

# Extensive Functional Diversification of the *Populus* Glutathione S-Transferase Supergene Family <sup>© | W</sup>

Ting Lan,<sup>a,b,1</sup> Zhi-Ling Yang,<sup>a,b,1</sup> Xue Yang,<sup>a,b</sup> Yan-Jing Liu,<sup>a</sup> Xiao-Ru Wang,<sup>a,c</sup> and Qing-Yin Zeng<sup>a,2</sup>

<sup>a</sup>State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

<sup>b</sup>Graduate School, Chinese Academy of Sciences, Beijing 100049, China

<sup>c</sup>Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, SE-901 87 Umeå, Sweden

**Identifying how genes and their functions evolve after duplication is central to understanding gene family radiation. In this study, we systematically examined the functional diversification of the glutathione S-transferase (GST) gene family in *Populus trichocarpa* by integrating phylogeny, expression, substrate specificity, and enzyme kinetic data. GSTs are ubiquitous proteins in plants that play important roles in stress tolerance and detoxification metabolism. Genome annotation identified 81 GST genes in *Populus* that were divided into eight classes with distinct divergence in their evolutionary rate, gene structure, expression responses to abiotic stressors, and enzymatic properties of encoded proteins. In addition, when all the functional parameters were examined, clear divergence was observed within tandem clusters and between paralogous gene pairs, suggesting that subfunctionalization has taken place among duplicate genes. The two domains of GST proteins appear to have evolved under differential selective pressures. The C-terminal domain seems to have been subject to more relaxed functional constraints or divergent directional selection, which may have allowed rapid changes in substrate specificity, affinity, and activity, while maintaining the primary function of the enzyme. Our findings shed light on mechanisms that facilitate the retention of duplicate genes, which can result in a large gene family with a broad substrate spectrum and a wide range of reactivity toward different substrates.**

## INTRODUCTION

In eukaryotes, most structural and regulatory genes are members of gene families that vary in size and genomic organization. In plant genomes, high frequencies of gene duplication are observed due to the frequent occurrence of genomic segmental duplications and polyploidization (*Arabidopsis* Genome Initiative, 2000; Tuskan et al., 2006). Duplication of individual genes, chromosomal segments, or whole genomes have long been thought to supply raw genetic material, allowing functional divergence and rapid biological evolution (Ohno, 1970; Lynch and Conery, 2000). Hence, understanding the genomic and functional evolution of gene families is essential for understanding the phenotypic diversification of organisms and their genetic systems.

Various functional fates have been proposed for duplicated genes (Ohno, 1970; Hughes, 1994; Force et al., 1999; Moore and Purugganan, 2005), including (1) retention of the original gene function; (2) loss of gene function by pseudogenization (non-

functionalization); (3) acquisition of a novel function through neofunctionalization; or (4) partitioning of the ancestral gene function by subfunctionalization. Theoretical and experimental studies have advanced our understanding of the possible retention mechanisms of duplicate genes, but large amounts of comparative biochemical or physiological data are required to reconstruct the evolutionary steps that have resulted in the functional diversification of gene families. Investigating the functional divergence of a whole family of genes is technically challenging even for model organisms. Hence, sufficient information has only been acquired for few gene families to date, and patterns of functional diversification and genetic factors governing the evolution of most classes of gene families in plants remain largely unknown.

Glutathione S-transferases (GSTs; EC 2.5.1.18) are multifunctional proteins encoded by a large gene family found in all cellular organisms. In plants, GSTs are divided into seven classes: theta, zeta, phi, tau, lambda, glutathione-dependent dehydroascorbate reductase (DHAR), and tetrachlorohydroquinone dehalogenase (Smith et al., 2004; Basantani and Srivastava, 2007). Recently, based on structural similarities, the  $\gamma$ -subunit of the eukaryotic translation elongation factor 1B (EF1B- $\gamma$ ) has also come to be regarded as a member of the GST family (Jeppesen et al., 2003; Oakley, 2005). In *Arabidopsis thaliana*, the GST gene family consists of 53 members (Dixon et al., 2002). Detoxification of xenobiotics is considered to be the main function of plant GSTs, but other functions include protecting cells from a wide range of biotic and abiotic stressors, including pathogen attack, heavy metal toxins, oxidative stress, and UV radiation

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Qing-Yin Zeng (qingyin.zeng@ibcas.ac.cn).

<sup>1</sup> These authors contributed equally to this work.

<sup>2</sup> Address correspondence to qingyin.zeng@ibcas.ac.cn.

Some figures in this article are displayed in color online but in black and white in the print edition.

Online version contains Web-only data.

www.plantcell.org/cgi/doi/10.1105/tpc.109.070219

(Kampranis et al., 2000; Loyall et al., 2000; Mueller et al., 2000; Agrawal et al., 2002). Since they are enzymatic proteins, studies of GST functional evolution can exploit the power afforded by combining genomic structure, gene expression, and enzyme biochemical analyses of GST variants in attempts to elucidate the effects of molecular changes on gene function.

Analysis of the *Populus trichocarpa* genome suggests a whole-genome duplication event occurred recently (in evolutionary terms) on the stem lineage of the Salicaceae, ~60 to 65 million years ago, in addition to another, much more ancient large-scale duplication event shared by *Populus* and *Arabidopsis* (Tuskan et al., 2006). The complex history of genome duplications and chromosomal rearrangements in *Populus* provides an opportunity to study the patterns of gene family expansion in the course of genome evolution. In this study, we conducted a genome-wide annotation of the GST gene family in *Populus* to identify the genetic events responsible for its expansion and organization. Functional diversification of the gene family was characterized by examining the gene expression responses to abiotic stresses and enzymatic properties of the encoded proteins. By combining phylogenetic reconstruction and functional assays, we quantitatively addressed the complex patterns of diversification of the gene family at three levels: among GST classes within the family, among members within tandem-arrayed GST clusters, and between pairs of paralogous duplicate genes. The genome-wide, multifaceted approach we employed provides new insights into the mechanisms of gene family expansion and functional evolution.

## RESULTS

### Large GST Gene Family in *Populus*

A total of 81 full-length genes encoding putative GST proteins were identified in the *P. trichocarpa* genome (see Supplemental Table 1 online), among which nine sequences were considered to be putative pseudogenes based on the presence of a frame shift disrupting the coding region or a stop codon occurring prematurely, resulting in a truncated protein. After removing these stop codons or revising the frame shifts by deleting one or two nucleotides, these nine full-length sequences were included in the phylogenetic and gene expression analyses. The predicted proteins encoded by these 81 genes were initially classified based on the National Center for Biotechnology Information's (NCBI) conserved domain analysis, which divided them into eight classes. The tau and phi GSTs were the most numerous, being represented by 58 and nine members, respectively. The lambda, DHAR, and EF1B $\gamma$  GST classes were each represented by three members, both the zeta and theta classes by two members, and the tetrachlorohydroquinone dehalogenase class by just one member.

Phylogenetic relationships among the *Populus* GSTs were reconstructed using a maximum likelihood (ML) procedure. On the ML tree, the 81 GSTs were grouped into eight distinct groups with high bootstrap support (Figure 1A). These groups corresponded to the classes identified by domain structures. Striking gene structure conservation was found within each GST class.

All 58 tau GST genes contain a single intron at a conserved position, except for *GSTU23*, 47, and 48, all of which have an insertion that disrupts their N-terminal domain, making them appear to be pseudogenes (Figure 1C). Intron length varies from 76 bp in *GSTU17* to 1106 bp in *GSTU8*. All of the nine phi GST genes have a two-intron/three-exon structure (Figure 1C) with exons of similar length and a highly conserved first intron position, except for the putative pseudogene *GSTF9*, which has a shorter sequence. In contrast with the tau and phi GSTs, the gene structures of the other minor GST classes are more variable (Figure 1C); members of the lambda and zeta classes contain nine exons, while members of the theta and DHAR have seven and six exons, respectively. The EF1B $\gamma$  GST genes consist of two parts: a GST domain and an EF1B $\gamma$  domain. In their GST domain, five introns were observed. The class-specific gene structure further supports the subfamily designations among the 81 GSTs.

In addition to full-length GST genes, 50 partial GST fragments (43 tau type, three phi type, two DHAR type, one theta type, and one lambda type) were identified in the *Populus* genome (see Supplemental Table 2 online); these were considered to be pseudogenes. The length of these fragments ranged from 36 to 175 amino acid residues. Domain structure analysis identified nine fragments that contained both partial N- and C-terminal domain sequences; 19 had only a partial N-terminal domain and 22 had only a partial C-terminal domain. We were unable to analyze the phylogenetic relationships of these fragments reliably because of their small size. These short and seemingly random remnants likely reflect more ancient pseudogenization events.

### Genomic Organization of the *Populus* GST Gene Family

The physical locations of 66 of the 81 full-length GSTs were assigned to 15 of the 19 *Populus* chromosomes (Figure 2A), while the other 15 were assigned to 14 as-of-yet unattributed scaffold fragments (see Supplemental Table 1 online). Of the 50 GST fragments, 32 were assigned to one of nine chromosomes and the other 18 on 16 scaffolds (see Supplemental Table 2 online). The distribution of the GST genes among the chromosomes appears to be uneven: chromosomes 7, 9, 17, and 18 harbor no GST genes or GST fragments, while relatively high densities of GSTs were discovered in some locations on chromosomes 1, 8, 10, 11, 14, and 19, where GSTs are arranged in clusters (Figure 2A). These clusters consist mainly of tau and phi GSTs; 37 of the full-length tau GSTs and 20 tau-type fragments are arranged in six clusters (clusters I to V and VII) on five chromosomes (1, 8, 10, 11, and 19), and four phi GSTs are organized in one cluster (cluster VI) on chromosome 14. Members of the minor GST classes are sparsely distributed at single loci on different chromosomes.

Previous analysis of the *Populus* genome has identified paralogous segments created by the whole-genome duplication event in the Salicaceae (salicoid duplication), ~60 to 65 million years ago (Figure 2A) (Tuskan et al., 2006). The distributions of GST genes relative to the duplicate genomic blocks are illustrated in Figure 2A. Of the 66 mapped GSTs, only 15 (clusters IV and VI, *GSTU22*, 34, 35, 41, 45, *GSTL2*, and *EF1B $\gamma$ 1*) are located

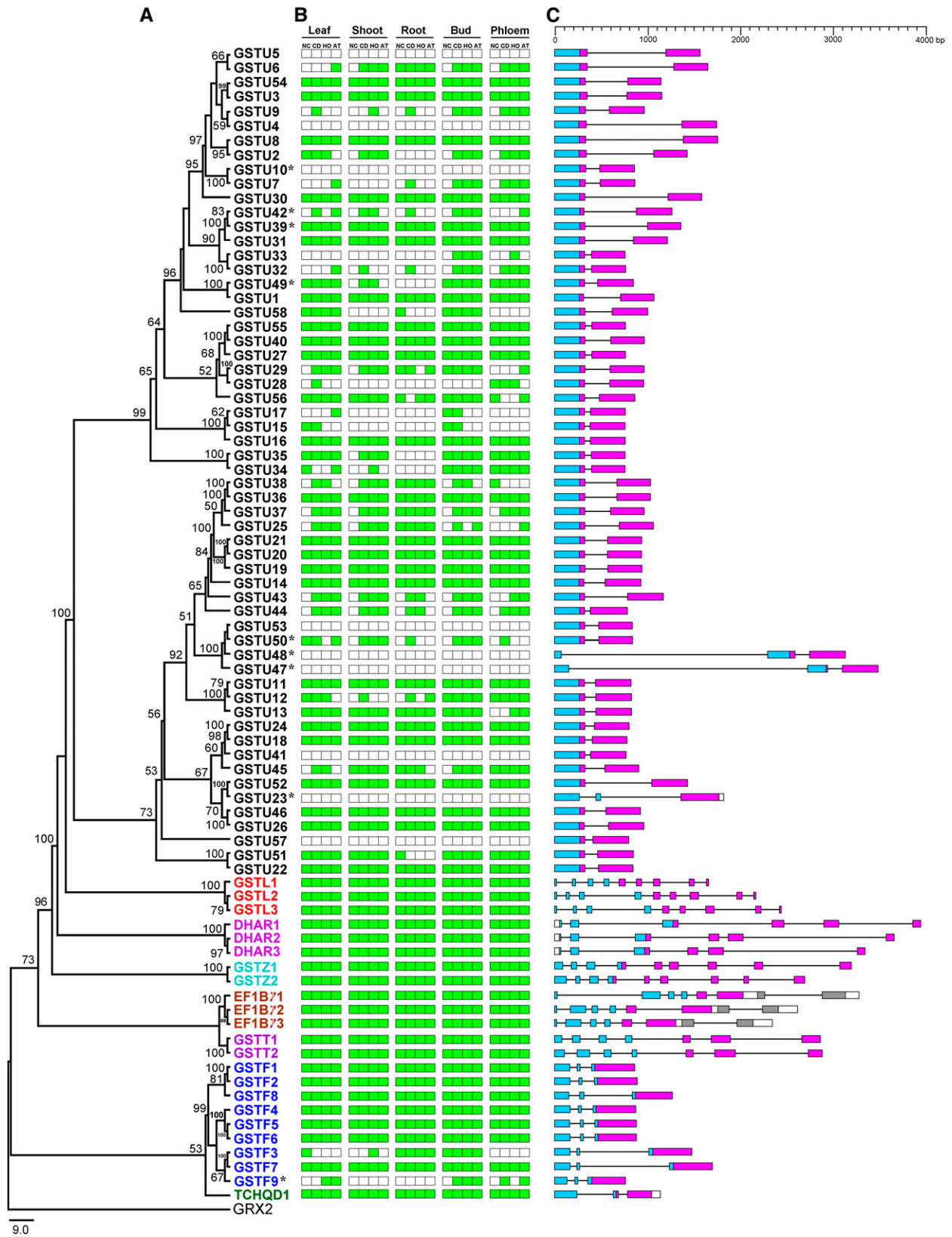


Figure 1. Phylogenetic Relationships among *Populus* GSTs, Their Expression Patterns, and Gene Structure.

