Heredity (2011) 106, 633–641
© 2011 Macmillan Publishers Limited All rights reserved 0018-067X/11
www.nature.com/hdy

ORIGINAL ARTICLE

Molecular evolution of *glycinin* and *β-conglycinin* gene families in soybean (*Glycine max* L. Merr.)

C Li and Y-M Zhang
State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University, Nanjing, PR China

There are two main classes of multi-subunit seed storage proteins, glycinin (11S) and β-conglycinin (7S), which account for approximately 70% of the total protein in a typical soybean seed. The subunits of these two protein classes are encoded by a number of genes. The genomic organization of these genes follows a complex evolutionary history. This research was designed to describe the origin and maintenance of genes in each of these gene families by analyzing the synteny, phylogenies, selection pressure and duplications of the genes in each gene family. The ancestral *glycinin* gene initially experienced a tandem duplication event; then, the genome underwent two subsequent rounds of whole-genome duplication, thereby resulting in duplication of the *glycinin* genes, and finally a tandem duplication likely gave rise to the *Gy1* and *Gy2* genes. The *β-conglycinin* genes primarily originated through the more recent whole-genome duplication and several tandem duplications. Purifying selection has had a key role in the maintenance of genes in both gene families. In addition, positive selection in the *glycinin* genes and a large deletion in a *β-conglycinin* exon contribute to the diversity of the duplicate genes. In summary, our results suggest that the duplicated genes in both gene families prefer to retain similar function throughout evolution and therefore may contribute to phenotypic robustness.
Heredity (2011) **106,** 633–641; doi:10.1038/hdy.2010.97; published online 28 July 2010

Keywords: β-Conglycinin; duplicate divergence; glycinin; molecular evolution; positive selection; soybean

Introduction

Glycinin (11S) and β-conglycinin (7S) are the two primary classes of multi-subunit seed storage proteins and make up approximately 70% of the total protein in a typical soybean seed (Krishnan, 2000). These protein families are of great nutritional and economic importance. To date, considerable effort has been dedicated to describing the genes encoding glycinin and β-conglycinin proteins (Fischer and Goldberg, 1982; Scallon *et al*., 1985; Harada *et al*., 1989; Nielsen *et al*., 1989; Beilinson *et al*., 2002; Yoshino *et al*., 2002; Wang *et al*., 2008), because these genes are important potential targets for improving seed quality.

Glycinin is a hexameric protein composed of six similar subunits, each of which comprises an acidic polypeptide chain linked by disulfide bonding to a basic polypeptide chain (Badley *et al*., 1975). Several genes were found to encode these subunits. Fischer and Goldberg (1982) described three highly homologous *glycinin* genes and designated these genes as *Gy1*, *Gy2* and *Gy3* (group I). Scallon *et al*. (1985) reported the existence of two additional *glycinin* genes: *Gy4* and *Gy5* (group II). These five genes in the above two groups encode the predominant subunits of glycinin hexamers. All of the genes are composed of four exons and three introns, and the sequence identity is about 45% between the above two groups and about 80% within each group (Nielsen *et al*., 1989). In addition, two extra *glycinin* genes have been identified: one is *Gy6*, a pseudogene, and the other is *Gy7*, a gene with weak expression (Nielsen *et al*., 1989; Beilinson *et al*., 2002).

β-Conglycinin is a trimeric protein composed of the subunits α, α′ and β (Thanh and Shibasaki, 1978). Fifteen genes relating to these subunits have been identified and are designated as *CG*-1 to *CG*-15. These genes are divided into two major groups that are clustered in several genetic regions (Harada *et al*., 1989). Of these genes, *CG*-1 encodes the α′-subunit, *CG*-4 encodes the β-subunit (Harada *et al*., 1989), and *CG*-2 and *CG*-3 encode the α-subunit (Yoshino *et al*., 2002). However, the remaining genes are largely unexplored; it is unknown as to which linkage groups these genes belong to, whether these genes are functionally active or which genes encode each subunit type (α, α′ and β).

Molecular genetic investigation has provided considerable insights into the molecular function of these protein families, and evolutionary analysis will further clarify the origin and histories of the genes. As Dobzhansky (1973) noted that nothing in biology makes sense, except in the light of evolution, an understanding as to how evolution shaped the characteristics of these proteins will provide new insights into the roles of the proteins in the evolutionary success of organisms.

We analyzed the *glycinin* and *β-conglycinin* genes in terms of genomic organization, gene duplication, gene divergence, selection pressure and evolutionary

Correspondence: Dr Y-M Zhang, State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, PR China.
E-mail: soyzhang@njau.edu.cn
Received 15 March 2010; revised 17 June 2010; accepted 24 June 2010; published online 28 July 2010

relationships to homologous genes in four other species. We reconstructed the gene duplication history and present the mechanisms by which the two families evolved.

Materials and methods

Sequence collection

BLASTp was conducted with a cutoff *E* value of $1e^{-20}$ in Phytozome (http://www.phytozome.net). The databases used were as follows: soybean (*Glycine max*; Glyma1.0, Soybean Genome Project, Joint Genome Inst, Walnut Creek, CA, USA), Arabidopsis (*Arabidopsis thaliana*; TAIR8.0, The Arabidopsis Genome Initiative), poplar (*Populus trichocarpa*; JGI1.1, The Joint Genome Institute), *Medicago truncatula* (Mt2.0, US/EU *Medicago truncatula* genome sequencing project) and *Vitis vinifera* (September 2007 release).

The sequences used to generate BLASTp results are the reported glycinin and β-conglycinin protein sequences: *Gy1* (GenBank accession no. P04776), *Gy2* (P04405), *Gy3* (P11828), *Gy4* (P02858), *Gy5* (P04347), *Gy7* (Q6DR94), β-conglycinin α-chain (P13916), α′-chain (P11827) and β-chain (P25974). All of the results were downloaded from Phytozome Gbrower.

Synteny analyses

Syntenic information was collected from the Plant Genome Duplication Database (PGDD; http://chibba.agtec.uga.edu/duplication) and Phytozome (http://www.phytozome.net). The PGDD was established using several steps. First, BLASTp was used to search for potential anchors ($E<1e^{-5}$, top 5 matches) between every possible pair of chromosomes across multiple genomes. Then, homologous pairs were used as the input for MCscan. The resulting syntenic chains were evaluated using a procedure in ColinearScan, and alignments with an *E*-value $<1e^{-10}$ were considered significant matches (Tang *et al*., 2008).

