glycine*-specific WGD event for the 25 LL model gene pairs (table 1). Regardless, however, the GST duplicate genes in the 25 gene pairs have now become pseudogenes. Thus, our data indicate that massive gene losses have occurred in this gene family. Analysis of the soybean genome suggests a rate of gene loss of 4.36% of genes per million years (approximately 56.68% duplicate genes were lost) following the *Glycine*-specific WGD (Schmutz et al. 2010). In *Raphanus raphanistrum* and *Brassica rapa*, 70% of the orthologous groups underwent gene losses post α' whole-genome triplication (Moghe et al. 2014). Comparative genomic studies of yeast species have suggested that the rate of paralogous gene loss is extremely rapid shortly after WGD events (Scannell et al. 2006). Thus, after a WGD event, duplicate gene loss may be the dominant trend over the course of subsequent genome evolution. Gene loss is considered an important mechanism in the generation of genome diversity among eukaryotic species (Li et al. 2005).

Newly formed polyploids may undergo rapid interchromosomal rearrangements and chromosomal losses following a WGD event, which can result in extensive gene loss from the genome through deletion (Tian et al. 2010; Chester et al. 2012). Pseudogenization is another model of duplicate gene loss. Following a WGD event, some duplicate genes may become pseudogenes by accumulating deleterious mutants. After a long evolutionary period, pseudogenes may be deleted from the genome or can diverge too extensively to be recognized (Zhang 2003). Nonetheless, some relatively young pseudogenes might have a high sequence similarity with the parental gene, and may be identifiable. Except for nine GST fragments, all GST genes and fragments in this study were localized in conserved syntenic blocks formed in the *Glycine*-specific WGD event (fig. 2), indicating that GST gene loss through pseudogenization was the dominant mechanism in the soybean genome. Among 72 GST duplicated gene pairs formed in the recent *Glycine*-specific WGD, some duplicates were deleted completely in the soybean genome, some duplicate genes maintained only a partial GST domain, and some duplicates contained frame shifts that disrupted the coding region or stop codons occurring prematurely, resulting in a truncated protein. These phenomena indicated that pseudogenization is still ongoing within the GST gene family in the soybean genome.

Pseudogenes are thought to evolve neutrally, free from selective constraints (Lynch and Conery 2000). By comparing d_N/d_S values between FP and FF gene pairs, this study revealed that the coding regions of GST pseudogenes are under more relaxed selective constraints than functional GSTs. We wondered whether the regulatory regions of GST pseudogenes were also free from selective constraints. In a WGD event, all of the functional elements (transcribed and regulatory) are included in the duplicated regions. In a newly formed polyploidy, the polyploidy-derived duplicate gene pairs should theoretically show identical expression patterns. However, by analyzing the expression patterns of ten FP gene pairs, we found that soybean GST pseudogenes had considerably higher degrees of expression divergence than functional GSTs after the *Glycine*-specific WGD event. A possible explanation is that the regulatory regions of some duplicated copies may be under relaxed selective constraint, which may result in higher mutation rates in the regulatory regions of the pseudogenes than in functional genes. Considered together, our results indicate that both the regulatory regions and protein-coding regions of polyploidy-derived pseudogenes might easily accumulate deleterious mutants under these relaxed selective constraints or under conditions of neutral evolution.

This study identified 14 ancestral GST clusters predating the recent *Glycine*-specific WGD event. After WGD, the GST genes in one cluster were preferentially removed, whereas the genes of its syntenic cluster were preferentially retained. For example, the ancestral cluster of clusters C11 and C26 contained ten GST genes. Following the *Glycine*-specific WGD event, cluster C11 contained nine genes, whereas its syntenic cluster C26 retained only three genes (fig. 3*A*). This preferential retention/loss of duplicate genes following WGD was also observed in *Zea mays*, *Arabidopsis thaliana*, *Tragopogon miscellus*, and *Brassica rapa* (Thomas et al. 2006; Schnable et al. 2011; Chester et al. 2012; Lou et al. 2012), and it is likely a general characteristic of posttetraploid eukaryotic genomes. This bias is thought to be the consequence of an initial inequality between the two paralogons, possibly due to epigenetic markers or gene expression differences (Semon and Wolfe 2007; Schnable et al. 2011).