outside any duplicate blocks. Four duplicate pairs (*GSTU18/24*, *GSTU46/26*, *DHAR2/3*, and *GSTT1/2*) are each located in a pair of paralogous blocks and can be considered to be direct results of the duplication event (Figure 2A). Similarly, tau class cluster pairs I/V and II/III correspond to paralogous blocks created by the salicoid duplication. The other two smaller clusters, IV and VII, lack corresponding duplicates (Figure 2A). In contrast with cluster IV, which is located in a single copy genome block on chromosome 11, cluster VII is located in a region that is thought to have been duplicated. However, no GSTs were found in the corresponding duplicate block, suggesting that the corresponding homolog may have been deleted after the duplication event. This observation corroborates the trend that the most abundant gene losses in eukaryotes occur following whole-genome duplication (Lynch, 2007). In contrast with the tau class, the expansion of the phi class seems not to have been affected by the salicoid duplication event. The only phi cluster, containing *GSTF4*, 5, 6, and 7 (cluster VI), is located in a nonduplicated region on chromosome 14.

Among the seven clusters, cluster I is the largest, consisting of 11 full-length tau and eight tau-type fragments arranged in tandem in a 119-kb region on chromosome 1. There are complex gene orientations among members of this cluster (Figure 2B). Cluster II, with four tau genes and four tau-type fragments, spans a region of 40 kb on chromosome 8. Nine tau and seven tau-type fragments are organized into cluster III, located in a 57-kb region on chromosome 10. On chromosome 11, there are two tau clusters (IV and V), one with four genes arranged in a head-to-head tandem and the other with six genes arranged in the same orientation in a nearby region. Cluster VII on chromosome 19 is small and contains *GSTU15*, 16, and 17. The phi cluster (VI), with four members tandemly arranged in the same orientation, spans a 34-kb region on chromosome 14 (Figure 2B).

We attempted to reconstruct the expansion history in each cluster by reconciliation of both the gene tree and the positions of genes within clusters. The most parsimonious scenario for gene duplication, loss, and rearrangement is presented in Figure 3. The events that led to the expansion of the four large clusters (I, II, III, and V) appear to have been complex, likely involving segmental duplication followed by a series of tandem duplications and rearrangements. Cluster I is located in a transposon-rich region (Figure 2B); hence, transposable elements could have driven gene duplications by inadvertently carrying copies of genes during transposition events and/or by facilitating unequal crossovers (Hancock, 2005). The large number of pseudogene fragments in this cluster also makes it difficult to delimit the detailed evolutionary steps leading to its present composition and structure. The expansion history of cluster V, on the other

hand, is simpler and seems to have involved four rounds of tandem duplications. Judging from the gene tree and the structure of cluster II and III, it is likely that a three-gene cluster was duplicated by the whole-genome duplication event. After that, four rounds of tandem duplications and a possible rearrangement took place in cluster III, and three rounds of tandem duplications probably occurred in cluster II (Figure 3A). The other two smaller tau clusters, IV and VII, formed two well-supported clades (Figure 3A). Each of these two clusters may be created by two rounds of localized duplication events. Phylogenetic relationships among members of the phi cluster suggest an ancestral copy gave rise to the progenitors of *GSTF3/7* and *GSTF4*, 5, and 6. *GSTF4*, 5, and 6 formed an array with the same orientation (Figure 2B) generated by tandem duplications (Figure 3B). The grouping of *GSTF3* and 7 into one well-supported branch suggests that either a segmental duplication between the two loci or a transposition or retroposition event occurred.

### Molecular Evolution

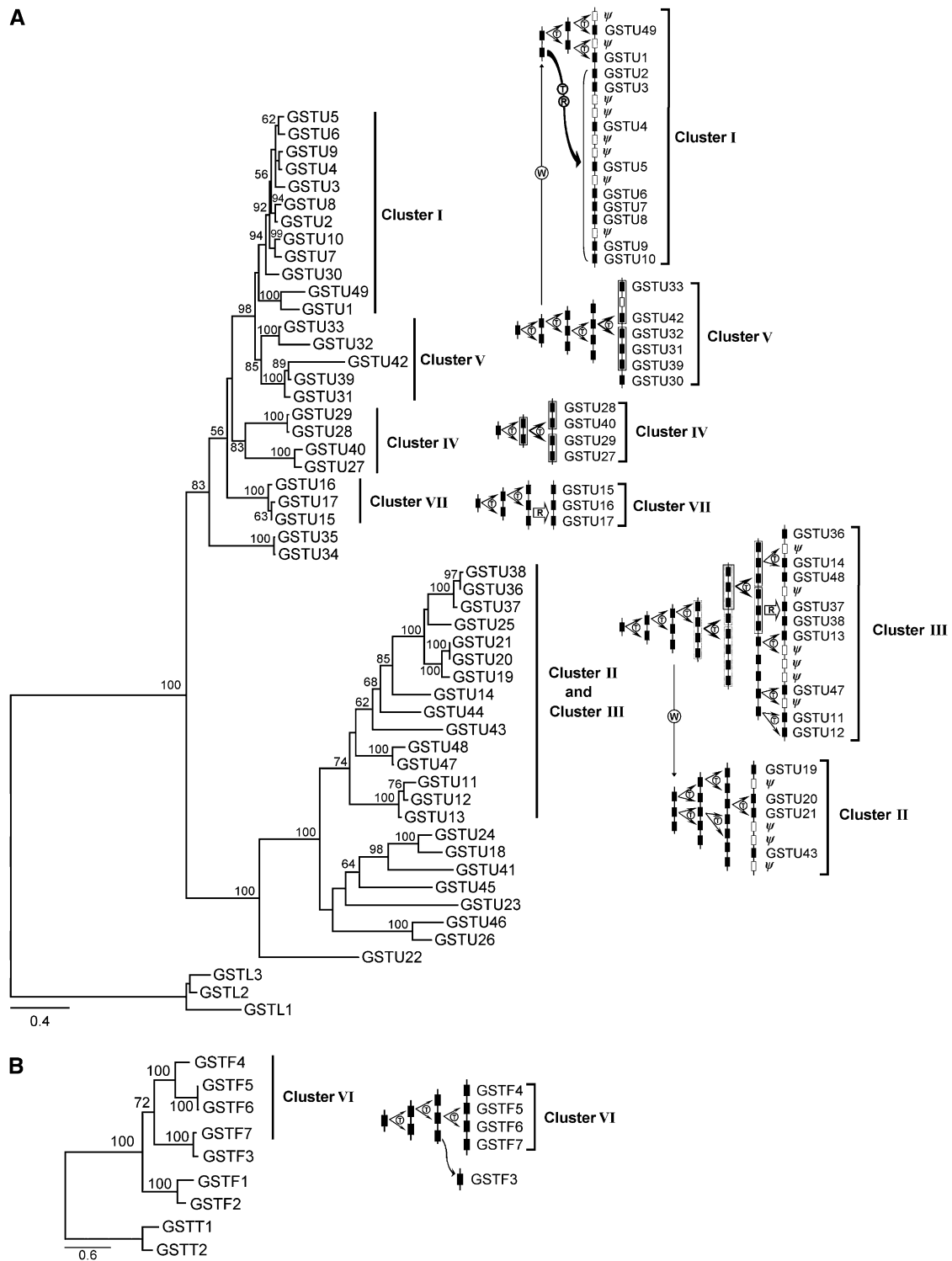
GST proteins consist of two well-defined domains, the N-terminal domain that binds the primary substrate glutathione (GSH) and the C-terminal domain that binds the secondary substrate (Edwards et al., 2000). To test for deviation in the substitution pattern of the two domains, we partitioned the 58 tau GST sequences into N- and C-terminal regions (the linker between the two domains was not included in this analysis). The ratio of nonsynonymous versus synonymous substitutions ( $\omega = d_N/d_S$ ) is an indicator of the history of selection acting on a gene or gene region. Ratios significantly  $<1$  are suggestive of purifying selection, whereas ratios  $>1$  suggest directional selection. A plot of  $d_N/d_S$  for the N- versus C-terminal domains is shown in Figure 4A. The results suggest that both domains have evolved mainly under the influence of purifying selection, but the selective constraint on the C-terminal domain was more relaxed than on the N-terminal domain ( $t$  tests,  $P < 0.001$ ). This pattern was also evident in the phi GST class (Figure 4B) ( $t$  tests,  $P < 0.003$ ).

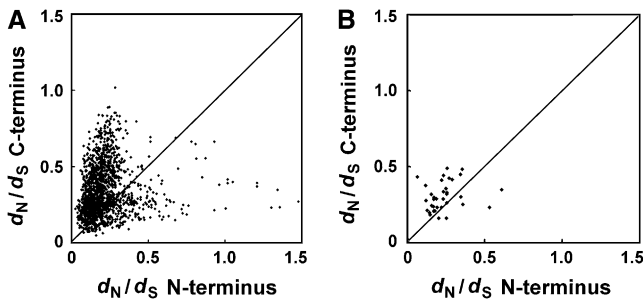
To infer the influence of selection on the expansion of the tau and phi classes, we estimated  $\omega (=d_N/d_S)$  values for all branches using ML codon models. Two assumptions were tested: the fixed one-ratio model that assumes the same  $\omega$  ratio for all branches and the free-ratio model that assumes an independent  $\omega$  for each branch in the gene tree. The log-likelihood values under the one-ratio model were  $\ln L = -20,771.8458$  and  $-6239.9762$ , with estimates of  $\omega = 0.27562$  and  $0.24328$  for the tau and phi GST classes, respectively. Under the free-ratio model, the log-likelihood values were  $\ln L = -20,613.6203$  and  $-6217.2478$  for

### Figure 1. (continued).

An ML procedure using the JTT model and 1000 bootstrap replicates was used in the phylogeny reconstruction; putative pseudogenes are indicated with asterisks; numbers at each node in the ML tree signify bootstrap values. GST genes designated as GSTU, F, T, Z, and L correspond to tau, phi, theta, zeta, and lambda class GSTs, respectively. The unmasked and masked sequence alignments used to reconstruct the ML tree are available as Supplemental Data Sets 2A and 2B online, respectively. In (B), the green box indicates positive detection of gene expression in the corresponding tissue under normal growth conditions (NC) and following CDNB (CD),  $H_2O_2$  (HO), and atrazine (AT) treatments. In (C), the GST N-terminal domain, C-terminal domain, and EF1B $\gamma$  domain are highlighted by the blue, purple, and gray boxes, respectively, while introns are indicated as lines.







**Figure 4.**  $d_N/d_S$  Plots for the C-Terminal Domain versus the N-Terminal Domain of 58 Tau and Nine Phi GSTs.

(A) Tau GSTs.

(B) Phi GSTs.

the rate of evolution between the two clades, a global clock model and a local clock two-rate model, one for each clade, were tested. The log likelihood values under the global clock model and two-rate model were  $\ln L = -21,062.213421$  and  $\ln L = -21,058.829828$ , respectively. LRTs of the two assumptions indicated that a global clock should be rejected by the two-rate clock ( $P < 0.01$ ), suggesting that evolutionary rate is different between the two clades: clade II evolved 1.5 times faster than clade I. To test if there were changes in selective pressure between the two clades, a two-ratio model in which the two clades were each assigned to a different ratio (Table 1) was explored. LRTs indicated that the null model, where the two clades have the same ratio, could not be rejected, thus suggesting that selective pressure has been similar between the two clades after the ancestral gene duplication.

The site model analyses indicated that log-likelihood values were significantly higher under the M8 model than under the M7 model for the tau GST genes, but not the phi GST genes (Table 1). In the tau class, 10 amino acids were identified by Bayes Empirical Bayes analysis (Yang et al., 2005) as candidates for positively selected sites (Table 1), including two with posterior probabilities  $>0.95$  (alignment positions 234 and 237 in Supplemental Figure 2 online). With one exception (alignment position 43 in Supplemental Figure 2 online), the amino acids were located in the C-terminal domain of the protein.