Phylogenic analyses

Amino-acid sequence alignment was performed using MUSCLE (Multiple Sequence Caparison by Log-Expectation) in Jalview (Waterhouse *et al*., 2009). The neighbor joining trees and bootstrapping analysis were calculated using the Molecular Evolutionary Genetics Analysis (MEGA) 4.0 program (Tamura *et al*., 2007). The parameter setups were as follows: model, *p*-distance; bootstrap, 1000 replicates; and gap/missing data, pairwise deletion. The maximum-likelihood trees were generated with Tree-Puzzle 5.2 (Schmidt *et al*., 2002) using the default settings.

Calculating K_S and dating the duplication event

Protein sequences of the gene pairs were aligned in Jalview, and the results were used to guide the coding sequence (CDS) alignments by Pal2Nal (Suyama *et al*., 2006). K_S, the number of synonymous substitutions per site, was determined using the aligned CDS by Codeml procedure in phylogenetic analysis by maximum likelihood (PAML) 4.3 (Yang, 2007) after all alignment gaps were eliminated.

In dating segmental duplication events, six consecutive anchor points on each side flanking the *glycinin* or *β-conglycinin* genes were chosen, and 10 of these were used to calculate the average K_S after the minimum and maximum were removed. For those with fewer than 12 anchor points, all available anchor points were used. The approximate date of the duplication event was calculated using the mean K_S values from $T=K_S/2\lambda$ (Nei and Kumar, 2000), where the mean synonymous substitution rate (λ) for Fabaceae is $6.1\times1e^{-9}$ (Lynch and Conery, 2000).

Positive selection analyses

Positive selection was investigated using a maximum likelihood approach with Codeml procedure in PAML 4.3 (Yang, 2007), under the situations of the branch model and branch-site model. In the branch model, an excess of non-synonymous substitutions over synonymous substitutions is an important indicator of positive selection at the molecular level. Lineages that underwent positive selection may show a non-synonymous/synonymous rate ratio (d_N/d_S, denoted ω) that differs from those of other lineages, analyses of such lineages are referred to as branch models. The one-ratio model assumes the same d_N/d_S ratio for all branches in the phylogeny, whereas the two-ratio model assumes that one of the branches (the 'foreground branch') has a d_N/d_S ratio different from the other branches ('background branches'). Thus, the likelihood values under the one-ratio model and the two-ratio model can be compared using a likelihood ratio test (LRT). The purpose of the LRT is to see if the data fit the second model significantly better than the first. Further, a significant LRT determines that the d_N/d_S ratios differ between the foreground and background branches (Yang, 1998).

The branch-site model assumes that the ω ratio varies between codon sites and that there are four site classes in the sequence. The first class of sites is highly conserved in all lineages with a small ω ratio, ω_0. The second class includes neutral or weakly constrained sites for which $\omega=\omega_1$, where ω_1 is near or smaller than 1. In the third and fourth classes, the background lineages show ω_0 or ω_1, but foreground branches have ω_2, which may be greater than 1. When constructing the LRTs, the null hypothesis fixes $\omega_2=1$, allowing sites evolving under negative selection in the background lineages to be released from constraint and to evolve neutrally on the foreground lineage; the alternative hypothesis constrains $\omega_2\geqslant1$ (Yang and Nielsen, 2002; Zhang *et al*., 2005). The posterior probabilities associated with specific codons falling into a site class affected by positive selection were calculated using the Bayes empirical Bayes method described by Yang *et al*. (2005).

Sequences outside the syntenic region were removed and manually pruned from the topology, because it was impossible to determine whether these sequences shared the common ancestor with the *glycinin* or *β-conglycinin* genes, and the presence of the these sequences could bias the positive selection analyses. The pseudogenes were also removed because they are selectively neutral and much shorter than functional genes. The gaps in the sequences were manually removed according to the PAML manual: we retained the sites for which data were available for all but one or two sequences and we removed sites at which all sequences but one or two had alignment gaps.

Results

Sequence collection and their chromosomal location analyses

Using the amino-acid sequences of *Gy1–7*, we performed BLASTp searches of the soybean database, and the resulting sequences were used as secondary BLASTp queries. Using the 'all against all' approach, we matched *Gy1–7* to their corresponding loci (for example, *Gy1* to Glyma03g32030, *Gy2* to Glyma03g32020) and identified the additional sequence Glyma10g04270 (Table 1). As described in previous studies, the *Gy1–7* genes are distributed within linkage groups N, L, O or F (Nielsen *et al*., 1989; Diers *et al*., 1994; Beilinson *et al*., 2002). Most of the genes are adjacent to one another, and Glyma10g04270 is immediately adjacent to *Gy4* (Figure 1a).

To find orthologs of the *glycinin* genes, the eight sequences identified above were used as the basis for BLASTp searches against the genomes of *M. truncatula*, poplar, Arabidopsis and *V. vinifera*. We identified three orthologous sequences in *M. truncatula* that are adjacent to one another on chromosome 1; eight sequences in poplar that are distributed on chromosomes 1, 2 and 5 and scaffolds 117 and 4434; four sequences in Arabidopsis, distributed on chromosomes 1, 4 and 5; and four sequences in *V. vinifera*, tandemly spaced on chromosome 7 (Table 1).

Similarly, using the amino-acid sequences of *CG-1* to *CG-4* as initial search parameters, we identified genes homologous to *β*-conglycinin. We found nine homologous sequences distributed on chromosomes 2, 10 and 20 in soybean; six sequences distributed on chromosomes 1 and 7 in *M. truncatula*; three sequences distributed on chromosomes 5, 6 and 10 in poplar; one sequence on chromosome 3 in Arabidopsis; and two sequences tandemly spaced in chromosome 7 in *V. vinifera* (Figures 1d, e and f). On the basis of the BLASTp results, Glyma02g16440 and Glyma10g03390 were found to encode the *sucrose binding protein* (*SBP*) genes (Grimes *et al*., 1992). Glyma10g39150 encodes the *α*′-subunit, whereas Glyma20g28650 and Glyma20g28660 encode the *α*-subunit, and Glyma20g28460 and Glyma20g28640 encode the *β*-subunit. These results are consistent with the results of Davies *et al*. (1985), who showed that the *α*- and *β*-subunit genes are tightly linked. Tsukada *et al*. (1986) showed that the *α*-and *α*′-subunit genes are inherited independently, and Thanh *et al*. (2004) showed that the *β*-subunit is controlled by two linked genes on the same chromosome. Our results corroborate these previous findings.

