Heredity

Heredity (2013) 110, 259–266
© 2013 Macmillan Publishers Limited All rights reserved 0018-067X/13
www.nature.com/hdy

ORIGINAL ARTICLE

The origin and functional transition of *P34*

Q-G Li[1,2] and Y-M Zhang[1]

P34, a storage protein and major soybean allergen, has undergone a functional transition from a cysteine peptidase to a syringolide receptor. An exploration of the evolutionary mechanism of this functional transition is made. To identify homologous genes of *P34*, syntenic network was constructed using syntenic relationships from the Plant Genome Duplication Database. The collected homologous genes, along with SPE31, a highly homologous protein to P34 from the seeds of *Pachyrhizus erosus*, were used to construct a phylogenetic tree. The results show that multiple gene duplications, exon shuffling and following granulin domain loss and some critical point mutations are associated with the functional transition. Although some tests suggested the existence of positive selection, the possibility that random fixation under relaxation of purifying selection results in the functional transition is also supported. In addition, the genes Glyma08g12340 and Medtr8g086470 may belong to a new group within the papain family.

Heredity (2013) **110,** 259–266; doi:10.1038/hdy.2012.81; published online 5 December 2012

Keywords: functional transition; molecular evolution; P34; SPE31; syntenic network

INTRODUCTION

The origination of new genes is a fundamental process in molecular evolution. To date, several molecular mechanisms have been known to be involved in the emergence of new genes, such as exon shuffling, gene duplication, retroposition and the action of mobile elements (Long *et al.*, 2003; Ding *et al.*, 2010; Zhan *et al.*, 2012). Some novel genes even undergo a transition from one function to another, for example, macrophage-stimulating protein (Patthy, 2008). To better understand the new function of protein, elucidating the functional transition of proteins is warranted.

P34 (Gly m Bd 30K, Glyma08g12270) is a moderately abundant protein in soybean seeds and cotyledons but its level in mature leaves is low (Herman *et al.*, 1990; Kalinski *et al.*, 1990, 1992; Ji *et al.*, 1998). P34 is processed from a 46-kDa glycoprotein precursor (Herman *et al.*, 1990), and specifically binds with syringolide, an elicitor that triggers the hypersensitive response specifically in soybean cultivars with the resistance gene *Rpg4* (Keen and Buzzell, 1991), indicating that P34 may be the receptor that mediates syringolide signaling (Ji *et al.*, 1998). P34 has also been shown to interact with vegetative storage protein (Ji *et al.*, 1998) and NADH-dependent hydroxypyruvate reductase (HPR) that was a potential second messenger for P34 (Okinaka *et al.*, 2002). In addition, P34 has been found to be a major soybean allergen that is most strongly and frequently recognized in soybean-sensitive patients (Ogawa *et al.*, 1993). Although amino-acid sequence analyses indicate that P34 belongs to a papain-type cysteine peptidase family (Herman *et al.*, 1990; Kalinski *et al.*, 1990, 1992), whose members contain a highly conserved catalytic triad (Cys–His–Asn; Kamphuis *et al.*, 1985), its peptidase activity has not been demonstrated and the replacement of catalytic cysteine with glycine makes P34 belong to a unique group of the papain family (Ji *et al.*, 1998; Okinaka *et al.*, 2002; Zhang *et al.*, 2006). Clearly, P34 has undergone a functional transition from a cysteine peptidase of the papain family to a syringolide receptor. However, it is not so clear about the evolutionary mechanism of the functional transition, for example, when, how and why the novel function of P34 was developed. Recently, the crystal structure of SPE31, close homolog to P34 from the seeds of *Pachyrhizus erosus*, was determined. Detailed analyses of the SPE31 structure bound to a natural peptide, probably from part of a second messenger for SPE31, revealed how catalytic activity of SPE31/P34 is lost and how SPE31/P34 may bind to other proteins and small molecules (especially syringolides; Zhang *et al.*, 2006).