In *Arabidopsis*, the GST gene family consists of 53 members, of which 28 are tau class (Dixon et al., 2002). Joint analysis of tau GSTs from *Populus* and *Arabidopsis* revealed three major *Populus*-specific and *Arabidopsis*-specific clades, respectively (Figure 5). This suggests that independent expansion of the tau class has occurred in the two lineages and at least 51 novel tau GSTs in *Populus* have been acquired, compared with 21 in *Arabidopsis*, after the most recent common ancestor of *Populus* and *Arabidopsis* 100 to 120 million years ago. Such a gene birth rate is much higher than the average rate of one gene per 100 million years reported for eukaryotes (Lynch and Conery, 2000). In addition, 43 tau-type pseudogene fragments were discovered in *Populus*, in contrast with very few in *Arabidopsis*. These results suggest that tau GSTs have been subject to faster rates of duplication in *Populus* than in *Arabidopsis*. This high duplication rate is partly attributable to the salicoid whole-genome duplica-

tion event followed by localized tandem duplications, since 69% of the tau GST genes and 56% of the tau fragments are located in the duplicated genome blocks. The existence of more fragments and pseudogenes in *Populus* could be indicative of a lower rate of complete deletion or divergence beyond the threshold for detection of homology. *Arabidopsis* and close relatives with small genomes have been suggested to possess as yet unidentified mechanisms impeding amplification, and/or they have an efficient mechanism of continuous removal of amplified sequences to counteract their proliferation (Lysak et al., 2009).

### Structure Modeling

Protein three-dimensional structure modeling illustrated that, in general, the structures of *Populus* tau and phi GSTs are conserved, especially in the N-terminal domain (Figure 6). The simulated structures of 55 tau and eight phi GSTs were superimposed to evaluate the goodness of fit of the overall topologies (Figure 6). This examination showed that all members within each class shared the same conformation of the structural elements of  $\alpha$ -helices and  $\beta$ -sheets, but structural modifications are present in loop regions of the C-terminal domain and in the linker between the N- and C-terminal domains, where the structural conservation is relatively weak. The protein sequence differentiation in these loop regions is high among members of both the tau and phi classes (see Supplemental Figures 2 and 3 online) due to either relaxed functional constraints or divergent directional selection. In fact, the majority of the putative positively selected sites (7 out of 10) are in the loops of the C-terminal domains (Figure 6).

### Expression of GST Genes under Normal Growth Conditions and Abiotic Stress

The expression patterns of GST gene family members were examined by RT-PCR and quantitative RT-PCR under normal growth conditions and in response to stress treatments (1-chloro-2,4-dinitrobenzene [CDNB],  $H_2O_2$ , and atrazine applications). Substantially more variation in expression patterns was found among the tau class members than among members of the other minor GST classes (Figure 1B). Of the 58 tau GSTs, 23 were expressed in all tissues under all growth conditions, while nine (*GSTU4*, 5, 10, 23, 41, 47, 48, 53, and 57) were neither expressed in any tissue nor in response to any treatment applied in this study. Thus, these nine genes are expressed at subdetectable levels, or they are only induced in response to treatments and/or in tissues not examined in our study, or they are pseudogenes. The other 26 tau members were selectively expressed either in response to a specific treatment and/or in a specific tissue. For instance, under normal growth conditions *GSTU6*, 25, and 37 showed root-specific expression, while under the stress treatments, they were expressed in all five examined tissues. *GSTU7*, 9, 32, 33, 42, 43, and 44 expression was not detected under normal growth conditions, but they were expressed in response to treatment with CDNB,  $H_2O_2$ , or atrazine, indicating that they are expressed in stress responses. The expression patterns of phi and other minor GST classes were more homogenous: except for *GSTF3* and 9, all of these genes

**Table 1.** Summary Statistics for Detection of Selection Using Branch and Site Models of PAML

Model		Estimates of Parameters <sup>a</sup>	-ln L	$\chi^2$	P	Positively Selected Sites <sup>b</sup>
Site model						
Tau GSTs	M0	0.28513	19,313.35			
	M1a	$p_0 = 0.68910, p_1 = 0.31090$	18,918.64			
	(nearly neutral)	$\omega_0 = 0.20084, \omega_1 = 1.00000$				
	M2a	$p_0 = 0.68910, p_1 = 0.15705, p_2 = 0.15385$	18,918.64	0		
	(positive selection)	$\omega_0 = 0.20084, \omega_1 = 1.00000, \omega_2 = 1.00000$				
	M3 (discrete)	$p_0 = 0.32505, p_1 = 0.46802, p_2 = 0.20692$	18,801.71	1,023.28	<0.001	
		$\omega_0 = 0.08436, \omega_1 = 0.30996, \omega_2 = 0.91017$				
	M7 ( $\beta$ )	$P = 0.89944, q = 1.62442$	18,801.59			
M8 ( $\beta$ & $\omega$ )	$p_0 = 0.90574, P = 1.24556, q = 3.15446$ $(p_1 = 0.09426) \omega = 1.28108$	18,779.36	44.46	<0.001	43, 114, 115, 116, 122, 123, 233, 234*, 236, 237*	
Phi GSTs	M0	$\omega = 0.22213$	4,832.43			
	M1a	$p_0 = 0.59667, p_1 = 0.40333$	4,788.14			
	(nearly neutral)	$\omega_0 = 0.15845, \omega_1 = 1.00000$				
	M2a	$p_0 = 0.59667, p_1 = 0.23067, p_2 = 0.17266$	4,788.14	0		
	(positive selection)	$\omega_0 = 0.15845, \omega_1 = 1.00000, \omega_2 = 1.00000$				
	M3 (discrete)	$p_0 = 0.16766, p_1 = 0.41839, p_2 = 0.41395$	4,750.95	162.96	<0.001	
		$\omega_0 = 0.00759, \omega_1 = 0.17187, \omega_2 = 0.48703$				
	M7 ( $\beta$ )	$P = 0.81835, q = 1.97496$	4,754.71			
M8 ( $\beta$ and $\omega$ )	$p_0 = 0.99999, P = 0.81836, q = 1.97502$ $(p_1 = 0.00001) \omega = 3.98485$	4,754.71	0			
Branch model						
One ratio		$\omega = 0.28513$ for all branches	19,313.35			
Two ratios		$\omega_0 = 0.26811$ for clade II	19,311.78	3.14		
		$\omega_1 = 0.30691$ for clade I				
Two ratios		$\omega_0 = 0.49690$ for clade I	19,312.76	1.18		
		$\omega_1 = 0.28427$ for clade II				

\*Posterior probability >95% of having  $\omega > 1$ .

<sup>a</sup>The proportion of sites ( $p_0, p_1$ , etc.) estimated to have  $\omega_0, \omega_1$ , etc.

<sup>b</sup>The numbering of residues identified by Bayes empirical bayes analysis (Yang et al., 2005) corresponds to their alignment positions in Supplemental Figure 2 online. Clade I and clade II in branch model refer to the two clades of tau class as shown in Supplemental Figure 1.

were expressed in all tissues under all growth conditions (Figure 1B).

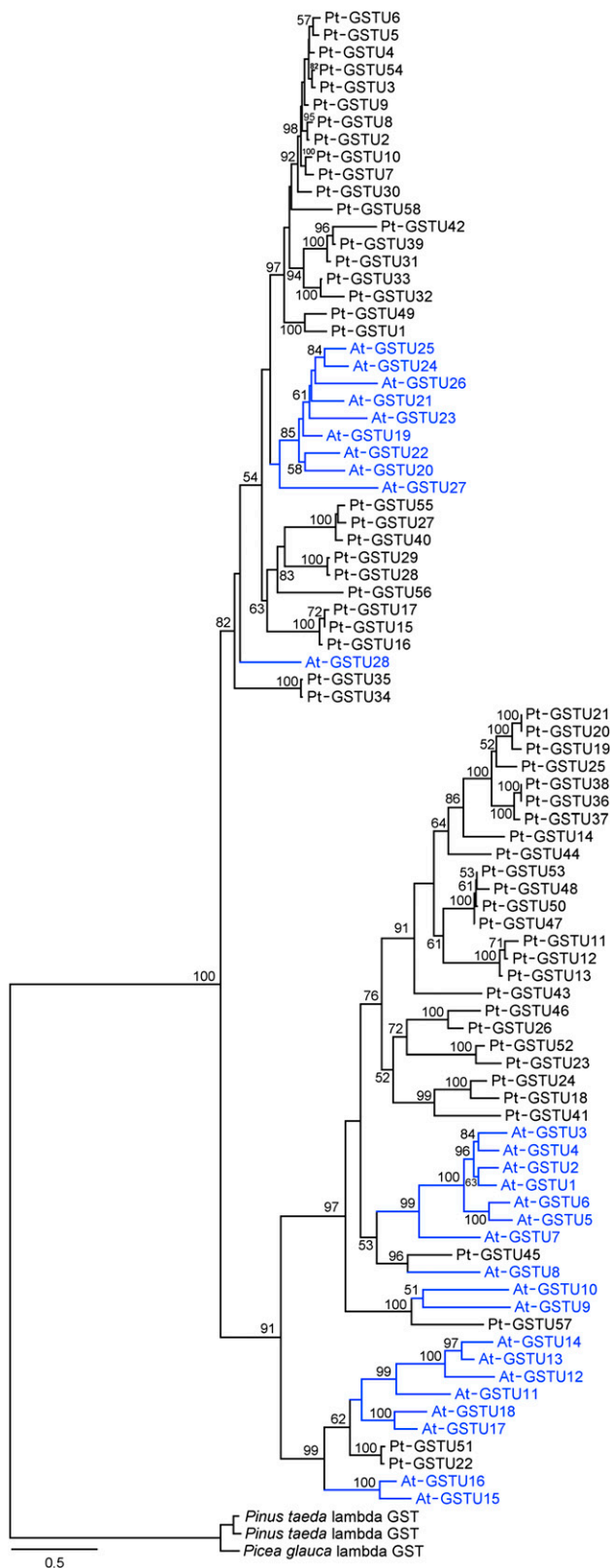
For each of the seven tandem-arrayed GST clusters, marked expression divergence was found among the members. In cluster I, which consists of 11 tau GSTs, three genes (*GSTU1*, 3, and 8) were expressed in all tissues, no transcripts of three others (*GSTU4*, 5, and 10) were detected in any of the tested tissues, and the other five (*GSTU2*, 6, 49, 7, and 9) showed restricted tissue-specific expression patterns under normal and/or stress conditions (Figure 1B). Furthermore, although *GSTU1*, 3, and 8 were expressed in all tested tissues under all conditions, their expression levels in response to the stress treatments differed significantly (Figure 7): *GSTU1* was upregulated 15- to 38-fold in shoots in response to all three treatments; *GSTU8* expression increased 77- and 105-fold in leaves in response to the  $H_2O_2$  and CDNB treatments, respectively; *GSTU3* expression increased 25-fold in leaves in response to atrazine and 9- to 15-fold in shoots in response to  $H_2O_2$  and atrazine.

Similarly, the four phi GSTs (*GSTF4*, 5, 6, and 7) in cluster VI, which were expressed in all tissues under all growth conditions, differed significantly in expression levels in response to the different treatments (Figure 7). In general, most of these genes

were upregulated in the tested tissues, for example, in leaves *GSTF4*, 5, and 6, expression increased 69- to 141-fold in response to CDNB, relative to the control, while *GSTF7* expression increased only fivefold. In response to  $H_2O_2$  and atrazine treatments, expression levels of these genes increased by 24- to 85-fold and 16- to 88-fold, respectively. In shoots, *GSTF4* showed the strongest response to the treatments, while in roots, *GSTF6* expression changed only slightly, but *GSTF5* and 7 were upregulated 20- to 86-fold. These results indicate that rapid divergence has occurred in the regulatory regions of the GST genes in the same clusters.

For duplicate gene pairs, four categories of expression pattern were observed. In the first category, found in 12 gene pairs, both of the duplicates were expressed in all tissues under normal growth and stress treatments (AA model in Table 2). In the second category, occurring in five gene pairs, one copy was expressed, while the other was not detected in any tissue type under any growth conditions (AN model in Table 2), suggesting that one duplicate gene may have become a pseudogene or evolved a new function not identified in our study. In the third category, found in seven gene pairs, one copy of each duplicate pair was expressed in all tissues under all growth conditions,





**Figure 5.** Phylogenetic Relationships of the *Populus* and *Arabidopsis* Tau GST Proteins.

while the other was expressed only following a specific treatment in a specific tissue (AI model in Table 2). In the fourth category, found in four gene pairs, both duplicates showed selective expression under different tissue treatment combinations (SE model in Table 2). Interestingly, duplicate gene pairs created by the whole-genome duplication event or segmental duplication all fell within the first category, with both of the duplicates showing a similar expression pattern. Duplicate pairs created by tandem duplication, on the other hand, spanned all four categories of expression pattern (e.g., *GSTU20/21*, *9/4*, *2/8*, and *28/29*).