Synteny analyses

Although the homologous sequences in soybean, *M. truncatula*, poplar, Arabidopsis and *V. vinifera* have been separated for at least 55 million years (Lavin *et al*., 2005), we observed strongly conserved synteny among the regions containing the genes (Table 1) across the five species. In regions containing *glycinin* genes and their homologs, the most strongly conserved gene synteny was found within four soybean chromosome segments (Figure 1a) containing homologous gene pairs (referred to as 'anchor points'). These gene pairs were found in sets of more than one hundred and up to more than one thousand (data not shown). We detected a considerable amount of conserved synteny between soybean and poplar, soybean and Arabidopsis and soybean and *V. vinifera*. The synteny was degraded between soybean and *M. truncatula*, indicating that a rearrangement occurred in *M. truncatula* that resulted in three homologous genes downstream from the syntenic region (Figures 1b and c). For the segments containing *β-conglycinin* genes and their homologs, the most strongly conserved gene synteny was found within four soybean chromosome segments that contained the greatest numbers of anchor points (Figures 1d, e and f).

On the basis of these syntenic alignments, we established the orthology of most sequences shown in Table 1, and we concluded that all of the syntenic regions shared a common ancestor. We further posit that the existence of two or more syntenic regions in one species is the result of chromosome segment duplication or whole-genome duplication (WGD). Moreover, we conclude that the segment from 28 460 to 28 630 on soybean chromosome 20 is a consequence of paracentric inversion (Figure 1d).

Table 1 Sequences identified by BLASTp in five species

Species	Glycinin *genes and their homologs*	β-Conglycinin *genes and their homologs*
Soybean	Glyma03g32010 (*Gy6*)[a], Glyma03g32020 (*Gy2*), Glyma03g32030 (*Gy1*), Glyma10g04270[a], Glyma10g04280 (*Gy4*), Glyma13g18450 (*Gy5*), Glyma19g34770 (*Gy7*), Glyma19g34780 (*Gy3*)	Glyma02g16440, Glyma10g03390, Glyma10g39150 (*α*′-subunit), Glyma10g39160[a], Glyma10g39170, Glyma20g28460 (*β*-subunit), Glyma20g28640 (*β*-subunit), Glyma20g28650 (*α*-subunit), Glyma20g28660 (*α*-subunit)
M. truncatula	Mt01g02315, Mt01g02316, Mt01g02318	Mt01g02215, Mt07g02220[b], Mt07g02223[b], Mt07g02224[b], Mt07g02226[b], Mt07g02227[b]
Poplar	Pt01g2364[b], Pt02g0360, Pt05g1327[a], Pt05g1330, Pt04434g0001[b], Pt00117g0071[b], Pt00117g0072[b], Pt00117g0073[b]	Pt05g1209, Pt10g0706[a], Pt06g0025[b]
Arabidopsis	At1g03880, At1g03890, At4g28520, At5g44120	At3g22640
V. vinifera	Vv7g0729[a], Vv7g0730[a], Vv7g0731, Vv7g0734	Vv7g0134, Vv7g0135

Note: The genes for *glycinin* or the subunits of the *β-conglycinin* genes within parentheses are the typical names in GenBank and are indicated by BLAST results.
[a]Pseudogenes, as indicated by previous results, cDNA searches and the NCBI Conserved Domain Database (Marchler-Bauer *et al*., 2009).
[b]Sequences that are absent in syntenic regions.