Retention and Functional Divergence of Duplicate Genes Post-WGD

For soybean GST gene family, many functional GST genes were clustered in the soybean genome (fig. 2); on the other hand, 20 polyploidy-derived functional GST duplicate gene pairs have still been retained in the soybean genome (RR model in table 1). A pivotal question is why so many functional duplicates have been retained for such a long time in the soybean genome. Plant GSTs play important roles in stress tolerance and detoxification metabolism. An increased number of GST copies from small-scale duplication (e.g., tandem duplication) or WGD likely results in increased protein abundance, which may be beneficial for the defense responses of the plant. This dosage selection is likely the driving force for early retention following small-scale duplication or WGD for duplicate genes such as plant GSTs involved in abiotic or biotic stresses. Over the longer term, functional divergences (e.g., subfunctionalization or neofunctionalization) are the most likely explanations for their retention over the course of genome evolution.

Clear divergence in expression patterns was observed among the soybean GST genes. Especially, 50% of polyploidy-derived GST gene pairs had expression divergence, suggesting that these gene pairs had undergone expression subfunctionalization or neofunctionalization. For example, for nine polyploidy-derived GST pairs (AS model in table 1), one duplicate copy was expressed in all tissues under all growth conditions, whereas the other was selectively expressed in response to a specific treatment and/or in a specific tissue, suggesting that partial subfunctionalization or

neofunctionalization occurred post-WGD. Divergence in expression has been reported for various functional categories of genes (Blanc and Wolfe 2004). Expression subfunctionization might be an early stage process that reduces the chance of nonfunctionalization of duplicate genes, and thereby increases the chance of duplicate genes being retained in a genome.

Plants are continually exposed to a multitude of environmental stresses because of their sessile nature. Thus, they have evolved various plastic mechanisms for responding to a wide range of potential threats. In this study, we investigated the activities and specificities of purified soybean GST proteins that may be related to their biological functions. A broad range of substrates was examined, including CDNB, NBD-Cl, DCNB, NBC, fluorodifen, Cum-OOH, and DHA, with four of the examined substrates (CDNB, NBD-Cl, DCNB, and NBC) related to the roles of GSTs in the detoxification reaction. Fluorodifen was a model substrate used to determine the activity of GST related to herbicide detoxification (Dixon et al. 2003). Some plant GST proteins with GSH-dependent peroxidase activity were found to have major roles in counteracting oxidative injury (Cummins et al. 1999). Cum-OOH has been used extensively as model substrate for the determination of GSH-dependent peroxidase activity. DHAR class GSTs can function as thioltransferases and reduce dehydroascorbate to ascorbic acid (Tang and Yang 2013). Ascorbic acid is an antioxidant, and in association with other components of the antioxidant system, it can protect plants against oxidative damage resulting from aerobic metabolism, photosynthesis, and a range of pollutants (Smirnoff 1996). DHA has been extensively used as a model substrate for the determination of thioltransferase activity (Edwards and Dixon 2005). In this study, soybean GST proteins showed different enzymatic activities and specificities toward different substrates (fig. 5), indicating divergence in their biochemical properties. An adaptive value likely exists in possessing numerous GSTs with diverse activities and specificities to a wide range of substrates, enabling plants to respond to diverse environmental challenges.