### Substrate Specificity and Activity of GST Enzymes

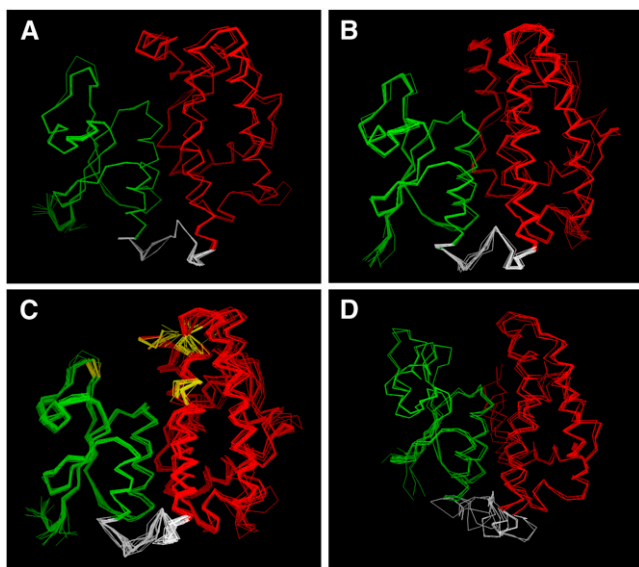
Forty-four *Populus* GSTs with differing evolutionary relationships and histories, including 30 tau, seven phi, three DHAR, two lambda, one theta, and one EF1B $\gamma$  GSTs, were selected for protein expression and purification (Table 3). All of these GSTs were expressed as soluble proteins in *Escherichia coli*, except for three tau genes (*GSTU7*, *18*, and *32*), two lambda genes (*GSTL1* and *2*), and the *GSTT2* gene, which were expressed as inclusion bodies.

The substrate specificity of the purified *Populus* GSTs was investigated to identify catalytic activities that may be related to their biological function, in assays with seven substrates: CDNB, 7-chloro-4-nitrobenzo-2-oxa-1,3-diazole (NBD-Cl), 1,2-dichloro-4-nitrobenzene (DCNB), 4-nitrobenzyl chloride (NBC), ethacrynic acid (ECA), 4-nitrophenyl acetate (4-NPA), and dehydroascorbate (DHA). Of the 30 purified tau GSTs listed in Table 3, 26 showed specific activity toward CDNB, 25 toward NBD-Cl, 13 toward DCNB, 12 toward NBC, and three toward 4-NPA, in accordance with the frequent use of CDNB and NBD-Cl for the detection and determination of plant tau GST activity. Each of two tau GSTs had enzymatic activity toward five substrates, six toward four substrates, 12 toward three substrates, and 10 toward only one to two substrates. Of the seven purified phi GSTs, all had activity toward CDNB and NBD-Cl, three toward both NBC and ECA, and none toward either DCNB or 4-NPA. The DHAR GSTs showed distinct diversification in enzyme specificity from the other GST classes; they showed high activity toward DHA, but no activity toward any of the other six substrates used in this study (Table 3). The GST domain in *EF1B $\gamma$ 1* showed GSH-conjugating activity with CDNB and DCNB. The lambda and theta GSTs were almost inactive toward all the substrates used in our study.

The enzyme assays revealed large variations in specific activities toward different substrates among the tandem-arrayed

Accession numbers for the *Arabidopsis* tau GSTs are presented in Supplemental Table 6 online. *Pinus taeda* lambda GSTs (GenBank accession numbers CV034086 and DR019281) and a *Picea glauca* lambda GST (GenBank accession number EX306134) were used as outgroups. The tree was reconstructed using an ML procedure with the JTT model and 1000 bootstrap replicates. Numbers at each node in the ML tree signify bootstrap values. The sequence alignments used to reconstruct the ML tree are available as Supplemental Data Sets 5A and 5B online.

[See online article for color version of this figure.]



**Figure 6.** Structural Analysis of *Populus* GSTs.

**(A)** Structural superposition of the 28 tau GSTs in clade I (except for GSTU32 and 49) from Supplemental Figure 1 online.

**(B)** The 27 tau GSTs in clade II (except for GSTU23) from Supplemental Figure 1 online.

**(C)** All 55 tau GSTs in **(A)** and **(B)**.

**(D)** Eight phi GSTs (GSTF1 to 8). In each case, the N- and C-terminal domain is illustrated in green and red, respectively, and the linker between the two domains is shown in white. The positions of the putative positively selected sites listed in Table 1 are illustrated in yellow.

members in the clusters. Among the seven examined members of cluster I, the tau GST enzyme activity toward CDNB and NBD-Cl varied from 0 to 12.9  $\mu\text{mol}/\text{min}$  per mg and from 0 to 6.93  $\mu\text{mol}/\text{min}$  per mg, respectively. Four members of this cluster (GSTU1, 2, 8, and 9) each showed activity toward four of the substrates, two members (GSTU3 and 4) showed activity toward three substrates, and GSTU7 showed activity toward only one substrate. GSTU8 and 9 had a similar substrate spectrum, but their specific activity toward each substrate varied twofold to ninefold. Thus, diversification in enzyme specificity and activity toward different substrates has apparently evolved among GSTs in the tandem arrays. This pattern was also observed in tau cluster III and V and phi cluster VI. Of the phi GSTs: although all were reactive with CDNB and NBD-Cl, their enzymatic activities varied 24- and 38-fold, from 0.15 to 3.62  $\mu\text{mol}/\text{min}$  per mg protein for CDNB and from 0.06 to 12.95  $\mu\text{mol}/\text{min}$  per mg protein for NBD-Cl (Table 3).

For duplicate gene pairs, differentiation in enzyme specificity could be categorized into three groups. In the first, the two duplicates showed a similar substrate spectrum, but differed by maximal 56-fold (GSTU18/24 to CDNB) in their enzymatic activity toward each substrate (SS model in Table 2). In the second, which is more encountered, the two duplicates showed a partially overlapping substrate spectrum (PS model in Table 2), suggesting that partial subfunctionalization occurred after gene duplication. In the third, the two duplicates have evolved a

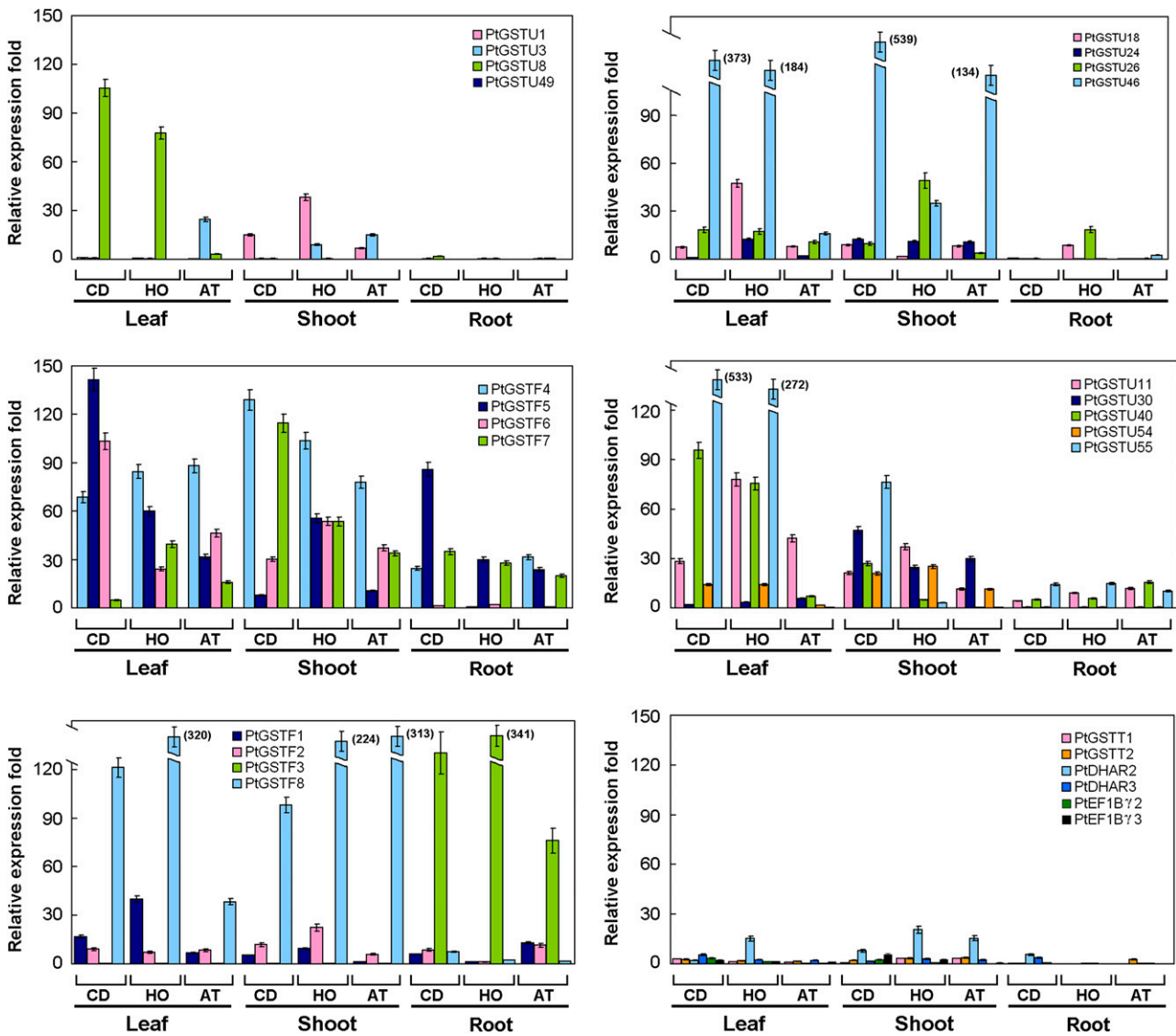
nonoverlapping substrate spectrum (NS model in Table 2), indicating that significant functional diversification in enzyme substrate specificity has occurred. Just one duplicate pair (GSTU26/46) showed this last pattern; GSTU26 was only reactive with CDNB, while GSTU46 was only reactive with ECA.

### Kinetics of the Conjugation Reaction

By definition, all GST enzymes catalyze reactions in which GSH is conjugated to a second substrate. The catalytic properties of 29 selected GST enzymes examined here were characterized in kinetic studies using CDNB, NBD-Cl, and DHA as the second substrate for tau, phi, and DHAR GSTs, respectively, and the resulting kinetic parameters are listed in Table 4. For all tau and phi GSTs, except GSTU4, GSTU16, and GSTF4, the apparent  $K_m^{\text{GSH}}$  values fell within the range of 0.1 to 0.8 mM, indicating that they have similar, high affinities for GSH. This is in accordance with the highly conserved nature of the N-terminal domain of the GSTs (see Supplemental Figures 2 and 3 online), where the enzyme conjugates GSH. The catalytic efficiency ( $k_{\text{cat}}/K_m$ ) for GSH, however, varied significantly among all GSTs (e.g., 1849-fold among tau and 1227-fold among phi GSTs) (Table 4). The C-terminal domain of GSTs, where the enzyme binds the 2nd hydrophobic substrate, is less conserved in plants (see Supplemental Figures 2 and 3 online). Accordingly, pronounced variations in substrate affinities ( $K_m^{\text{CDNB}}$  and  $K_m^{\text{NBD-Cl}}$ ) and catalytic efficiency ( $k_{\text{cat}}/K_m$ ) were observed among the tau GSTs ( $K_m^{\text{CDNB}}$  0.15 to 123.6 mM and  $k_{\text{cat}}/K_m^{\text{CDNB}}$  0.46 to 4152.87  $\text{mM}^{-1} \text{s}^{-1}$ ) and phi GSTs ( $K_m^{\text{NBD-Cl}}$  0.31 to 2.10 mM,  $k_{\text{cat}}/K_m^{\text{NBD-Cl}}$  4.39 to 1361.25  $\text{mM}^{-1} \text{s}^{-1}$ ), suggesting that their C-terminal domains have different conformations.

The tandem-arrayed GSTs varied in both their affinity and catalytic efficiency toward different substrates. For instance, among the six tau GSTs (GSTU1, 2, 3, 4, 8, and 9) examined in cluster I,  $K_m^{\text{CDNB}}$  varied  $\sim 42$ -fold, and their catalytic efficiency ( $k_{\text{cat}}/K_m$ ) for GSH and CDNB varied  $\sim 224$ -fold and 488-fold, respectively. Similar diversification patterns in enzyme kinetic parameters were observed for GSTU11, 12, and 14 in cluster III and GSTU15, 16, and 17 in cluster VII and phi cluster VI.