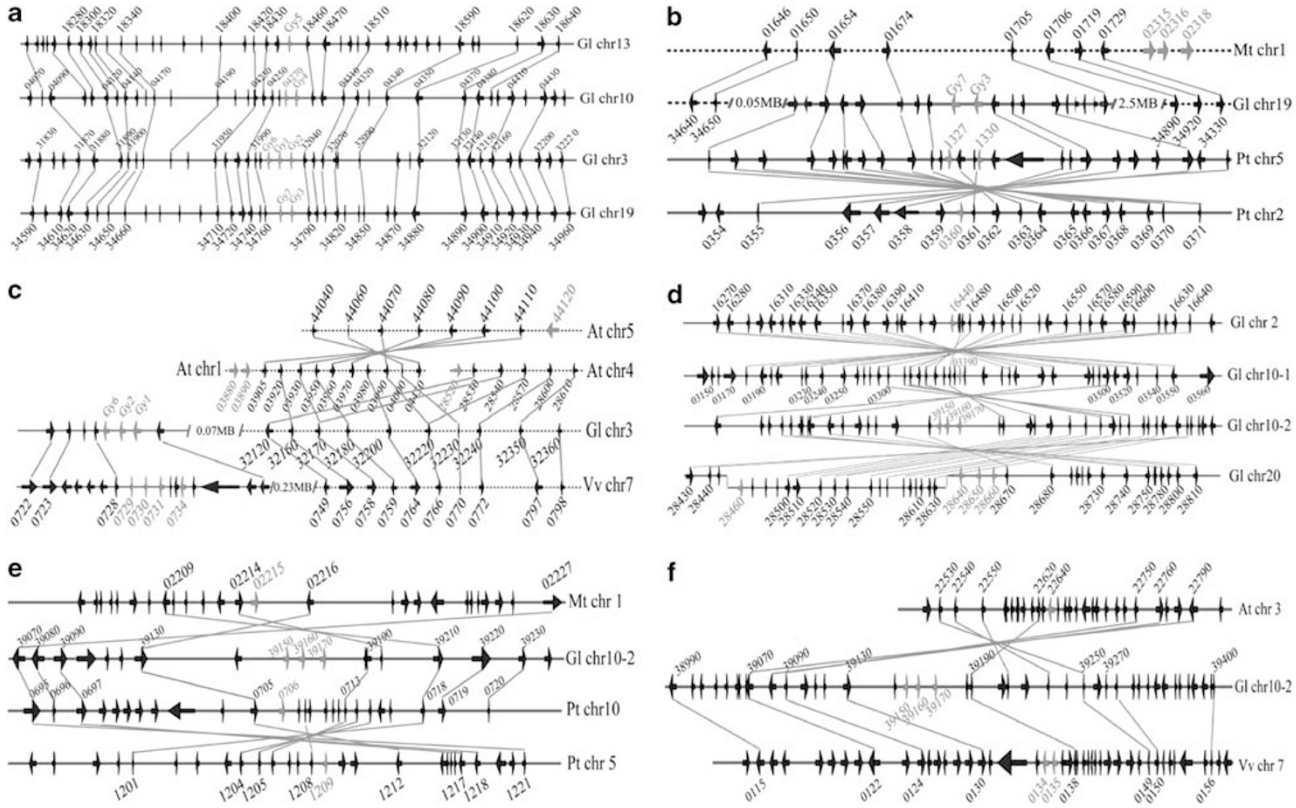

Figure 1 Synteny analyses for the sequences in Table 1 among the segments on soybean chromosomes 3, 10, 13 and 19 (**a**); on soybean chromosome 19, *M. truncatula* chromosome 1, poplar chromosomes 2 and 5 (**b**); on soybean chromosome 3, *M. truncatula* chromosome 1, Arabidopsis chromosomes 4, 5 and *V. vinifera* chromosome 7 (**c**); on soybean chromosomes 2, 10 and 20 (**d**); on soybean chromosome 10-2, *M. truncatula* chromosome 1 and poplar chromosomes 5 and 10 (**e**); and on soybean chromosome 10-2, Arabidopsis chromosome 3 and *V. vinifera* chromosome 7 (**f**). Homologous gene pairs are connected with lines. The chromosome segment is indicated by horizontal line, and the broad line with arrowhead represents gene and its transcriptional orientation. Horizontal line is drawn roughly to scale; broken line indicates segment not drawn to scale. The text besides the gene is the locus identifier prefix. The *glycinin* and *β-conglycinin* genes and their homologs are shown in red. Figures in (**a**), (**d**) and (**e**) are drawn roughly to 0.2 MB resolution, and in (**b**), (**c**) and (**f**) to 0.1 MB. The *β-conglycinin* genes lie in two separate regions of soybean chromosome 10; the two regions are denoted Gl chr 10-1 and Gl chr 10-2. A full color version of this figure is available at the *Heredity* journal online.

Phylogenetic analyses

To investigate the molecular evolution of, and phylogenetic relationships among, the sequences of these five species, we constructed two phylogenetic trees using the amino-acid sequences (Figure 2). Two subfamily clades, indicated by bold horizontal lines, are evident in each phylogenetic tree. On the *glycinin* gene tree (Figure 2a), the upper clade contains sequences from soybean and *M. truncatula*. The soybean sequences from three subfamilies are as follows: the first family contains *Gy1*, *Gy2* and *Gy3*; the second family contains *Gy6* and *Gy7*; and the final family contains *Gy4* and *Gy5*. These results are consistent with those presented by Beilinson *et al*. (2002). The three sequences from *M. truncatula* are sister to *Gy4* and *Gy5*. The lower clade contains sequences from the other three species, and genes from the same species group together with high bootstrap support (except for one poplar sequence, pt05g1372). On the *β-conglycinin* gene tree (Figure 2b), the upper clade contains the sequences encoding the *β*-conglycinin subunits with Glyma10g39160 and Glyma10g39170 forming sister clades. The lower clade contains the soybean sequences encoding the *SBP* genes, with the sequences from the other four species lying outside.

In the above two trees, genes from the same species mostly form monophyletic groups. Comparing the relationships of the groups with the species phylogenetic tree ((((soybean, *M. truncatula*), poplar), Arabidopsis), *V. vinifera*) (Cannon, 2009), we can infer that the above two trees should be midpoint-rooted, and the genes in each monophyletic group likely arose from lineage-specific gene duplication. For instance, the four genes in Arabidopsis likely arose from chromosome segment and tandem duplications after Arabidopsis and *V. vinifera* diverged, the *Gy6* or *Gy7* likely arose from the tandem duplication of an ancestral gene after soybean diverged from poplar or *M. truncatula* and the nine soybean genes in the *β*-conglycinin tree dividing into separate monophyletic groups is a consequence of functional divergence after ancestral gene duplications.

Gene duplications in the *glycinin* and *β-conglycinin* gene families

We hypothesize that large-scale segmental duplication events occurred early in the evolution of the two gene families. On the basis of the syntenic alignments, two rounds of segmental duplication led to the establishment of the glycinin and *β*-conglycinin families (Figures 1a and d). Using the K_S value as a proxy measure, we dated these duplication events (Table 2). The segment pair on chromosomes 3 and 19 and those on chromosomes 10