To understand the process of enzymatic divergence, we reconstructed the most recent common ancestral proteins of the polyploidy-derived duplicate gene pairs and examined their biochemical characteristics. By comparing the enzymatic activity of the ancestral protein and its descendant, we discovered different divergent patterns for enzymatic activity, which further confirmed that the divergences of biochemical functions had occurred in soybean GST duplicates. The divergences of biochemical properties of coding proteins might contribute to the retention of GST genes in the soybean genome. Examining the enzymatic activities of mutant proteins provided further evidence that nonsynonymous substitutions of key amino acid sites in duplicate genes resulted in the divergence of enzymatic functions. For example, the ancestral protein AnGSTU3-6 showed lower enzymatic activity with substrates CDNB and NBD-Cl than the daughter protein GSTU6. When Trp169 of AnGSTU3-6 was replaced with Phe present in GSTU6, the mutants had increased enzymatic activity with these two substrates (fig. 6*K*). This indicated that the substitution of this residue in the daughter protein GSTU6 contributed to the increased enzymatic activity. In GST proteins, this Trp residue is considered a hydrophobic substrate-binding site, and the side chain of this residue forms part of the wall of the hydrophobic substrate-binding pocket (supplementary fig. S6, Supplementary Material online). Because the Trp residue had a large indole (benzopyrrole) side chain, when Trp is replaced by a residue with a smaller side chain (e.g., Phe), the hydrophobic substrate-binding pocket can enlarge, resulting in a structural change within the protein that affects its enzymatic activity.

Conclusions

WGD or polyploidy is a widespread feature of plant genomes, which facilitates evolutionary innovation and adaptation. The evolutionary mechanisms responsible for the retention and subsequent functional divergence of polyploidy-derived duplicate genes after a WGD event were poorly understood. In this study, we conducted a genome-wide annotation of the GST gene family and reconstructed the evolutionary history of this large gene family in the soybean genome. By examining the gene expression responses to abiotic stresses and the enzymatic properties of the ancestral, current, and mutant proteins, this study revealed the evolutionary and functional dynamics of GST duplicate genes formed by the recent *Glycine*-specific WGD. Our findings provide new insights into the functional fates of polyploidy-derived duplicate genes.

Materials and Methods

Identification and Nomenclature of Soybean GST Genes

To identify soybean GST genes, the *Glycine max* var. Williams 82 genome database version 1.1 (http://www.phytozome.net/, last accessed August 6, 2015) was searched with 55 full-length GST protein sequences of *Arabidopsis thaliana* (Dixon and Edwards 2010) and 81 of *Populus trichocarpa* (Lan et al. 2009) using the TBLASTN program with default algorithm parameters. *Glycine max* GST candidates were analyzed using an NCBI conserved domain search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, last accessed August 6, 2015) to confirm the presence of typical GST N- and C-terminal domains in their protein structure. The predicted GST genes were then amplified from the mRNA from *Glycine max* var. Williams 82, cloned into the *pEASY*-T3 vector (TransGen Biotech, Beijing, China), and sequenced in both directions to verify the gene sequence. For genes that PCR did not detect (7 of 101 in this study), their structures were assumed to be identical to those of their closest phylogenetic relatives. This approach was adapted from other studies (Meyers et al. 2003). A univocal name was assigned to each soybean GST gene (supplementary table S1, Supplementary Material online), consisting of a letter for the subfamily class (e.g., GSTU, F, T, Z, and L corresponding to tau, phi, theta, zeta and lambda classes, respectively) and a progressive number for each gene (e.g., GSTU1).

Phylogenetic Analysis

The full-length GST sequences were aligned using MUSCLE software (Edgar 2004). The optimal substitution model of amino acid substitution was selected using the modelGenerator version 0.84 program (Keane et al. 2006). The phylogenetic trees were constructed using the maximum-likelihood procedure with PHYML software (Guindon and Gascuel 2003). One hundred bootstrap replicates were performed in each analysis to obtain the confidence support. Cytosolic GSTs are thought to be derived from the GRX2 protein (Oakley 2005). Thus, the GRX2 protein was used as an outgroup for phylogenetic analysis of the soybean GST family.