For duplicate gene pairs, three divergent patterns in kinetic properties for hydrophobic (2nd) substrates were observed. In the first, the two duplicates had similar  $K_m$  values but differed markedly in their catalytic efficiency ( $k_{\text{cat}}/K_m$ ) due to pronounced differences in their  $k_{\text{cat}}$  values. For example, GSTF1/2 showed similar affinity ( $K_m^{\text{NBD-Cl}}$ ) to NBD-Cl, but the catalytic efficiency of GSTF1 was 156-fold higher than that of GSTF2; similar patterns applied to the duplicate pair GSTU3/54. In the second pattern, there was differentiation in the affinity for a substrate between the two duplicates, which correlated positively with their catalytic efficiency. Duplicate pairs GSTU51/22, GSTU4/9, GSTU2/8, and GSTU11/12 fell within this category of kinetic properties. In the third pattern, the two duplicates differed in their affinity for a substrate, and their catalytic efficiencies showed an inverse relationship with affinity; i.e., one had lower affinity (higher  $K_m$ ) but higher catalytic efficiency (higher  $k_{\text{cat}}/K_m$ ) due to compensation by a higher turnover rate ( $k_{\text{cat}}$ ). Duplicate pairs GSTU40/55, GSTU15/17, GSTF5/6, and DHAR2/3 displayed this pattern.



**Figure 7.** Quantitative RT-PCR Analysis of Relative Expression Levels of *Populus* GST Genes under Abiotic Stress.

The expression levels of the assayed genes were normalized to the expression level under normal growth conditions, which was set to 1.0. CD, HO, and AT indicate the CDNB, H<sub>2</sub>O<sub>2</sub>, and atrazine stress treatments, respectively.

**DISCUSSION**

Functional diversification among gene family members is viewed as an important source of evolutionary innovation in complex organisms, and various theoretical models have been proposed to explain the mechanisms involved (Ohno, 1970; Hughes, 1994; Force et al., 1999; Walsh, 2003; Moore and Purugganan, 2005). The most plausible models proposed for the retention of duplicated genes invoke sub- or neofunctionalization. Nevertheless, some previous studies have found these models to be inadequate because many duplicate genes have been found to have little divergence in their sequence and expression (Barakat et al., 2009; Matsumura et al., 2009) or substrate specificity (Zhang et al., 1998; Dean et al., 2008). This has led to the assumption that

the duplicate genes could be redundant in some cases. However, these studies have not examined all axes of potential subfunctionalization due to the difficulties associated with acquiring sufficient systematic biochemical data. Relatively few studies have empirically analyzed the functional diversification of a gene family at different levels of genomic organization: among subfamily classes, within tandem clusters, and in paralogous gene pairs. Our study systematically explored the radiation of a gene family while integrating phylogenetic, expression pattern, substrate specificity, and enzyme kinetic data. Through this integrated approach, we demonstrated that all GST genes appeared to have diverged along at least one of these axes. When all the functional parameters were examined, very few of

**Table 2.** Divergence between Paralogous GST Gene Pairs in *Populus*

No.	Gene 1	Gene 2	$K_s$	$K_a$	$K_a/K_s$	Gene Expression	Substrate Specificities
1 T	GSTU20	GSTU21	0.000	0.000	–	AA	–
2 T	GSTU34	GSTU35	0.009	0.008	0.889	SE	–
3 T	GSTU7	GSTU10	0.015	0.028	1.867	AN	–
4 T	GSTU15	GSTU17	0.015	0.012	0.800	SE	PS
5 T	GSTU36	GSTU38	0.023	0.000	0.000	AI	PS
6 T	GSTU39	GSTU42	0.023	0.021	0.913	AI	–
7 O	GSTU22	GSTU51	0.024	0.015	0.625	AI	PS
8 T	GSTU28	GSTU29	0.030	0.006	0.200	SE	–
9 O	GSTU3	GSTU54	0.034	0.010	0.294	AA	PS
10 O	GSTU50	GSTU53	0.035	0.006	0.171	AN	–
11 T	GSTU4	GSTU9	0.044	0.019	0.432	AN	PS
12 T	GSTU5	GSTU6	0.050	0.030	0.600	AN	–
13 T	GSTU32	GSTU33	0.050	0.046	0.920	SE	–
14 T	GSTF5	GSTF6	0.058	0.004	0.069	AA	PS
15 T	GSTU11	GSTU12	0.062	0.040	0.645	AI	PS
16 O	GSTU40	GSTU55	0.063	0.029	0.460	AA	PS
17 T	GSTU2	GSTU8	0.126	0.018	0.143	AI	SS
18 W	GSTU18	GSTU24	0.170	0.067	0.394	AA	SS
19 O	EF1B $\gamma$ 2	EF1B $\gamma$ 3	0.180	0.050	0.278	AA	–
20 T	GSTU1	GSTU49	0.183	0.078	0.426	AI	–
21 O	GSTF7	GSTF3	0.197	0.037	0.188	AI	PS
22 W	GSTU26	GSTU46	0.212	0.083	0.391	AA	NS
23 W	DHAR2	DHAR3	0.247	0.072	0.291	AA	SS
24 W	GSTT1	GSTT2	0.259	0.104	0.401	AA	–
25 T	GSTF1	GSTF2	0.277	0.078	0.282	AA	SS
26 O	GSTU23	GSTU52	0.322	0.057	0.177	AN	–
27 O	GSTZ1	GSTZ2	0.649	0.095	0.146	AA	–
28 O	GSTL2	GSTL3	0.762	0.078	0.102	AA	–

These gene pairs were identified at the terminal nodes of the gene tree shown in Figure 1. Gene pairs created by tandem duplication (T), whole-genome duplication (W), or other (O) events are indicated in the first column of the table. Synonymous ( $k_s$ ) and nonsynonymous substitution ( $k_a$ ) rates are presented for each pair. Observed gene expression patterns are categorized into four classes: AA, both duplicates were expressed in all tissues under all growth conditions; AN, one duplicate was expressed, while the other was not detected in any tissues under any growth conditions; AI, one duplicate was expressed in all tissues under all growth conditions, while the other was induced-expressed in response to a specific treatment in a specific tissue; SE, both duplicates were selectively expressed either in response to a specific treatment and/or in a specific tissue. The encoded enzyme activity patterns are categorized as follows: SS, both duplicates displayed a similar substrate spectrum; PS, the two duplicates displayed a partially overlapping substrate spectrum; NS, the two duplicates displayed a nonoverlapping substrate spectrum.

the GSTs had identical properties. Even the most recently diverged paralogs differed in their expression, substrate spectrum, and/or catalytic efficiency, suggesting that duplicates have a relatively high rate of diverging rapidly in function and (partial) subfunctionalization has indeed taken place. This may explain why so many GST duplicates have been retained in the poplar genome.

Clear divergence in expression patterns was observed among the *Populus* GSTs in response to different stress treatments. Within the large tau class, 45% of the members showed distinct selective expression under a specific treatment and/or in a specific tissue. Most of the duplicate GSTs created by the whole-genome duplication appear to have maintained similar expression patterns, while GSTs from genes created by tandem duplications have diverged rapidly. A similar pattern has been found in *Arabidopsis* (Casneuf et al., 2006; Ganko et al., 2007). A possible explanation for this pattern is that large-scale segmental duplication results in the duplication of multiple genes with their promoter and/or enhancer elements (Casneuf et al., 2006; Kim et al., 2006), while tandem duplication may disrupt the

regulatory regions of target genes, resulting in considerable variation in expression responses. Divergence in expression has been reported for various functional categories of genes. Indeed, >60% of rice gene families exhibit higher expression diversity between members than randomly selected gene pairs (Yim et al., 2009), and >50% of duplicated gene pairs formed by the most recent polyploidy event in *Arabidopsis* have divergent expression profiles (Blanc and Wolfe, 2004). These findings clearly support the assertion that expression divergence is often the first step in the functional divergence between duplicate genes and thereby increases the chance of duplicate genes being retained in a genome (Ohno, 1970).

Lineage-specific expansions and contractions of gene families have been shown to be driven by diversifying selection (Mondragon-Palomino et al., 2002; Shiu et al., 2004, 2006; Gingerich et al., 2007). Gene families involved in immune defense, stress responses, metabolism, cell signaling, chemoreception, and reproduction are well represented among those shown to have diversified under positive selection (Lespinet et al., 2002; Demuth and Hahn, 2009). However, evidence from

**Table 3.** Specific Activities of the *Populus* GSTs toward Seven Substrates (Means  $\pm$  SD Obtained from at Least Three Independent Determinations)

Cluster	Gene Pair	GST	Specific Activity ( $\mu\text{mol}/\text{min}$ per mg) to Each Substrate						
			CDNB	NBD-CI	DCNB	NBC	ECA	4-NPA	DHA
I	T	GSTU2	12.90 $\pm$ 0.04	6.93 $\pm$ 0.03	0.18 $\pm$ 0.09	1.51 $\pm$ 0.08	n.d.	nd	n.t.
I		GSTU8	6.02 $\pm$ 0.36	3.59 $\pm$ 0.02	0.01 $\pm$ 0.01	2.09 $\pm$ 1.40	n.d.	nd	n.t.
I	T	GSTU4	0.76 $\pm$ 0.02	n.d.	n.d.	n.d.	0.72 $\pm$ 0.04	0.01 $\pm$ 0.01	n.t.
I		GSTU9	3.92 $\pm$ 0.05	1.42 $\pm$ 0.06	0.09 $\pm$ 0.01	0.63 $\pm$ 0.34	n.d.	n.d.	n.t.
I	O	GSTU3	8.49 $\pm$ 0.31	6.14 $\pm$ 0.16	0.26 $\pm$ 0.03	n.d.	n.d.	n.d.	n.t.
–		GSTU54	20.99 $\pm$ 0.01	13.53 $\pm$ 0.13	n.d.	n.d.	n.d.	n.d.	n.t.
I		GSTU1	0.99 $\pm$ 0.01	0.08 $\pm$ 0.01	n.d.	0.25 $\pm$ 0.01	0.63 $\pm$ 0.03	n.d.	n.t.
I		GSTU7	n.d.	0.01 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	n.d.
II		GSTU19	0.04 $\pm$ 0.01	n.d.	n.d.	0.15 $\pm$ 0.02	n.d.	n.d.	n.t.
III	T	GSTU11	1.67 $\pm$ 0.01	1.70 $\pm$ 0.02	n.d.	n.d.	0.02 $\pm$ 0.01	n.d.	n.d.
III		GSTU12	7.38 $\pm$ 0.17	3.59 $\pm$ 0.06	0.12 $\pm$ 0.01	n.d.	n.d.	n.t.	n.d.
III	T	GSTU36	0.02 $\pm$ 0.01	n.d.	n.d.	n.d.	0.14 $\pm$ 0.03	n.d.	n.t.
III		GSTU38	n.d.	0.04 $\pm$ 0.01	n.d.	n.d.	0.77 $\pm$ 0.04	n.d.	n.t.
III		GSTU14	2.47 $\pm$ 0.04	1.38 $\pm$ 0.10	n.d.	n.d.	n.d.	n.d.	n.t.
IV	T	GSTU40	15.64 $\pm$ 0.16	7.30 $\pm$ 0.51	0.29 $\pm$ 0.03	0.26 $\pm$ 0.01	n.d.	n.d.	n.t.
–		GSTU55	3.00 $\pm$ 0.07	5.18 $\pm$ 0.02	n.d.	n.d.	0.15 $\pm$ 0.01	n.d.	n.t.
V	T	GSTU32	n.d.	0.03 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	n.t.
V		GSTU33	2.62 $\pm$ 0.01	3.14 $\pm$ 0.03	0.98 $\pm$ 0.04	1.61 $\pm$ 0.08	n.d.	n.d.	n.t.
V		GSTU30	15.63 $\pm$ 0.03	6.92 $\pm$ 0.02	0.22 $\pm$ 0.01	n.d.	n.d.	n.d.	n.t.
V		GSTU31	0.62 $\pm$ 0.01	1.46 $\pm$ 0.21	n.d.	0.88 $\pm$ 0.46	n.d.	n.d.	n.t.
VI	T	GSTF5	0.48 $\pm$ 0.04	1.79 $\pm$ 0.03	n.d.	0.63 $\pm$ 0.13	0.38 $\pm$ 0.08	n.d.	n.t.
VI		GSTF6	0.60 $\pm$ 0.02	12.95 $\pm$ 0.11	n.d.	n.d.	n.d.	n.d.	n.t.
–	O	GSTF3	0.32 $\pm$ 0.01	0.06 $\pm$ 0.01	n.d.	n.d.	0.58 $\pm$ 0.02	n.d.	n.d.
VI		GSTF7	0.15 $\pm$ 0.01	0.34 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	n.t.
VI		GSTF4	3.62 $\pm$ 0.02	1.10 $\pm$ 0.01	n.d.	2.94 $\pm$ 1.03	n.d.	n.d.	n.t.
VII	T	GSTU15	0.62 $\pm$ 0.01	0.63 $\pm$ 0.01	n.d.	1.38 $\pm$ 0.20	n.d.	n.d.	n.t.
VII		GSTU17	0.77 $\pm$ 0.01	0.86 $\pm$ 0.02	n.d.	n.d.	n.d.	n.d.	n.t.
VII		GSTU16	0.93 $\pm$ 0.02	0.49 $\pm$ 0.02	n.d.	0.10 $\pm$ 0.01	n.d.	n.d.	n.t.
–	W	GSTU18	0.01 $\pm$ 0.01	0.01 $\pm$ 0.01	0.03 $\pm$ 0.01	n.d.	n.d.	n.t.	n.t.
–		GSTU24	0.56 $\pm$ 0.01	0.05 $\pm$ 0.01	0.05 $\pm$ 0.01	n.d.	n.d.	n.d.	n.t.
–	O	GSTU22	28.05 $\pm$ 0.64	1.11 $\pm$ 0.38	0.11 $\pm$ 0.07	n.d.	n.d.	n.d.	n.t.
–		GSTU51	43.83 $\pm$ 1.23	5.24 $\pm$ 0.43	0.01 $\pm$ 0.01	4.34 $\pm$ 1.81	n.d.	0.02 $\pm$ 0.02	n.t.
–	W	GSTU26	0.05 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	n.d.	n.t.
–		GSTU46	n.d.	n.d.	n.d.	n.d.	0.06 $\pm$ 0.01	n.d.	n.t.
–	T	GSTF1	1.94 $\pm$ 0.02	10.18 $\pm$ 0.03	n.d.	4.49 $\pm$ 0.02	0.88 $\pm$ 0.02	n.d.	n.t.
–		GSTF2	0.75 $\pm$ 0.01	0.94 $\pm$ 0.03	n.d.	0.41 $\pm$ 0.01	0.13 $\pm$ 0.01	n.d.	n.t.
–	W	DHAR2	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	40.24 $\pm$ 0.12
–		DHAR3	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	16.79 $\pm$ 0.08
–		DHAR1	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	7.84 $\pm$ 0.14
–		GSTU35	8.01 $\pm$ 0.41	3.37 $\pm$ 0.06	0.14 $\pm$ 0.01	1.17 $\pm$ 0.74	n.d.	0.01 $\pm$ 0.01	n.t.
–		GSTL1	n.d.	0.03 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	n.d.
–		GSTL2	n.d.	0.03 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	0.01 $\pm$ 0.01
–		EF1B $\gamma$ 1	0.03 $\pm$ 0.01	n.d.	0.13 $\pm$ 0.04	n.d.	n.d.	n.d.	0.01 $\pm$ 0.01
–		GSTT2	0.04 $\pm$ 0.01	0.02 $\pm$ 0.01	n.d.	n.d.	n.d.	n.d.	0.08 $\pm$ 0.04