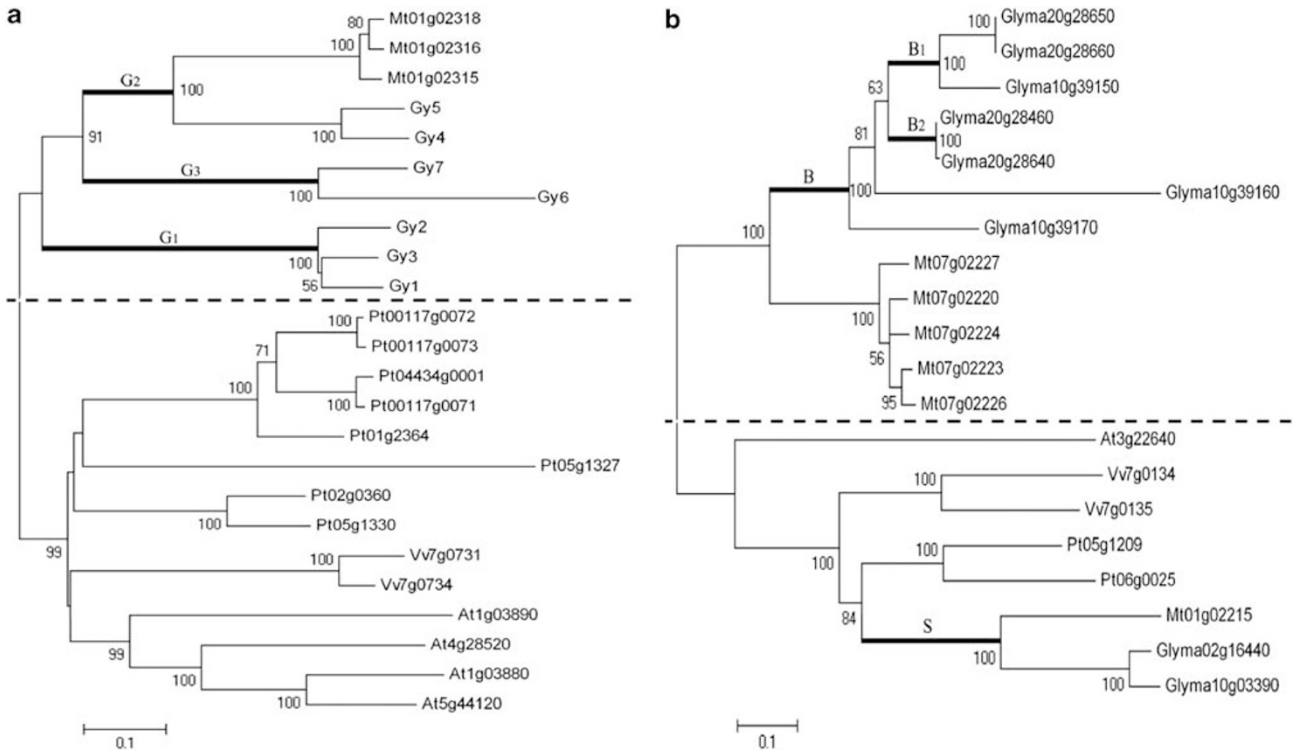

Figure 2 Phylogenetic analyses of the *glycinin* genes and their homologs (**a**) and the *β-conglycinin* genes and their homologs (**b**). The trees were constructed using the neighbor joining (NJ) method implemented in MEGA 4.0. The numbers beside the branches represent bootstrap values (>50%) based on 1000 replications. Sequences Glyma10g04270, Vv7g0729, Vv7g0730 and Pt10g0706 are excluded owing to their short lengths. The same analyses were performed in Tree-Puzzle (manual delete gaps; 434 and 416 amino acids were used in (**a**) and (**b**), respectively), which yielded highly similar results (data not shown). The horizontal dashed lines were manually drawn to assist in reading the trees, and the branches that were drawn in thick lines and labeled with letters G1, G2 or S, and so on, were supposed to be under different selection pressures.

Table 2 Estimates of the dates for the segmental duplication events in soybean[a]

Segment pairs	*Number of anchors*	K_S *(mean ± s.d.)*	*Estimated time (mya)*
Segments containing glycinin genes			
Chr 3 & chr 19	10	0.16 ± 0.07	13
Chr 10 & chr 13	10	0.15 ± 0.05	12
Chr 3 & chr 10	10	0.65 ± 0.13	53
Chr 3 & chr 13	10	0.64 ± 0.19	52
Chr 19 & chr 10	10	0.60 ± 0.12	49
Chr 19 & chr 13	10	0.61 ± 0.23	50
Segments containing β-conglycinin genes			
Chr 10-2 & chr 20	10	0.16 ± 0.08	13
Chr 2 & chr 10-2	5	1.82 ± 0.42	149
Chr 2 & chr 20	3	1.77 ± 0.31	145
Chr 10-1 & chr 10-2	5	1.75 ± 0.29	143
Chr 10-1 & chr 20	3	1.81 ± 0.18	148

Abbreviation: mya, million years ago.
[a]Highly similar K_S means were achieved for the segments on chr 3, chr 10, chr 13, chr 19, chr 10-2 and chr 20 when 20 anchors were used (data not shown).

and 13 all yielded similar K_S values (~0.15), corresponding to an event 12–13 million years ago (mya). The remaining four segment pairs give similar K_S values (~0.60), corresponding to an event 49–53 mya. Similarly, the two rounds of segmental duplication in the *β*-conglycinin family likely occurred at ~13 mya and more than 100 mya.

Several tandem duplicates were also identified in the glycinin and *β*-conglycinin families, such as the duplications that yielded the *Gy6–Gy1–Gy2*, *Gy7–Gy3* and *β-conglycinin* genes on chromosome 10-2 or 20 (Figures 1a and d). From the apparent synteny and the phylogenetic analyses, we believe that the duplication that yielded *Gy7–Gy3* likely also yielded *Gy6–Gy1*. This duplication must have occurred after soybean split with poplar because no such tandem duplication was detected in poplar (Figure 1b). The duplication that yielded *Gy1–Gy2* occurred after the segmental duplications of chromosomes 3 and 19 because only *Gy3* was found in this syntenic position. The duplication that originated from Glyma10g39150–Glyma10g39160–Glyma10g39170 is ~40 mya because Glyma10g39150 and Glyma10g39170 show a K_S value of 0.5. The duplication that gave rise to Glyma20g28460–Glyma20g28640 and Glyma20g28650–Glyma20g28660 should have occurred <10 mya because of the high sequence similarities and small K_S values of the duplicate genes.

Positive selection analyses

On the *glycinin* gene tree, three branches were independently defined as the foreground branch, one leading to *glycinin group I genes* (G1), one leading to the *glycinin group II* genes (G2) and one leading to the *Gy6* and *Gy7*

Table 3 Summary statistics for detecting selection using branch and branch-site models of PAML

Model	*Glycinin family*		*β-conglycinin family*	
	$-ln$ L	*Parameter estimates*[a]	$-ln$ L	*Parameter estimates*[a]
M0 (one-ratio)	15385.72	$\omega_0 = 0.28$	10554.78	$\omega_0 = 0.34$
Branch-specific model (two-ratio)	*Branch G1*		*Branch B*	
	15381.58*	$\omega_0 = 0.26$, $\omega_1 = 0.39$	10550.61*	$\omega_0 = 0.30$, $\omega_1 = 0.45$
	Branch G2		*Branch B1*	
	15385.72	$\omega_0 = 0.28$, $\omega_1 = 0.28$	10554.46	$\omega_0 = 0.33$, $\omega_1 = 0.39$
	Branch G3		*Branch B2*	
	15373.97*	$\omega_0 = 0.26$, $\omega_1 = 0.53$	10552.63**	$\omega_0 = 0.33$, $\omega_1 = 0.62$
Branch-site model[b]	*Branch G1*		*Branch S*	
A_{null}	14669.80	$p_0 = 0.57$, $p_1 = 0.23$ ($p_2+p_3 = 0.20$) $\omega_0 = 0.19$, $\omega_1 = \omega_2 = 1$	10471.21	$p_0 = 0.58$, $p_1 = 0.25$($p_2+p_3 = 0.17$) $\omega_0 = 0.22$, $\omega_1 = \omega_2 = 1$
A	14663.61*	$p_0 = 0.62$, $p_1 = 0.24$ ($p_2+p_3 = 0.12$) $\omega_0 = 0.20$, $\omega_2 = 2.56$	10468.21**	$p_0 = 0.63$, $p_1 = 0.27$ ($p_2+p_3 = 0.10$) $\omega_0 = 0.23$, $\omega_2 = 2.08$

Abbreviation: PAML, phylogenetic analysis by maximum likelihood.
Note: Pseudogene *Gy6* was included in the branch-specific model for glycinin family; 426 codons for the glycinin family and 411 codons for the *β*-conglycinin family were used. Branches G1, G2, G3 and B, B1, B2, S are shown in Figure 2.
*$P<0.01$; **$P<0.05$ (χ^2 test).
[a]The proportion of sites (p_0, p_1, and so on) estimated to have ω_0, ω_1, and so on.
[b]Positive sites for foreground lineages with posterior probability above 65% are 14S, 55K, 76R, 85E, 89Q, 104S, 117S, 127Y, 143W, 202T, 207E, 208H, 211S, 217A, 227K, 229A, 241K, 245D, 247S, 263H, 268T, 325L, 350R, 379T, 381M, 414I, 420F and 421K in the glycinin family and 134R, 198S, 201Q, 246E, 250Q, 302Q, 305E, 351N, 402K and 404E in the *β*-conglycinin family.