Identification of Duplicate Gene Pairs Formed by a Recent Glycine-Specific WGD Event and Construction of the Most Recent Common Ancestors

We examined the distribution of the GST genes and fragments on the soybean chromosomes and found that some GST genes/fragments were clustered together (fig. 2). In this study, if two or more full-length GST genes or fragments were separated by no more than three intervening genes, they were collectively considered a GST gene cluster. Based on this criterion, 29 GST gene clusters were observed on 16 soybean chromosomes (fig. 2). GST fragments FR-F4 and FR-F5 had 6 and 27 gene intervals, respectively, with cluster C15 (supplementary fig. S2G, Supplementary Material online), but the two GST fragments and FR-F6 in cluster C15 were grouped together in the phylogenetic tree (fig. 3G). Thus, GST fragments FR-F4 and FR-F5 were considered as the members of cluster C15 in this study.

Previous analysis identified the paralogous segments created by the recent *Glycine*-specific WGD event provided by PLAZA v.2.5 (Proost et al. 2009; Schmutz et al. 2010). We mapped all the GST genes and fragments to the soybean genomes. To identify the duplicate gene pairs formed by a recent *Glycine*-specific WGD event, the following three criterions were used in this study: 1) Duplicated GST pairs were each located in a pair of paralogous blocks created by *Glycine*-specific WGD event (fig. 2 and supplementary fig. S1, Supplementary Material online), 2) duplicated GST pairs were grouped together in the soybean GST phylogenetic tree (figs. 1A and 3), and 3) collinearity analysis showed that the flanking regions of candidate duplicated gene pairs contained at least ten paralogous gene pairs formed by *Glycine*-specific WGD provided by PLAZA v.2.5 (supplementary fig. S2, Supplementary Material online).

The most recent common ancestor of each duplicate gene pair was reconstructed using CODEML in the Phylogenetic Analysis Using Maximum Likelihood (PAML) software package (Yang 2007). For each duplicate pair, two GST sister genes of the same clades in the phylogenetic trees of soybean GSTs were selected as outgroup for ancestor construction. The accuracy of predicted ancestors was measured by the highest posterior probability of each site, and the ancestral sequences were accepted only when every site of the sequence has a probability higher than 0.99.

Expression of Soybean GST Genes under Normal Conditions and Abiotic Stress

To investigate the expression patterns of the soybean GST genes under both normal conditions and abiotic stress, soybean plants (*Glycine max* var. Williams 82) were cultured in soil to the V5 growth stage, and then four chemical treatments were adopted: 0.5% H_2O_2, 1 mM salicylic acid, 0.1% atrazine, and 1 mM CDNB as cultivation solutions and sprays for 12, 24, 12, and 12 h, respectively. *Glycine max* of the V5 growth stage without any treatment was used as a control. Each treatment consisted of three replicates. After treatment, total RNA was isolated from root, stem, mature leaf, immature leaf, and terminal bud tissues.

Soybean plants at the VE, V3, and R3 growth stages were selected to explore the expression patterns of the soybean GST genes at different growth stages. Total RNAs were isolated from six soybean tissues at the VE growth stage. The total RNAs were isolated from 14 and 15 soybean tissues at the V3 and R3 growth stages, respectively (supplementary fig. S4, Supplementary Material online). Each experiment included three biological replicates.

The total RNAs were isolated using an Aurum Total RNA Kit (Bio-Rad Laboratories, CA), then treated with RNase-free DNase I (Promega, Madison, WI) and reverse transcribed into cDNA using a RNA PCR Kit (AMV) version 3.0 (TaKaRa, Dalian, China). One hundred and one specific primer pairs were designed based on multiple sequence alignment of all soybean GST sequences (supplementary table S3, Supplementary Material online). The soybean actin gene (Genome locus name: Glyma08g19420) was used as an internal control (primer pair: 5′-CCAAAGGCCAACAGAGAAAAG-3′ and 5′-CTTCTGGGCAACGGAATCTC-3′). PCR was performed in a volume of 25 μl containing 3 μl of first-strand cDNA, 2.5 μl of TaKaRa 10× PCR buffer, 0.125 μl of TaKaRa Ex Taq (5 units μl^{-1}), 2 μl of deoxyribonucleotide triphosphate (2.5 mM each), and 10 pmol of each primer. PCR conditions were as follows: 94 °C for 3 min, followed by cycles of 94 °C for 30 s, 60 °C annealing for 40 s, and 72 °C for 1 min, with a final extension at 72 °C for 10 min. Each PCR was performed for 35 cycles. The PCR products from each sample were analyzed by 1% agarose gel electrophoresis, and then validated by DNA sequencing.