To trace the evolution of a gene over a time period of interest, we needed to obtain homologous sequences from plants diversified in that time period; these sequences could be used to reconstruct a phylogenetic tree to infer the evolutionary history of a gene. At present, homology searches (frequently BLAST) based on similarities between the query and target sequences are widely used. However, it is difficult to determine a suitable significance threshold to filter the numerous returned hits. An unsuitable threshold may result in numerous unnecessary sequences or the loss of some necessary sequences. Notably, synteny (conserved gene order, collinearity) provides additional direct evidence for the common origin of two genes with a syntenic relationship. With rapidly increasing amounts of genome sequencing data, more and more syntenic blocks have been identified. To date, the Plant Genome Duplication Database (PGDD) has identified and cataloged plant genes from 19 plant genomes in terms of intra-genome or cross-genome syntenic relationships (Tang *et al.*, 2008a, b). With the available abundant database resources, directly mining the database to find homologous sequences rather than simply using it to locate syntenic blocks after a homology search is required. This idea is same as that in recent MCScanX packages (Wang *et al.*, 2012).

The goal of this paper was to explore the evolutionary history of *P34* and to understand when and how the function of P34 was

[1]State Key Laboratory of Crop Genetics and Germplasm Enhancement, Department of Crop Genetics and Breeding, College of Agriculture, Nanjing Agricultural University, Nanjing, China and [2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China
Correspondence: Dr Y-M Zhang, College of Agriculture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, China.
E-mail: soyzhang@njau.edu.cn or soyzhang@hotmail.com

Received 10 February 2012; revised 20 September 2012; accepted 1 October 2012; published online 5 December 2012

transformed from a papain-like cysteine peptidase to a syringolide receptor or an allergen. In this study, we first obtained the homologous sequences using syntenic relationships from the PGDD. We then combined these sequences, gene expression and crystal structure data into the framework of evolution of P34. To understand what drove the functional change of P34, we examined variations in molecular pressure along phylogenetic branches and performed a series of tests for selection on branches of interest.

MATERIALS AND METHODS

Data collection

The syntenic data for further sequence collection were downloaded from the PGDD (http://chibba.agtec.uga.edu/duplication/). Genome sequences and other annotated data were collected from Phytozome (http://www.phytozome.net/). The details of these genomes, such as the release versions for genome annotations used in the study, are available in Table 1. The expression data for *Glycine max*, obtained from next-generation sequencing (Severin *et al.*, 2010), was downloaded from the Soybase (http://soybase.org/soyseq/). The mRNA sequence and crystal structure of SPE31 were obtained from GenBank (DQ152924) and the RCSB Protein Data Bank (2B1N), respectively.

Syntenic network analyses

Syntenic network analyses were conducted using the open-source graph manipulation software 'igraph' in the R platform (Csardi and Nepusz, 2006). This program, named syntenic_network.R (Supplementary Table S1), included two steps. The first step was to create a large undirected network from the above syntenic data, with vertices representing genes and edges representing syntenic relationships. The second one was to extract a subnetwork of *P34* in which all vertices were reachable via some paths but did not connect with any other vertex in the large network. The genes in the subnetwork are syntenic homologous genes of *P34*. This approach is similar to that in MCScanX packages (Wang *et al.*, 2012). The time of divergence of a pair of genes with a syntenic relationship can be roughly estimated by computing mean Ks values for all gene pairs located in the same syntenic blocks (Lavin *et al.*, 2005), where all the Ks values were also downloaded from PGDD along with syntenic data.

Coding sequence examination and pseudogene identification

To validate gene annotations, we manually confirmed each coding sequence by referring to the gene models (exon–intron structures) of *P34*. To identify pseudogenes that are potentially misannotated, we adopted the criteria that pseudogenes commonly contain nonsense mutations, frameshift mutations or partial nucleotide deletions causing a loss of function and rare expression. These sequence manipulations were performed in MEGA5 (Tamura *et al.*, 2011).