(G3) (Figure 2a). A two-ratio model was used to determine whether there were different selective pressures, indicated by different ω values, on these lineages. The two-ratio model, with $\omega_1 = 0.39$ or 0.53 for branch G1 or G3 as the foreground branch, fits the data significantly better than the one-ratio model with a single ω ($\omega_0 = 0.28$) ($P<0.01$). The other two-ratio model, which defined branch G2 as the foreground branch, was no better than the one-ratio model ($P>0.05$) (Table 3). Significant positive selection was detected using the branch-site model when the branch G1 was considered the foreground branch ($P<0.01$). The estimated parameters suggested that about 12% of sites were under positive selection with $\omega_2 = 2.56$ (Table 3). When either branch G2 or G3 was brought to the foreground, no significant positive selection was detected (data not shown).

In the phylogenetic tree of the *β-conglycinin* genes, either branch B (leading to the *β*-glycinin family) or branch B2 (leading to *β*-subunit gene subfamily) could be considered as the foreground branch (Figure 2b); the two-ratio model with $\omega_1 = 0.45$ or 0.62 fit the data significantly better than the one-ratio model with a $\omega_0 = 0.34$ ($P<0.01$ or 0.05). However, using branch B1 (leading to the α/α'-subunit gene subfamily; Figure 2b), the two-ratio model fit the data no better than the one-ratio model (Table 3). No significant positive selection was detected along the three branches B, B1 and B2 using the branch-site model. However, significant positive selection was detected ($P<0.05$) when branch S (leading to the three non-nutrient genes *Glyma10g03390*, *Glyma02g16440* and *Mt01g02215*; Figure 2b) was defined as the foreground branch. These results suggest that the branches G1, G3, B and B2 are under relaxed purifying selection and the branches G2 and B1 are under strong purifying selection, whereas along the branches G1 and S, a few amino acids seem to have been under positive selection.

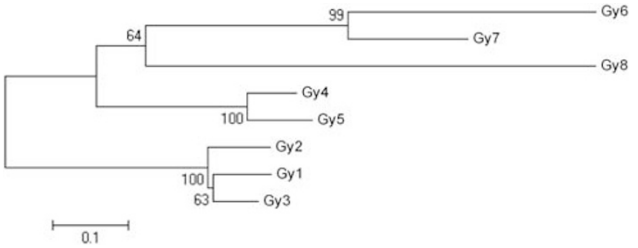

Figure 3 Phylogenetic analyses of the genes *Gy1–8*. The trees were constructed using the neighbor joining (NJ) method implemented in MEGA 4.0. The tree is rather crude owing to the highly divergent sequences of the pseudogenes *Gy6* and *Gy8*.

Discussion

Gy7 is losing its function

Beyond the five main glycinin genes *Gy1–5*, Nielsen *et al*. (1989) identified two other genes immediately downstream from the *Gy2* and *Gy3*, which they called 'glycinin-related' or *G**, because they hybridize weakly with the five genes. Nielsen *et al*. (1989) further suggested that the glycinin-related genes could encode other glycinin subunit families whose members accumulate in minor amounts in seeds. Beilinson *et al*. (2002) more recently renamed the glycinin-related genes *Gy6* and *Gy7*, and showed that *Gy6* is a pseudogene and *Gy7* is expressed several orders of magnitude lower than the other five *glycinin* genes (Nielsen *et al*., 1989). Both of these results suggest that the glycinin-related genes form a third group distinct from groups I and II. In this study, we found another pseudogene, *Glyma10g04270*, and both syntenic and phylogenic analyses suggested that it is a new member of the third group (Figure 3). We thus designated this pseudogene as *Gy8*. Our results suggest

that the third group arose from a tandem duplication of the common ancestor of the *glycinin* genes and that two rounds of large-scale gene duplication followed this duplication.

Systematic analysis has revealed that, in eukaryotic genomes, a gene's propensity to be lost is significantly negatively correlated with both gene expression level and the ongoing selection pressure (Krylov *et al*., 2003; Davis and Petrov, 2004; Wolf *et al*., 2006). In accordance with the fact that *Gy7* has a weak expression activity (Nielsen *et al*., 1989; Beilinson *et al*., 2002) and is subject to relaxed selection, we deduce that the *Gy7* gene has a higher propensity to be lost. On the other hand, the pseudogenation of both *Gy6* and *Gy8* in the same group suggests that *Gy7* may eventually meet the same fate. Therefore, we predict that the soybean no longer needs the genes in the third group and that all of the third group genes will become pseudogenes and eventually disappear from the genome or adopt novel gene functions.

The glycinin and β-conglycinin families arose through WGD and tandem duplication

Gene duplication can occur by a variety of mechanisms, including WGD, chromosomal segmental duplication, tandem duplication and duplicative transposition. In soybean, researchers postulate that two rounds of WGD have occurred since the Arabidopsis/soybean split, at 50–60 and 10–15 mya (Shoemaker *et al*., 2006; Fawcett *et al*., 2009; Schmutz *et al*., 2010). Our syntenic analyses show that the *glycinin* genes arose primarily from two rounds of segmental duplication that occurred during the two soybean WGD events.

Several lines of evidence support the hypothesis that the *glycinin* genes arose from the two WGD events. First, we dated the two rounds of segmental duplication using K_S values of paralogous genes, one at 49–53 mya and another at 12–13 mya (Table 2), and the times are consistent with the times when the soybean WGD occurred. Second, the large regions of syntenic segments extend over 13 megabases (MB) between chromosomes 3 and 19, over 2.6 MB between chromosomes 10 and 13 and over 3.3 MB between chromosomes 10 and 19. These results indicate that the duplications arose through WGD rather than chromosomal segmental duplications. Third, although some evidence shows that gene duplication is a continuous and frequently occurring process, mounting genomic data indicate that many duplications are formed during major, large-scale gene duplication events (Lynch and Conery, 2000; Raes and Van de Peer, 2003). For similar reasons, the β-conglycinin segments, which were duplicated at around 13 mya and extend over 10 MB, originated during the WGD as well. In the analyses of synteny, a segment on soybean chromosome 1 without genes homologous to the β-conglycinin protein was found to share strong synteny with the segments containing *β-conglycinin* genes (Figure 4). The duplications between the segment on chromosome 1 and its syntenic region on chromosome 10-2 or 20 occurred around 50 mya, suggesting that the segments indeed arose through the soybean WGD.