Dispersed GSTs not grouped into any cluster are indicated with “–”; gene pairs created by tandem duplication, whole-genome duplication, and other events are indicated as T, W, and O, respectively. n.d., no activity detected; n.t., not detected.

genome-wide nucleotide substitution analyses suggests that retention of paralogs driven by positive selection accounts for only a fraction of the expansion of gene families (Bergthorsson et al., 2007; Demuth and Hahn, 2009). In this study, we found evidence that purifying selection has predominated across the tau and phi GST classes, with episodic positive selection. A theory that may partly explain why many duplicate genes bear the signature of continued purifying selection after duplication is that deleterious mutations may occur in different domains in

duplicates of genes with multiple independent domain subfunctions, allowing the conservation of both duplicates since they retain different permutations of subfunctions (Force et al., 1999). Purifying selection against deleterious loss-of-function mutations increases the fixation probability of a new duplicate gene and enhances the preservation of functional alleles at both duplicate loci (Tanaka et al., 2009). Mutations that affect modular domains or molecular surfaces that control distinct subfunctions have the potential to alter one aspect of a protein's function

**Table 4.** Steady State Kinetic Constants of the *Populus* GSTs for CDNB, NBD-Cl, and DHA Conjugation Reactions (Means  $\pm$  SD Obtained from at Least Three Independent Determinations)

Cluster	Gene pair	GSTs	$K_m^{\text{GSH}}$ (mM)	$k_{\text{cat}}^{\text{GSH}}$ ( $\text{s}^{-1}$ )	$(k_{\text{cat}}/K_m)^{\text{GSH}}$ ( $\text{mM}^{-1} \text{s}^{-1}$ )	$K_m^{\text{CDNB}}$ (mM)	$k_{\text{cat}}^{\text{CDNB}}$ ( $\text{s}^{-1}$ )	$(k_{\text{cat}}/K_m)^{\text{CDNB}}$ ( $\text{mM}^{-1} \text{s}^{-1}$ )
I	T	GSTU2	0.33 $\pm$ 0.02	130.85	396.52	0.22 $\pm$ 0.01	231.92	1054.18
I		GSTU8	0.08 $\pm$ 0.01	23.23	290.37	4.29 $\pm$ 1.06	124.93	29.12
I	T	GSTU4	4.16 $\pm$ 0.77	9.47	2.27	8.75 $\pm$ 2.75	18.87	2.16
I		GSTU9	0.30 $\pm$ 0.03	14.65	48.83	0.59 $\pm$ 0.03	18.46	31.29
I	O	GSTU3	0.25 $\pm$ 0.03	127.29	509.16	0.29 $\pm$ 0.22	152.04	524.27
–		GSTU54	0.48 $\pm$ 0.01	490.55	1021.98	0.24 $\pm$ 0.01	412.24	1717.67
I		GSTU1	0.17 $\pm$ 0.06	1.76	10.35	0.21 $\pm$ 0.01	2.26	10.76
III	T	GSTU11	0.17 $\pm$ 0.01	1.88	11.06	123.6 $\pm$ 6.18	174.36	1.41
III		GSTU12	0.20 $\pm$ 0.01	36.96	184.80	18.76 $\pm$ 9.14	562.23	29.97
III		GSTU14	0.16 $\pm$ 0.01	66.72	417.00	12.57 $\pm$ 4.53	269.09	21.41
V		GSTU30	0.72 $\pm$ 0.01	543.56	754.94	0.21 $\pm$ 0.01	341.21	1624.81
VII	T	GSTU15	0.35 $\pm$ 0.05	0.48	1.37	3.39 $\pm$ 0.42	2.16	0.64
VII		GSTU17	0.27 $\pm$ 0.01	0.82	3.04	9.31 $\pm$ 3.49	6.89	0.74
VII		GSTU16	2.45 $\pm$ 0.57	2.64	1.08	2.49 $\pm$ 0.74	2.94	1.18
–	O	GSTU22	0.56 $\pm$ 0.10	1118.39	1997.12	1.72 $\pm$ 0.22	574.6	334.07
–		GSTU51	0.43 $\pm$ 0.03	710.15	1651.51	0.15 $\pm$ 0.02	622.93	4152.87
–	O	GSTU40	0.18 $\pm$ 0.02	112.42	624.56	3.08 $\pm$ 0.70	486.21	157.86
–		GSTU55	0.21 $\pm$ 0.02	3.61	17.19	1.82 $\pm$ 0.16	7.89	4.33
–		GSTU24	0.33 $\pm$ 0.02	0.54	1.64	3.90 $\pm$ 0.32	1.81	0.46
–		GSTU35	0.45 $\pm$ 0.02	48.311	107.35	0.41 $\pm$ 0.03	33.39	81.44
			$K_m^{\text{GSH}}$ (mM)	$k_{\text{cat}}^{\text{GSH}}$ ( $\text{s}^{-1}$ )	$(k_{\text{cat}}/K_m)^{\text{GSH}}$ ( $\text{mM}^{-1} \text{s}^{-1}$ )	$K_m^{\text{NBD-Cl}}$ (mM)	$k_{\text{cat}}^{\text{NBD-Cl}}$ ( $\text{s}^{-1}$ )	$(k_{\text{cat}}/K_m)^{\text{NBD-Cl}}$ ( $\text{mM}^{-1} \text{s}^{-1}$ )
VI	T	GSTF5	0.09 $\pm$ 0.01	25.27	280.78	0.62 $\pm$ 0.29	14.64	23.61
VI		GSTF6	0.10 $\pm$ 0.01	147.73	1448.31	1.62 $\pm$ 0.32	1383.73	854.15
VI		GSTF4	2.02 $\pm$ 0.56	16.83	8.33	2.10 $\pm$ 0.21	12.64	6.02
VI		GSTF7	0.71 $\pm$ 0.14	0.84	1.18	0.43 $\pm$ 0.07	1.89	4.39
–	T	GSTF1	0.51 $\pm$ 0.07	246.54	483.42	0.32 $\pm$ 0.03	435.60	1361.25
–		GSTF2	0.82 $\pm$ 0.04	1.82	2.22	0.31 $\pm$ 0.03	2.70	8.71
			$K_m^{\text{GSH}}$ (mM)	$k_{\text{cat}}^{\text{GSH}}$ ( $\text{s}^{-1}$ )	$(k_{\text{cat}}/K_m)^{\text{GSH}}$ ( $\text{mM}^{-1} \text{s}^{-1}$ )	$K_m^{\text{DHA}}$ (mM)	$k_{\text{cat}}^{\text{DHA}}$ ( $\text{s}^{-1}$ )	$(k_{\text{cat}}/K_m)^{\text{DHA}}$ ( $\text{mM}^{-1} \text{s}^{-1}$ )
–	W	DHAR2	1.39 $\pm$ 0.28	10349.44	7445.64	0.45 $\pm$ 0.09	8698.75	19330.56
–		DHAR3	2.94 $\pm$ 0.19	5300.45	1802.87	0.16 $\pm$ 0.04	1265.55	7909.69
–		DHAR1	1.72 $\pm$ 0.51	172.43	100.25	0.30 $\pm$ 0.01	115.05	383.50

Dispersed GSTs not grouped into any cluster are indicated with “–”; gene pairs created by tandem duplication, whole-genome duplication, and other events are indicated as T, W, and O, respectively.

without disrupting its interactions with at least some molecular partners. It is therefore likely that purifying selection would be partially relaxed after duplication (Bridgham et al., 2008).

Plant GSTs perform diverse catalytic and noncatalytic functions in the detoxification of xenobiotics, prevention of oxidative damage, and endogenous metabolism (Frova, 2003; Basantani and Srivastava, 2007), some of which are fulfilled by conjugating electrophilic substrates to glutathione. Most GSTs are active as dimers, composed of either homogeneous (the most prevalent form) or heterogeneous subunits (Dixon et al., 1999; Edwards et al., 2000). Both within and among the GST classes, the relatively small thioredoxin-like N-terminal domain that binds to GSH is conserved in all classes (Dixon et al., 2002). By contrast, the C-terminal domain that provides structural elements for the recognition of xenobiotic substrates harbors much more diversity within and among classes (Edwards et al., 2000; Basantani and Srivastava, 2007). Among *Populus* tau GSTs, the C-terminal domain appears to be under more relaxed functional constraints than the N-terminal domain, which could lead to diversification in

substrate selectivity and specificity among the members, while preserving the enzymes' primary function. Given the role of the C-terminal domain in substrate recognition, the overrepresentation of putative positively selected sites in the tau class suggests pressure for diversification, possibly linked to selective advantages conferred by the ability to recognize evolving targets and enhance the metabolism of substances encountered in the environment. In addition, our study revealed that a subset of the *Populus* tau GSTs appears to have evolved significantly faster than the others, suggesting that the substrate recognition module in this subset is diversifying in response to a set of changing substrates, while the other subset recognizes targets defined by the ancestral functions of the enzyme. This pattern of gene family diversification has been previously observed in proteins that function as substrate recognition factors (Gingerich et al., 2007).

The potential of a gene to evolve new function upon duplication may also depend on its ability to accept mutations without losing thermodynamic stability of the protein domain that it



encodes. This protein structure-constrained potential has an impact, at least in part, on the sequence variability and functional diversity and the size of a gene family (Shakhnovich et al., 2005). Proteins often display a certain degree of structural flexibility that allows conformational changes, and it is widely assumed that protein functional plasticity correlates with structural flexibility (Hou et al., 2007; Kobilka and Deupi, 2007). Mutations in the core structures can disrupt catalytic functions or dramatically alter the shape of the binding pocket, while mutations outside the active sites that subtly alter the substrate binding conformation might be more effective for achieving functional flexibility. Structure modeling of tau and phi GSTs illustrated that divergence is particularly visible in the C-terminal domain. Interestingly, 7 out of 10 of the putative positively selected sites are in the loop regions of this domain. The mutations accumulated in loop regions may not have a dramatic effect on substrate binding, but they could result in subtle conformational changes to enzyme structure and, thus, substrate selectivity and kinetics, resulting in a broad range of biochemical properties among the gene family members.