Phylogeny reconstruction

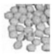
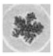
After checking coding sequences and filtering out pseudogenes, the coding sequences of remaining genes along with *SPE31* were used to reconstruct a phylogenetic tree in which topological structure was inferred by the MrBayes program (Ronquist and Huelsenbeck, 2003), and the branch lengths were computed using the CODEML program in PAML with model M0 (Yang, 2007). Multiple sequence alignments were conducted using MUSCLE (Edgar, 2004).

GABranch and selection tests

The ratio of nonsynonymous to synonymous substitution rates (d_N/d_S, ω) is commonly considered to be a measure of selection at the protein level, with values of $\omega<1, =1$ and >1 indicating negative purifying selection, neutral evolution and positive selection, respectively. We applied the GABranch method to investigate the variation of ω along various lineages. The GABranch method uses a genetic algorithm to fit data and does not need to specify particular lineages *a priori* (Pond and Frost, 2005).

A number of codon substitution models to test for positive selection have been implemented in CODEML of PAML (Yang, 2007). First, two pairs of site models, which allow ω to vary among codons, M1a vs M2a and M7 vs M8,

Table 1 The materials used to collate P34 genes

Species name		*Common name*	*Release version*	*Gene number*	*Access*	*Gene ID*
	*Glycine max*	Soybean	Release 1 (December 2008)	66 153	JGI	Glyma-
	Medicago truncatula	Barrel medic	Mt 3.5.1 (December 2010)	45 108	JCVI	Medtr-
	Prunus persica	Peach	Version 1.0	27 864	JGI	ppa004381m
	Fragaria vesca	Strawberry	December 2010	34 809	PFR	gene29675
	*Carica papaya*	Papaya	December 2007	25 536	Hawaii	evm.TU.supercontig_7.74
	Theobroma cacao	Cacao	Release 0.9 (September 2010)	28 798	CIRAD	Tc06_g014280
	Vitis vinifera	Grape vine	Genoscope (August 2007)	26 346	Genoscope	GSVIVT01021223001

were used. Then, we used the branch-site models to detect positive selection that affects only a few sites on prior lineages. Each pair of models was compared by likelihood ratio test. When the likelihood ratio test suggested a positive selection, finally, the Bayes Empirical Bayes method was implemented to calculate posterior probabilities for site classes under positively selective models.

Ancestral sequences reconstruction

To dissect the evolutionary details of individual sites, ancestral amino acid of interior nodes were reconstructed using MEGA5 (Tamura *et al.*, 2011), only maintaining the sites with maximum probabilities of >0.9. The coding sequences for positive selection test in MEGA5 were reconstructed using ANC-GENE program (Zhang *et al.*, 1998).

RESULTS

Syntenic network and phylogeny reconstruction

Using syntenic relationship data downloaded from PGDD, a syntenic network of *P34* was constructed, and 13 homologous genes from seven species were found on the network (Figure 1). After filtering out three pseudogenes (Glyma15g08950, Glyma05g29130 and Glyma 05g29180) and one gene with incomplete sequence (Medtr2g15920), the coding sequences of the remaining nine genes, along with that of *SPE31* from *P. erosus*, were aligned using the MUSCLE (Edgar, 2004), and used to reconstruct a phylogenetic tree (Figure 2) using the MrBayes program (Ronquist and Huelsenbeck, 2003) and the CODEML program in PAML (Yang, 2007).