Purification and Activity Assays of Soybean GST Proteins

To investigate the enzymatic functions of the soybean GST proteins, except for the pseudogenes all tau, phi and DHAR GSTs were selected for protein expression analysis and purification. Each full-length cDNA was subcloned into a pET-30a expression vector (Novagen) to obtain a recombinant protein with an N-terminal 6× His-tag. The primers used to construct the GST expression vectors are listed in supplementary table S4, Supplementary Material online. Colonies containing appropriate inserts were identified by sequencing.

Escherichia coli BL21 (DE3) cells harboring pET-30a/GST plasmids were cultured overnight, diluted 1:100, and grown until the optical density (A_{600}) reached 0.6. Isopropyl-β-D-

thiogalactopyranoside was added to each culture at a final concentration of 0.1 mM to induce synthesis of the recombinant GST proteins. After inducing for 10 h at 37 °C, the bacteria were harvested by centrifugation (8,000 × g, 3 min, 4 °C), resuspended in a binding buffer (20 mM sodium phosphate, 0.5 M NaCl, and 20 mM imidazole, pH 7.4), and disrupted by cold sonication. The homogenate was then subjected to centrifugation (10,000 × g, 10 min, 4 °C). The supernatant was loaded onto a Nickel-Sepharose High Performance column (GE Healthcare Bio-Sciences) that was pre-equilibrated with binding buffer. The GST proteins that bound to the Nickel-Sepharose High Performance column were eluted with elution buffer (20 mM sodium phosphate, 0.5 M NaCl, and 500 mM imidazole, pH 7.4). Protein concentrations were determined by measuring A_{280} (Layne 1957).

The activity of recombinant GST proteins toward CDNB, DCNB, and NBC was measured according to the description of Habig et al. (1974), that toward NBD-Cl was measured as described by Ricci et al. (1994), and that toward DHA, fluorodifen, and Cum-OOH was measured as described by Edwards and Dixon (2005). All assays were performed at 25 °C.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the four anonymous reviewers for valuable comments and suggestions that helped to improve the manuscript. This work was supported by the National Science Foundation of China (91231103 and 31425006) and Chinese Academy of Sciences (KSCX2-EW-J-1).

References

Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell* 23:1719–1728.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* 19:395–402.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16:1679–1691.

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN Jr, Rolf M, et al. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol Biol Evol.* 32:193–210.

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc Natl Acad Sci U S A.* 109:1176–1181.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol.* 19:91–98.

Cummins I, Cole DJ, Edwards R. 1999. A role for glutathione transferases functioning as glutathione peroxidases in resistance to multiple herbicides in black-grass. *Plant J.* 18:285–292.

Dixon DP, Edwards R. 2010. Glutathione transferases. The Arabidopsis Book. 8:e0131; doi:10.1199/tab.0131.

Dixon DP, Hawkins T, Hussey PJ, Edwards R. 2009. Enzyme activities and subcellular localization of members of the *Arabidopsis* glutathione transferase superfamily. *J Exp Bot.* 60:1207–1518.

Dixon DP, McEwen AG, Lapthorn AJ, Edwards R. 2003. Forced evolution of a herbicide detoxifying glutathione transferase. *J Biol Chem.* 278:23930–23935.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Edwards R, Dixon DP. 2005. Plant glutathione transferases. *Methods Enzymol.* 401:169–186.

Egan AN, Doyle J. 2010. A comparison of global, gene-specific, and relaxed clock methods in a comparative genomics framework: dating the polyploid history of soybean (*Glycine max*). *Syst Biol.* 59:534–547.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

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.