How genes and their functions evolve after duplication is a central, long-standing question in evolutionary biology. One limitation for understanding gene family evolution is a lack of understanding of the mode and tempo of functional diversification in different functional categories of genes. As we gain insights into divergence in gene sequences, structures, and functions, the patterns of evolutionary dynamics of gene families are just beginning to emerge. Our findings provide evidence for the evolutionary partitioning of ancestral functions among duplicated genes and protein domains, accompanied by specialization and only partially overlapping enzymatic properties. Both purifying and directional selection played a role in the observed functional diversification. This mechanism could facilitate the retention of the duplicate genes and result in a large gene family that has a broad substrate spectrum and a wide range of reactivity toward different substrates. Thus, when gene families generally, and the GST family in particular, have functions that are potentially subspecialized, and the structure of the proteins they encode allows rapid changes in specificity, affinity, and activity, they are likely to expand in response to environmental shifts at various time scales.

## METHODS

### Genomic Data Mining, GST Gene Identification, and Nomenclature

To identify GST genes in *Populus trichocarpa*, TBLASTN searches of the *P. trichocarpa* genome database were performed using 53 full-length GST protein sequences of *Arabidopsis thaliana* (Dixon et al., 2002), 61 of rice (*Oryza sativa*; Soranzo et al., 2004), and 575 of other plants, animals, fungi, and bacteria (see Supplemental Data Set 1 online). These 689 full-length GSTs represent 35 classes defined by the NCBI conserved domain database (Marchler-Bauer et al., 2005). Analysis of the collected *P. trichocarpa* GST candidates indicated that some sequences were partially misannotated during the automated genome annotation process. Thus, manual reannotation was performed to rectify incorrect start codon predictions, splicing errors, missed or extra exons, fused genes, split genes, and incorrectly predicted pseudogenes. The reannotated sequences were further analyzed using an NCBI conserved domain search

(<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) to confirm the presence of typical GST N- and C-terminal domains in their protein structure. The predicted GST genes were then amplified from genomic DNA and mRNA, cloned into the pGEM-T Easy Vector (Promega), and sequenced in both directions to verify the gene sequence and structure. For genes that RT-PCR did not detect (nine out of 81 in this study), their structure was assumed to be identical to that of their closest relative on the phylogenetic tree; this approach was adapted from other studies (Meyers et al., 2003). Noncoding exons in the 5' and 3' untranslated regions were not considered in this study. Primers used in gene amplifications are listed in Supplemental Table 3 online.

The nomenclature for *Populus* GSTs followed the system suggested by Edwards et al. (2000) for plant GSTs; a univocal name was assigned to each *Populus* GST gene (see Supplemental Table 1 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 and Molecular Evolution Analyses

GST protein sequences were aligned using a BLOSUM 30 matrix by the Clustal X 1.83 program (Thompson et al., 1997), with an open gap penalty of 10 and an extend gap penalty of 0.1 in pairwise alignments, an extend gap penalty of 0.05 in the multiple alignment, and a delay divergent setting of 40%. The protein alignment was further adjusted manually using BioEdit (Hall, 1999). Phylogenetic relationships among the *Populus* GSTs were reconstructed using an ML procedure by PHYML (Guindon and Gascuel, 2003) with the JTT (Jones, Taylor, and Thornton) amino acid substitution model. Cytosolic GSTs are regarded as having been derived from the GRX2 protein (Holm et al., 2006); thus, GRX2 was used as an outgroup in the phylogenetic analysis of the whole GST family. For analysis of each GST class, members of the sister class were used as an outgroup. One thousand bootstrap replicates were performed in each analysis to obtain the confidence support. The synonymous ( $k_s$ ) and nonsynonymous substitution ( $k_a$ ) rates between paralogous gene pairs were calculated by the K-Estimator program (Comeron, 1999).

The  $\omega$  values ( $\omega = d_N/d_S$ ) among all pairwise comparisons within each of the tau and phi GST classes were calculated by the YN00 program in the PAML 4.3 package (Yang, 2007). The calculations were performed separately using the N-terminal domain (from the start codon to alignment position 83 in Supplemental Figure 2 for the tau GSTs and to alignment position 79 in Supplemental Figure 3 for the phi GSTs), and the C-terminal domain sequences (from alignment position 94 to the end in Supplemental Figure 2 for the tau GSTs and from alignment position 96 to the end in Supplemental Figure 3 for the phi GSTs). Two-sample  $t$  tests were performed to determine whether the  $\omega$  values of the two domains are significantly different. To evaluate variation in selective pressure over a phylogeny, the branch models of CODEML, PAML, were used to estimate  $\omega$  under different assumptions. Analyses were conducted under three a priori assumptions: a one-ratio model in which one  $\omega$  value was assumed for the entire tree, a free-ratio model that allows  $\omega$  to vary over all branches of the tree, and a two-ratio model for the tau class, in which  $\omega$  values were allowed to vary between the two major clades of the class (see Supplemental Figure 1 online), by selecting each clade as foreground, respectively. To determine whether positive selection had acted at specific sites in the GST sequences, six models in the PAML package were explored: the one-ratio model (M0), the discrete model (M3), the nearly neutral model (M1a), the positive-selection model (M2a), the beta model (M7), and the beta and  $\omega$  model (M8). To evaluate rate variation between the two major clades of tau class (see Supplemental Figure 1 online), a local clock model was compared with a global clock model using PAML. Two different rates were assigned to each clade in the local clock model, while one the same rate was assumed for both clades in the global clock model.

To verify which of the models best fitted the data, LRTs were performed by comparing twice the difference in log likelihood values between pairs of the models using a  $\chi^2$  distribution, with the degrees of freedom equal to the differences in the number of parameters between the models (Yang et al., 2000).

### Homology Modeling

The crystal structures of *Glycine max* Gm GSTU4, wheat (*Triticum aestivum*) Ta GSTU4, and rice Os GSTU1 (Protein Data Bank code numbers 2VO4, 1GWC, and 1OYJ, respectively) were used as templates for constructing the structure models of *Populus* tau GSTs. The crystal structures of two *Zea mays* phi GSTs (PDB: 1AW9 and 1BYE) and an *Arabidopsis* phi GST (PDB: 1GNW) were used as templates for structure modeling of *Populus* phi GSTs. Sequences were aligned by the Align 2D structure alignment program (homology module, InsightII; Accelrys). Structures were automatically built by the MODELER module of InsightII. MODELER uses a spatial restraint method to build a three-dimensional image of protein structure and is capable of generating a reliable predicted structure using probability density functions derived from homologous structures and general features of known proteins (Fiser and Sali, 2003). All optimized structures were evaluated by the Profile-3D program of InsightII to select the best model for a given protein.

### Expression of GST Genes under Different Treatments

To investigate the expression patterns of GSTs under normal growth conditions and abiotic stress, cuttings of *P. trichocarpa* were cultivated in water for ~2 months and then three chemical treatments were applied: 1.0 mM CDNB, 5.0% H<sub>2</sub>O<sub>2</sub>, and 1.5% atrazine as cultivation solutions and sprays. Cultivation in water was used as a control. Each treatment consisted of five replicates. Twelve hours after the chemical treatments, total RNA was isolated from leaf, shoot, bud, phloem, and root tissues using an Aurum Total RNA Kit (Bio-Rad Laboratories). Total RNA was treated with RNase-free DNase I (Promega) and reverse transcribed into cDNA using an RNA PCR Kit (AMV) version 3.0 (TaKaRa). Based on the multiple sequences alignment of all *Populus* GST sequences, 80 specific primer pairs were designed (see Supplemental Table 4 online). In the RT-PCR analysis, the *Populus Actin* gene (GenBank number XM\_002316253) was used as an internal control. After RT-PCR, PCR products from each sample were validated by DNA sequencing. Quantitative RT-PCR (qRT-PCR) was performed using an Mx3000P real-time PCR system (Stratagene). In qRT-PCR analysis, the reverse transcription products were used as templates, and Brilliant SYBR Green QPCR Master Mix (Stratagene) was used in all qPCR reactions. The *Populus Actin* gene (GenBank number XM\_002316253) was used as an internal control in the qRT-PCR analyses, with the forward primer 5'-GAGACCTT-CAAGACTCCTGCTATG-3' and reverse primer 5'-CAGGTCAGACGAA-GAATGGC-3'. Specific qRT-PCR primers for *Populus* GSTs are listed in Supplemental Table 4 online. The qPCR conditions, following optimization, consisted of an initial denaturation step of 10 min at 95°C followed by 40 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s. A melt-curve analysis immediately followed the final amplification to observe the melting characteristics of each amplicon and confirm that a specific product had been amplified, consisting of incubation at 95°C for 60 s, cooling to 60°C for 30 s, and a slow rise in temperature to 95°C with continuous measurement of the decline in fluorescence. The relative expression of specific genes was quantified by  $2^{-\Delta\Delta Ct}$ , where  $\Delta Ct$  is the difference in threshold cycles between the target and housekeeping gene *Actin*, and  $\Delta\Delta Ct$  is the difference between the  $\Delta Ct$  of the samples exposed to abiotic stress and the  $\Delta Ct$  of the control plants. The mean threshold cycle values for each GST were obtained from three independent PCR experiments.

### Molecular Cloning and Purification of Recombinant *Populus* GST Proteins

To investigate the enzymatic functions of *Populus* GSTs, 30 tau, seven phi, three DHAR, two lambda, one theta, and one EF1B $\gamma$  GST were selected for protein expression analysis and purification (Table 3). These GSTs represent different evolutionary relationships and histories. Except for EF1B $\gamma$ 1, the full-length cDNA of each GST was subcloned into a pET30a expression vector (Novagen), which provides the correct reading frame and a 6 $\times$ His-tag at the N terminus. EF1B $\gamma$ 1 consists of two parts: a GST domain and an EF1B $\gamma$  domain (Figure 1C). In this study, only the GST domain of EF1B $\gamma$ 1 was used to construct the expression vector. Primers used to construct GST expression vectors are listed in Supplemental Table 5 online. Colonies containing the appropriate insert were identified by sequencing.

Overnight cultures of *E. coli* BL21, transformed with each GST gene expression plasmid, were diluted 1:100 and grown until the optical density ( $A_{600}$ ) reached 0.5. isopropyl- $\beta$ -D-thiogalactopyranoside was then added to each culture at a final concentration of 0.1 mM, and the cultures were incubated at 37°C overnight. The bacteria were harvested by centrifugation at 8000g for 3 min at 4°C, resuspended in binding buffer (20 mM sodium phosphate, 0.5 M NaCl, and 20 mM imidazole, pH 7.4), and disrupted by cold sonication. In each case, the homogenate was then subjected to centrifugation at 10,000g for 10 min at 4°C. The resultant particulate material and a small portion of the supernatant were analyzed by SDS-PAGE. The rest of the supernatant was loaded onto a Ni Sepharose High Performance column (GE Healthcare Bio-Sciences) that had been preequilibrated with binding buffer. The overexpressed protein that bound to the Ni Sepharose High Performance column was eluted with elution buffer (20 mM sodium phosphate, 0.5 M NaCl, and 500 mM imidazole, pH 7.4). The purified recombinant protein was desalted using a PD-10 column (GE Healthcare Bio-Sciences) in 10 mM Tris-HCl buffer, pH 7.5. In this study, six GSTs (GSTU7, 18, and 32, GSTL1 and 2, and GSTT2) were expressed as inclusion bodies. Inclusion bodies were dissolved in 10 mM Tris-HCl, pH 8.0, 10 mM DTT, and 8 M urea and then refolded by serial dilutions in 10 mM Tris-HCl, pH 8.0, and 10 mM DTT. The refolded proteins were further purified by Ni Sepharose High Performance columns.

### Specific Activity and Kinetics of GST Enzymes

GST activities (at 25°C) toward the substrates CDNB, DCNB, ECA, NBC, and 4-NPA were measured using the method described by Habig et al. (1974), while activity toward NBD-Cl was measured using the method described by Ricci et al. (1994), and activity toward DHA was measured as described by Edwards and Dixon (2005). Protein concentrations in the enzyme preparations were determined by measuring the absorbance at 280 nm. The apparent  $K_m$  values for GSH were determined using GSH concentrations ranging from 0.02 to 1.0 mM and a fixed CDNB, NBD-Cl, or DHA concentration of 1.0 mM. The apparent  $K_m$  values for CDNB, NBD-Cl, or DHA were determined using concentrations of CDNB, NBD-Cl, or DHA, respectively, ranging from 0.04 to 1.0 mM and a fixed GSH concentration of 1.0 mM. The kinetic parameters were derived from nonlinear regression analysis by the Hyper32 program available at <http://www.liv.ac.uk/~jse/software.html>.

### Accession Numbers

Sequence data from this article can be found in the Arabidopsis Genome Initiative or GenBank/EMBL databases under the accession numbers listed in Supplemental Tables 1 and 6 online for *P. trichocarpa* and *Arabidopsis*, respectively. Other sequences used are *Populus Actin* gene (GenBank number XM\_002316253), two *Pinus taeda* lambda GSTs (GenBank numbers CV034086 and DR019281), *Picea glauca* lambda



GST (GenBank number EX306134), and GRX2 (GenBank number NP\_287198).