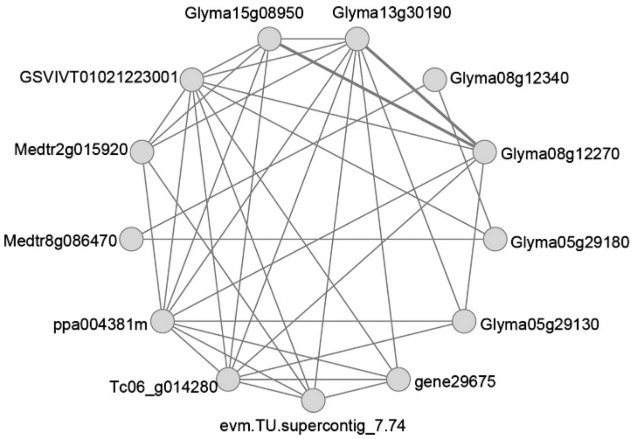


Figure 1 Syntenic network of P34. Each node represents a gene, and two genes with a syntenic relationship are linked by an edge (regular or bold lines). A regular line indicates a syntenic relationship in the PGDD, and a bold line reflects a true syntenic relationship but not identified in the PGDD probably due to pseudogenization (Glyma15g08950) or incorrect gene annotation (Glyma13g30190). The details of syntenic blocks can be searched in the PGDD.

The Ks values for each pair of genes with a syntenic relationship (Table 2) were used to interpret duplication nodes. The Ks values for three pairs of genes, *P34* vs Glyma05g29130, Glyma08g12340 vs Glyma05g29180 and Glyma15g08950 vs Glyma13g30190, ranged from 0.17 ± 0.13 to 0.19 ± 0.17. Their common ancestor nodes, marked by blue triangles (Figure 2), represent the duplication event that arose during the recent whole-genome duplication (WGD) of soybean, corresponding to the recent soybean lineage-specific paleo-tetraploidization, which occurred 13 million years ago (Lavin *et al.*, 2005; Bertioli *et al.*, 2009; Gill *et al.*, 2009; Schmutz *et al.*, 2010). The Ks value between *P34* and Glyma13g30190 is 0.80 ± 0.29 and its common ancestor node IV, marked by red triangles (Figure 2), represents a duplication event during the ancient WGD of soybean, corresponding to the early legume WGD, which occurred 59 million years ago (Herman *et al.*, 1990; Kalinski *et al.*, 1990, 1992; Lavin *et al.*, 2005). Notably, *P34* is as close to Glyma08g12340 as Glyma05g29130 is to Glyma05g29180, but no such paralogy of Glyma13g30190 has been identified. Thus, node III, marked by a red diamond (Figure 2), represents a tandem duplication event that occurred between the two rounds of WGD. After all duplication nodes were identified, the remaining nodes were considered natural speciation nodes. Comparing the gene tree and the species tree, the overall topologies of both are consistent (Figure 3). More importantly, the root of all the species in this study approximately locates at the basal of rosids (Figure 3), corresponding to the γ event, a whole-genome triplication event that is probably shared by all core eudicots (Jaillon *et al.*; 2007; Tang *et al.*, 2008a, b), so the phylogeny of *P34* in this study likely stems from one of the trifurcating branches formed by the γ gene triplication. Therefore, we determined the evolutionary history of *P34* approximately as far back as the root of rosids, and every clade of the phylogenetic tree has a biological interpretation.

Evolution of P34

According to the above phylogeny, there are three duplication events associated with the evolutionary path that leads to *P34* with the current function (Figure 2). The most recent duplication event generated three pseudogenes. Along with the other two duplication events, all the above genes were divided into groups A and B with