On the basis of previous data and our new analyses, we can now present the entire history of the evolution of the glycinin and β-conglycinin families: for the glycinin family, the ancestor genes initially experienced a tandem duplication event at the time between the poplar/soybean split event and the more ancient WGD event of soybean; then, the genome underwent two subsequent rounds of WGD that occurred at 50–60 and 10–15 mya, thereby resulting in duplication of the *glycinin* genes; and finally, a tandem duplication likely gave rise to the *Gy1* and *Gy2* genes. For the β-conglycinin family, the ancestor of the *β-conglycinin* and *SBP* genes split before 100 mya through a segmental duplication. The common ancestor of the *β-conglycinin* gene first created two copies through the ancient soybean WGD. One copy lost its function and then disappeared, whereas the other copy first underwent two tandem duplications around 40 mya and then WGD at 10–15 mya. One of the two segments arising during the 10–15 mya WGD became the segment on chromosome 10 accompanied by gene pseudogenization, whereas the other one became the segment on chromosome 20 accompanied by gene loss, tandem duplications and paracentric inversion (Figure 5).

The maintenance and divergence of duplicate genes

When a duplicate copy of a gene is present, mutations in one copy are often selectively neutral and will turn most genes into non-functional pseudogenes (the process of pseudogenization); alternatively, the mutations may create a novel function for the gene (neofunctionalization),

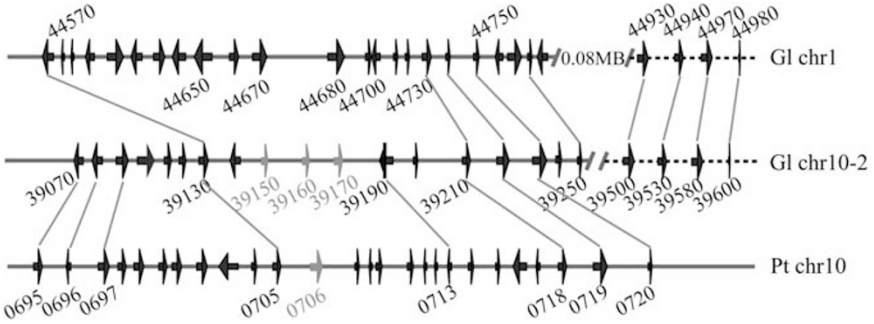

Figure 4 The segment, without homologous genes of β-conglycinin protein, on soybean chromosome 1 is found to share strong synteny with the segments containing *β-conglycinin* genes. The time of the duplication of the segment on chromosome 1 and its syntenic region on chromosome 10-2 or 20 is around 50 mya, suggesting that the segments arose during the more ancient WGD (chr 1 and chr 10-2, $K_S = 0.64$, s.d. = 0.20, $n = 10$, time = 52 mya; chr 1 and chr 20, $K_S = 0.58$, s.d. = 0.10, $n = 5$, time = 48 mya).

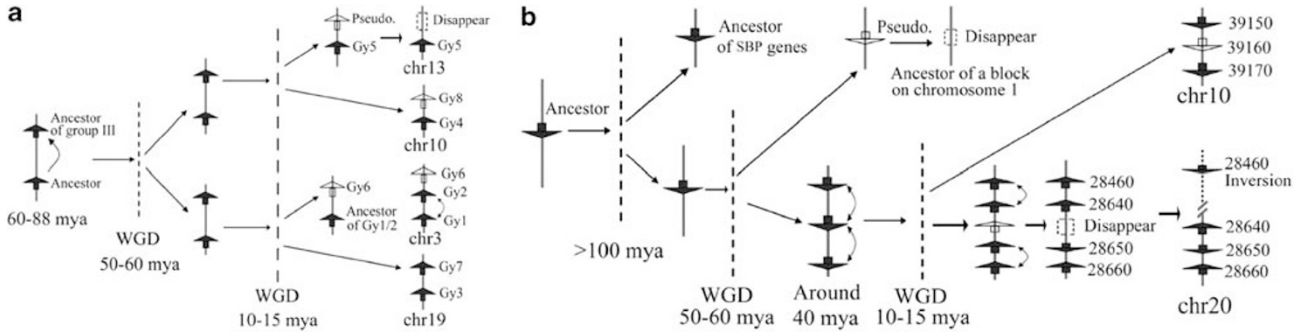

Figure 5 Gene duplication history of glycinin (**a**) and *β*-conglycinin (**b**). The determined and pending directions of gene duplication events are indicated by single and double arrows at both ends, respectively. Genes and chromosomes follow the legend in Figure 1, but the genes shaded in white indicate pseudogenes and broken line boxes indicate the gene deletion events. The time 60–88 mya in (**a**) is the time between the poplar/soybean split event and the more ancient WGD event of soybean (Sanderson *et al*., 2004).

or the duplicates may be maintained and specialized to perform complementary ancestral functions (subfunctionalization) (Demuth and Hahn, 2009). However, if the presence of duplicate genes is beneficial owing to the additional RNA transcribed, duplicates may maintain the same function (Zhang, 2003). This phenomenon can occur by gene conversion or purifying selection. The main differences between gene conversion and purifying selection are that: (i) gene conversion removes synonymous differences, while purifying selection does not, and (ii) the conditions for gene conversion are more restrictive, while purifying selection is more common and important (Hurst and Smith, 1998; Nei *et al*., 2000). We detected strong evolutionary purifying selection (along the branch G2) and relaxed purifying selection (along the branch G1 or G3) in the history of the glycinin family. Therefore, we conclude that the soybean benefits from the presence of these duplicate genes because more RNA or protein products are produced; thus, purifying selection maintains these genes, although some duplicates probably do experience a period of relaxed selection after duplication.

In addition to the purifying selection, significant branch-site positive selection was also detected along the branch leading to the *group I glycinin* gene subfamily (Table 3), indicating that changes in the amino acids through mutation are selected upon (positively or negatively). This offers a probable explanation as to why the *group I glycinin* genes encode proteins containing substantially more nutritionally important sulfur-containing amino acids, such as methionine and cysteine, compared with *group II glycinin* genes. Nevertheless, the positive selection may be a primary cause of the divergence between groups I and II.