### Supplemental Data

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

**Supplemental Figure 1.** Phylogenetic Tree of the 58 Full-Length *Populus* Tau GST Proteins.

**Supplemental Figure 2.** Sequence Alignment of *Populus* Tau GSTs and the Predicted Secondary Structure Elements.

**Supplemental Figure 3.** Sequence Alignment of *Populus* Phi GSTs and the Predicted Secondary Structure Elements.

**Supplemental Table 1.** Full-Length GST Genes Identified from the *Populus trichocarpa* Genome.

**Supplemental Table 2.** GST Fragments Identified from the *Populus trichocarpa* Genome.

**Supplemental Table 3.** Primers Used to Amplify *Populus* GST Genes.

**Supplemental Table 4.** RT-PCR Primers Used to Detect the Expression of *Populus* GST Genes.

**Supplemental Table 5.** Primers Used to Construct the *Populus* GST Protein Expression Vector.

**Supplemental Table 6.** Tau GSTs Identified from *Arabidopsis thaliana* Genome.

**Supplemental Data Set 1.** GSTs Used in the Homology Search of *Populus* GST Candidates.

**Supplemental Data Set 2A.** Text File of Unmasked Alignment Corresponding to the Phylogenetic Tree in Figure 1.

**Supplemental Data Set 2B.** Text File of Masked Alignment Corresponding to the Phylogenetic Tree in Figure 1.

**Supplemental Data Set 3A.** Text File of Unmasked Alignment Corresponding to the Phylogenetic Tree in Figure 3A.

**Supplemental Data Set 3B.** Text File of Masked Alignment Corresponding to the Phylogenetic Tree in Figure 3A.

**Supplemental Data Set 4A.** Text File of Unmasked Alignment Corresponding to the Phylogenetic Tree in Figure 3B.

**Supplemental Data Set 4B.** Text File of Masked Alignment Corresponding to the Phylogenetic Tree in Figure 3B.

**Supplemental Data Set 5A.** Text File of Unmasked Alignment Corresponding to the Phylogenetic Tree in Figure 5.

**Supplemental Data Set 5B.** Text File of Masked Alignment Corresponding to the Phylogenetic Tree in Figure 5.

**Supplemental Data Set 6A.** Text File of Unmasked Alignment Corresponding to the Phylogenetic Tree in Supplemental Figure 1.

**Supplemental Data Set 6B.** Text File of Masked Alignment Corresponding to the Phylogenetic Tree in Supplemental Figure 1.

### ACKNOWLEDGMENTS

We thank Bo Zheng of Umeå Plant Science Centre, Umeå University for providing plant material and Song Ge, Xiao-Quan Wang, and Daming Zhang of the Institute of Botany, Chinese Academy of Sciences, Tao Sang of Michigan State University, Hong Ma of Pennsylvania State University, and Folmer Bokma of Umeå University for valuable comments on an earlier draft of this work. This study was supported by grants from the Natural Science Foundation of China (NSFC 30830010

and 30770149), the National Basic Research Program of China (2009CB119104), and the Chinese Academy of Sciences (KZCX2-YW-414).

Received July 24, 2009; revised October 31, 2009; accepted November 16, 2009; published December 8, 2009.

### REFERENCES

- Agrawal, G.K., Jwa, N.S., and Rakwal, R.** (2002). A pathogen-induced novel rice (*Oryza sativa* L.) gene encodes a putative protein homologous to type II glutathione S-transferases. *Plant Sci.* **163**: 1153–1160.
- Arabidopsis Genome Initiative** (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.
- Barakat, A., Bagniewska-Zadworna, A., Choi, A., Plakkat, U., DiLoreto, D.S., Yellanki, P., and Carlson, J.E.** (2009). The cinnamyl alcohol dehydrogenase gene family in *Populus*: Phylogeny, organization, and expression. *BMC Plant Biol.* **9**: 26.
- Basantani, M., and Srivastava, A.** (2007). Plant glutathione transferases - A decade falls short. *Can. J. Bot.* **85**: 443–456.
- Bergthorsson, U., Andersson, D.I., and Roth, J.R.** (2007). Ohno's dilemma: Evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. USA* **104**: 17004–17009.
- Blanc, G., and Wolfe, K.H.** (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**: 1679–1691.
- Bridgham, J.T., Brown, J.E., Rodriguez-Mari, A., Catchen, J.M., and Thornton, J.W.** (2008). Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet.* **4**: e1000191.
- Casneuf, T., De Bodt, S., Raes, J., Maere, S., and Van de Peer, Y.** (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol.* **7**: R13.
- Cameron, J.M.** (1999). K-Estimator: Calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15**: 763–764.
- Dean, E.J., Davis, J.C., Davis, R.W., and Petrov, D.A.** (2008). Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet.* **4**: e1000113.
- Demuth, J.P., and Hahn, M.W.** (2009). The life and death of gene families. *Bioessays* **31**: 29–39.
- Dixon, D.P., Cole, D.J., and Edwards, R.** (1999). Dimerisation of maize glutathione transferases in recombinant bacteria. *Plant Mol. Biol.* **40**: 997–1008.
- Dixon, D.P., Laphorn, A., and Edwards, R.** (2002). Plant glutathione transferases. *Genome Biol.* **3**: reviews 3004.3001–3004.3010.
- Edwards, R., and Dixon, D.P.** (2005). Plant glutathione transferases. *Methods Enzymol.* **401**: 169–186.
- Edwards, R., Dixon, D.P., and Walbot, V.** (2000). Plant glutathione S-transferases: Enzymes with multiple functions in sickness and in health. *Trends Plant Sci.* **5**: 193–198.
- Fiser, A., and Sali, A.** (2003). Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol.* **374**: 461–491.
- 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.
- Frova, C.** (2003). The plant glutathione transferase gene family: Genomic structure, functions, expression and evolution. *Physiol. Plant.* **119**: 469–479.

- Ganko, E.W., Meyers, B.C., and Vision, T.J.** (2007). Divergence in expression between duplicated genes in *Arabidopsis*. *Mol. Biol. Evol.* **24**: 2298–2309.
- Gingerich, D.J., Hanada, K., Shiu, S.H., and Vierstra, R.D.** (2007). Large-scale, lineage-specific expansion of a bric-a-brac/tramtrack/broad complex ubiquitin-ligase gene family in rice. *Plant Cell* **19**: 2329–2348.
- Guindon, S., and Gascuel, O.** (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**: 696–704.
- Habig, W.H., Pabst, M.J., and Jakoby, W.B.** (1974). Glutathione S-transferases. The first enzymatic step in mercapturic acid formation. *J. Biol. Chem.* **249**: 7130–7139.
- Hall, T.A.** (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**: 95–98.
- Hancock, J.M.** (2005). Gene factories, microfunctionalization and the evolution of gene families. *Trends Genet.* **21**: 591–595.
- Holm, P.J., Bhakat, P., Jegerschold, C., Gyobu, N., Mitsuoka, K., Fujiyoshi, Y., Morgenstern, R., and Hebert, H.** (2006). Structural basis for detoxification and oxidative stress protection in membranes. *J. Mol. Biol.* **360**: 934–945.
- Hou, L., Honaker, M.T., Shireman, L.M., Balogh, L.M., Roberts, A.G., Ng, K.-c., Nath, A., and Atkins, W.M.** (2007). Functional promiscuity correlates with conformational heterogeneity in A-class glutathione S-transferases. *J. Biol. Chem.* **282**: 23264–23274.
- Hughes, A.L.** (1994). The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256**: 119–124.
- Jeppesen, M.G., Ortiz, P., Shepard, W., Kinzy, T.G., Nyborg, J., and Andersen, G.R.** (2003). The crystal structure of the glutathione S-transferase-like domain of elongation factor 1B gamma from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **278**: 47190–47198.
- Kampranis, S.C., Damianova, R., Atallah, M., Toby, G., Kondi, G., Tschlis, P.N., and Makris, A.M.** (2000). A novel plant glutathione S-transferase/peroxidase suppresses Bax lethality in yeast. *J. Biol. Chem.* **275**: 29207–29216.
- Kim, J., Shiu, S.H., Thoma, S., Li, W.H., and Patterson, S.E.** (2006). Patterns of expansion and expression divergence in the plant polygalacturonase gene family. *Genome Biol.* **7**: R87.
- Kobilka, B.K., and Deupi, X.** (2007). Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol. Sci.* **28**: 397–406.
- Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L.** (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12**: 1048–1059.
- Loyall, L., Uchida, K., Braun, S., Furuya, M., and Frohnmeyer, H.** (2000). Glutathione and a UV light-induced glutathione S-transferase are involved in signaling to chalcone synthase in cell cultures. *Plant Cell* **12**: 1939–1950.
- Lynch, M.** (2007). *The Origins of Genome Architecture*. (Sunderland, MA: Sinauer Associates).
- Lynch, M., and Conery, J.S.** (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.
- Lysak, M.A., Koch, M.A., Beaulieu, J.M., Meister, A., and Leitch, I.J.** (2009). The dynamic ups and downs of genome size evolution in *Brassicaceae*. *Mol. Biol. Evol.* **26**: 85–98.
- Marchler-Bauer, A., et al.** (2005). CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res.* **33**: D192–D196.
- Matsumura, Y., Iwakawa, H., Machida, Y., and Machida, C.** (2009). Characterization of genes in the *ASYMMETRIC LEAVES2/LATERAL ORGAN BOUNDARIES (AS2/LOB)* family in *Arabidopsis thaliana*, and functional and molecular comparisons between AS2 and other family members. *Plant J.* **58**: 525–537.
- Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W.** (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15**: 809–834.
- Mondragon-Palomino, M., Meyers, B.C., Michelmore, R.W., and Gaut, B.S.** (2002). Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* **12**: 1305–1315.
- Moore, R.C., and Purugganan, M.D.** (2005). The evolutionary dynamics of plant duplicate genes. *Curr. Opin. Plant Biol.* **8**: 122–128.
- Mueller, L.A., Goodman, C.D., Silady, R.A., and Walbot, V.** (2000). AN9, a petunia glutathione S-transferase required for anthocyanin sequestration, is a flavonoid-binding protein. *Plant Physiol.* **123**: 1561–1570.
- Oakley, A.J.** (2005). Glutathione transferases: New functions. *Curr. Opin. Struct. Biol.* **15**: 716–723.
- Ohno, S.** (1970). *Evolution by Gene Duplication*. (Heidelberg, Germany: Springer-Verlag).
- Ricci, G., Caccuri, A.M., Lo Bello, M., Pastore, A., Piemonte, F., and 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.
- Shakhnovich, B.E., Deeds, E., Delisi, C., and Shakhnovich, E.** (2005). Protein structure and evolutionary history determine sequence space topology. *Genome Res.* **15**: 385–392.
- Shiu, S.H., Byrnes, J.K., Pan, R., Zhang, P., and Li, W.H.** (2006). Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc. Natl. Acad. Sci. USA* **103**: 2232–2236.
- Shiu, S.H., Karlowski, W.M., Pan, R., Tzeng, Y.H., Mayer, K.F., and Li, W.H.** (2004). Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* **16**: 1220–1234.
- Smith, A.P., DeRidder, B.P., Guo, W.J., Seeley, E.H., Regnier, F.E., and Goldsbrough, P.B.** (2004). Proteomic analysis of *Arabidopsis* glutathione S-transferases from benoxacor- and copper-treated seedlings. *J. Biol. Chem.* **279**: 26098–26104.
- Soranzo, N., Sari Gorla, M., Mizzi, L., De Toma, G., and Frova, C.** (2004). Organisation and structural evolution of the rice glutathione S-transferase gene family. *Mol. Genet. Genomics* **271**: 511–521.
- Tanaka, K.M., Takahasi, K.R., and Takano-Shimizu, T.** (2009). Enhanced fixation and preservation of a newly arisen duplicate gene by masking deleterious loss-of-function mutations. *Genet. Res.* **91**: 267–280.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G.** (1997). The Clustal X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24**: 4876–4882.
- Tuskan, G.A., et al.** (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.
- Walsh, B.** (2003). Population-genetic models of the fates of duplicate genes. *Genetica* **118**: 279–294.
- Yang, Z.** (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**: 1586–1591.
- Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M.** (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.
- Yang, Z., Wong, W.S., and Nielsen, R.** (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**: 1107–1118.
- Yim, W.C., Lee, B.M., and Jang, C.S.** (2009). Expression diversity and evolutionary dynamics of rice duplicate genes. *Mol. Genet. Genomics* **281**: 483–493.
- Zhang, W., Bone, J.R., Edmondson, D.G., Turner, B.M., and Roth, S.Y.** (1998). Essential and redundant functions of histone acetylation revealed by mutation of target lysines and loss of the Gcn5p acetyltransferase. *EMBO J.* **17**: 3155–3167.