Frova C. 2003. The plant glutathione transferase gene family: genomic structure, functions, expression and evolution. *Physiol Plant.* 119:469–479

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

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704

Habig WH, Pabst MJ, Jakoby WB. 1974. Glutathione S-transferases. The first enzymatic step in mercapturic acid formation. *J Biol Chem.* 249:7130–7139.

Jain M, Ghanashyam C, Bhattacharjee A. 2010. Comprehensive expression analysis suggests overlapping and specific roles of rice glutathione S-transferase genes during development and stress responses. *BMC Genomics* 11:73.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97–100.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol.* 6:29.

Lan T, Yang ZL, Yang X, Liu YJ, Wang XR, Zeng QY. 2009. Extensive functional diversification of the *Populus* glutathione *S*-transferase supergene family. *Plant Cell* 21:3749–3766.

Layne E. 1957. Spectrophotometric and turbidimetric methods for measuring proteins. *Methods Enzymol.* 3:447–455.

Li HM, Rotter D, Bonos SA, Meyer WA, Belanger FC. 2005. Identification of a gene in the process of being lost from the genus *Agrostis*. *Plant Physiol.* 138:2386–2395.

Liu YJ, Han XM, Ren LL, Yang HL, Zeng QY. 2013. Functional divergence of the GST supergene family in *Physcomitrella patens* reveals complex patterns of large gene family evolution in land plants. *Plant Physiol.* 161:773–786.

Lou P, Wu J, Cheng F, Cressman LG, Wang X, McClung CR. 2012. Preferential retention of circadian clock genes during diploidization following whole genome triplication in *Brassica rapa*. *Plant Cell* 24:2415–2426.

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

McClean PE, Mamidi S, McConnell M, Chikara S, Lee R. 2010. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC Genomics* 11:184.

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. 2003. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15:809–834.

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu SH. 2014. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish *Raphanus raphanistrum* and three other Brassicaceae species. *Plant Cell* 26:1925–1937.

Oakley AJ. 2005. Glutathione transferases: new functions. *Curr Opin Struct Biol.* 15:716–723.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21:3718–3731.

Ricci G, Caccuri AM, Lo Bello M, Pastore A, Piemonte F, Federici G. 1994. Colorimetric and fluorometric assays of glutathione transferase based on 7-chloro-4-nitrobenzo-2-oxa-1,3-diazole. *Anal Biochem.* 218:463–465.

Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, May G, Stacey G, Doerge RW, Jackson SA. 2013. The fate of duplicated genes in a polyploid plant genome. *Plant J.* 73:143–153.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440:341–345.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47:868–876.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A.* 108:4069–4074.

Semon M, Wolfe KH. 2007. Consequences of genome duplication. *Curr Opin Genet Dev.* 17:505–512.

Smirnoff N. 1996. The function and metabolism of ascorbic acid in plants. *Ann Bot.* 78:661–669.

Soltis PS, Liu X, Marchant DB, Visger CJ, Soltis DE. 2014. Polyploidy and novelty: Gottlieb's legacy. *Philos Trans R Soc Lond B Biol Sci.* 369:20130351.

Tang ZX, Yang HL. 2013. Functional divergence and catalytic properties of dehydroascorbate reductase family proteins from *Populus tomentosa*. *Mol Biol Rep.* 40:5105–5114.

Thomas BC, Pedersen B, 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 E, Jiang Y, Chen L, Zou J, Liu F, Meng J. 2010. Synthesis of a Brassica trigenomic allohexaploid (*B. carinata* × *B. rapa*) de novo and its stability in subsequent generations. *Theor Appl Genet.* 121:1431–1440.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* 24:1334–1347.

Yang Q, Liu YJ, Zeng QY. 2014. Biochemical functions of the glutathione transferase supergene family of *Larix kaempferi*. *Plant Physiol Biochem.* 77:99–107.

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

Zhang JZ. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.