The duplicated genes in the *β-conglycinin* family seem to be retained in a similar way as those in the glycinin family. The *β*- and the *α*/*α*′-subunit genes are differentiated mainly by the presence or absence of a specific DNA segment in an exon. Some researchers suggest that this segment is the result of an insertion event in *α*/*α*′-subunit genes (Casey and Domoney, 1987), but there is no evidence of a transponson-like element flanking the insert (Doyle *et al*., 1986). In this study, we found that the segment was present in the *β-conglycinin* ancestor gene and orthologous genes and absent only in the *β*-subunit gene, suggesting that the segment resulted from a deletion event.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful to the Editor and two anonymous referees for their constructive comments and suggestions that significantly improved the presentation of the manuscript. The work was supported in part by the National Basic Research Program of China (2006CB101708), the National Natural Science Foundation of China (30971848), Jiangsu Natural Science Foundation (BK2008335), NCET (NCET-05-0489) and the 111 Project (B08025) to YM Zhang.

References

Badley RA, Atkinson D, Hauser H, Oldani D, Green JP, Stubbs JM (1975). The structure, physical and chemical properties of the soybean protein glycinin. *Biochim Biophys Acta* **412**: 214–228.

Beilinson V, Chen Z, Shoemaker RC, Fischer RL, Goldberg RB, Nielsen NC (2002). Genomic organization of *glycinin* genes in soybean. *Theor Appl Genet* **104**: 1132–1140.

Cannon S (2009). Genetics and genomics of soybean. In:Stacey G (ed). *Legume Comparative Genomics*. Springer: Berlin. pp 35–54.

Casey R, Domoney C (1987). The structure of plant storage protein genes. *Plant Mol Biol* **5**: 261–281.

Davies CS, Coates JB, Nielsen NC (1985). Inheritance and biochemical analysis of four electrophoretic variants of *β*-conglycinin from soybean. *Theor Appl Genet* **71**: 351–358.

Davis JC, Petrov DA (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* **2**: 318–326.

Demuth JP, Hahn MW (2009). The life and death of gene families. *Bioessays* **31**: 29–39.

Diers BW, Beilinson V, Nielsen NC, Shoemaker RC (1994). Genetic mapping of the *Gy4* and *Gy5* glycinin genes in soybean and the analysis of a variant of *Gy4*. *Theor Appl Genet* **89**: 297–304.

Dobzhansky T (1973). Nothing in biology makes sense except in the light of evolution. *Am Biol Teach* **35**: 125–129.

Doyle JJ, Schuler MA, Godette WD, Zenger V, Beachy RN, Slightom JL (1986). The glycosylated seed storage proteins of *Glycine max* and *Phaseolus vulgaris*. Structural homologies of genes and proteins. *J Biol Chem* **261**: 9228–9238.

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

Fischer RL, Goldberg RB (1982). Structure and flanking regions of soybean seed protein genes. *Cell* **29**: 651–660.
Grimes HD, Overvoorde PJ, Ripp K, Franceschi VR, Hitz WD (1992). A 62-kD sucrose binding protein is expressed and localized in tissues actively engaged in sucrose transport. *Plant Cell* **4**: 1561–1574.
Harada JJ, Barker SJ, Goldberg RB (1989). Soybean *β-conglycinin* genes are clustered in severa1 DNA regions and are regulated by transcriptional and posttranscriptional processes. *Plant Cell* **1**: 415–425.
Hurst LD, Smith NGC (1998). The evolution of concerted evolution. *Proc R Soc Lond Ser B* **265**: 121–127.
Krishnan HB (2000). Biochemistry and molecular biology of soybean seed storage proteins. *J New Seeds* **2**: 1–25.
Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13**: 2229–2235.
Lavin M, Herendeen PS, Wojciechowski MF (2005). Evolutionary rates analysis of *Leguminosae* implicates a rapid diversification of lineages during the Tertiary. *Syst Biol* **54**: 530–549.
Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.
Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH *et al.* (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**: D205–D210.
Nei M, Kumar S (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press: Oxford.
Nei M, Rogozin IB, Piontkivska H (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci USA* **97**: 10866–10871.
Nielsen NC, Dickinson CD, Cho TJ, Thanh VH, Scallon BJ, Fischer RL *et al.* (1989). Characterization of the *glycinin* gene family in soybean. *Plant Cell* **1**: 313–328.
Raes J, Van de Peer Y (2003). Gene duplications, the evolution of novel gene functions, and detecting functional divergence of duplicates *in silico*. *Appl Bioinform* **2**: 92–101.
Sanderson MJ, Thorne JL, Wikström N, Bremer K (2004). Molecular evidence on plant divergence times. *Am J Bot* **91**: 1656–1665.
Scallon B, Thanh VH, Floener LA, Nielsen NC (1985). Identification and characterization of DNA clones encoding group-II glycinin subunits. *Theor Appl Genet* **70**: 510–519.
Schmidt HA, Strimmer K, Vingron M, Haeseler A (2002). Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502–504.
Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W *et al.* (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183.
Shoemaker RC, Schlueter J, Doyle JJ (2006). Paleopolyploidy and genome duplication in soybean and other legumes. *Curr Opin Plant Biol* **9**: 104–109.
Suyama M, Torrents D, Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609–W612.
Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.
Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008). Synteny and collinearity in plant genomes. *Science* **320**: 486–488.
Thanh VC, Trang DTX, Liu SS, Zhou JM, Hirata Y (2004). Evaluation of 7S *β*-subunit deficiency and its inheritance among soybeans *Glycine max* L. in the Mekong DeltaVietnam. *Biosphere Conser* **6**: 1–5.
Thanh VH, Shibasaki K (1978). Major proteins of soybean seeds: subunit structure of beta-conglycinin. *J Agric Food Chem* **26**: 692–695.
Tsukada Y, Kitamura K, Harada K, Kaizumu N (1986). Genetic analysis of subunits of two major storage protein (*β*-conglycinin and glycinin) in soybean seeds. *Jpn J Breed* **36**: 390–400.
Wang CM, Wu XL, Jia FX, Zhang JS, Chen SY (2008). Genetic variations of glycinin subunit genes among cultivated and wild type soybean species. *Prog Nat Sci* **18**: 33–41.
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.
Wolf YI, Carmel L, Koonin EV (2006). Unifying measures of gene function and evolution. *Proc Biol Sci* **273**: 1507–1515.
Yang Z (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**: 568–573.
Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.
Yang Z, Nielsen R (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**: 908–917.
Yang Z, Wong WSW, Nielsen R (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**: 1107–1118.
Yoshino M, Kanazawa A, Tsutsumi K, Nakamura I, Takahashi K, Shimamoto Y (2002). Structural variation around the gene encoding the α subunit of soybean *β*-conglycinin and correlation with the expression of the α subunit. *Breed Sci* **52**: 285–292.
Zhang J, Nielsen R, Yang Z (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472–2479.
Zhang JZ (2003). Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292–298.