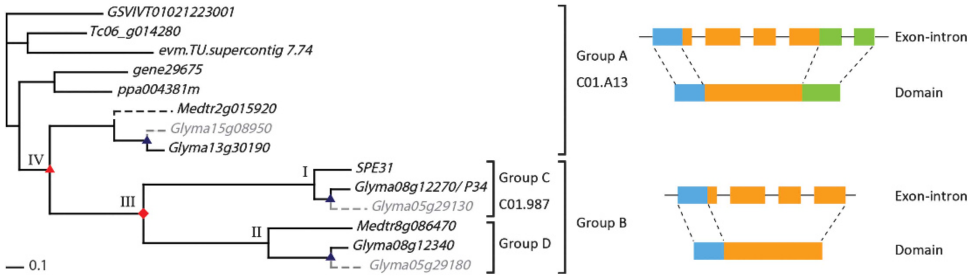


Figure 2 A phylogenetic tree (left) and topological structure (right) of 14 genes. Among these genes, 13 originate from the syntenic network, and the last gene is *SPE31*. The topological structure was estimated using the MrBayes program, and the branch lengths were computed using CODEML under a one ω model, M0. Pseudogenes are indicated in gray, and their branch lengths are not true. Medtr2g015920 was not completely sequenced, and its branch was also not true. Nodes marked with triangles or diamonds represent gene duplication events. The right-hand side contains gene exon–intron and protein domain structures. Papain-like proteins are first synthesized as inactive or less inactive precursors, including an N-terminal inhibitor region (blue), a mature region (orange), and at times, a C-terminal extension containing a granulin domain (green).

Table 2 Ks values of gene pairs in syntenic blocks

Locus 1	*Locus 2*	Ks		*Anchors*	*Locus 1*	*Locus 2*	Ks		*Anchors*
		Mean	*s.d.*				*Mean*	*s.d.*	
P34	Glyma05g29130	0.1730	0.1344	528	Glyma13g30190	gene29675	1.3006	0.3139	53
Glyma08g12340	Glyma05g29180	0.1730	0.1344	528	gene29675	Tc06_g014280	1.3344	0.3699	60
Glyma15g08950	Glyma13g30190	0.1924	0.1742	367	GSVIVT01021223001	Glyma05g29180	1.3504	0.3596	28
ppa004381m	gene29675	0.6364	0.2275	102	Tc06_g014280	evm.TU.supercontig_7.74	1.3599	0.4289	24
Glyma15g08950	*P34*	0.7053	0.2048	16	gene29675	evm.TU.supercontig_7.74	1.3877	0.3943	32
Medtr8g086470	Glyma08g12340	0.7500	0.4672	90	GSVIVT01021223001	Glyma13g30190	1.3978	0.4100	78
P34	Glyma13g30190	0.79525	0.2901	18	GSVIVT01021223001	*P34*	1.4085	0.5355	25
Glyma05g29130	Glyma13g30190	0.8030	0.2548	20	GSVIVT01021223001	evm.TU.supercontig_7.74	1.4221	0.5045	29
Medtr8g086470	Glyma05g29180	0.8063	0.4833	91	ppa004381m	evm.TU.supercontig_7.74	1.4778	0.6000	38
Glyma15g08950	Medtr2g015920	0.9212	0.5621	10	GSVIVT01021223001	Glyma15g08950	1.4824	0.4867	56
Glyma13g30190	Medtr2g015920	1.0159	0.6159	10	*P34*	Tc06_g014280	1.5333	0.3886	28
ppa004381m	GSVIVT01021223001	1.0398	0.3728	254	GSVIVT01021223001	Medtr2g015920	1.5532	0.4414	20
GSVIVT01021223001	Tc06_g014280	1.1483	0.3948	142	Glyma13g30190	Tc06_g014280	1.6108	0.4674	108
GSVIVT01021223001	gene29675	1.1526	0.3864	82	Glyma13g30190	evm.TU.supercontig_7.74	1.6171	0.5439	21
ppa004381m	Tc06_g014280	1.1822	0.3872	346	Glyma05g29130	Tc06_g014280	1.6475	0.5648	26
ppa004381m	*P34*	1.1935	0.3644	41	Glyma15g08950	Tc06_g014280	1.6484	0.5227	106
ppa004381m	Glyma05g29130	1.2153	0.3644	39	ppa004381m	Medtr2g015920	1.6527	0.5289	49
ppa004381m	Glyma13g30190	1.2548	0.3707	149	Medtr2g015920	evm.TU.supercontig_7.74	1.7275	0.2396	8
ppa004381m	Glyma15g08950	1.2566	0.3503	141					

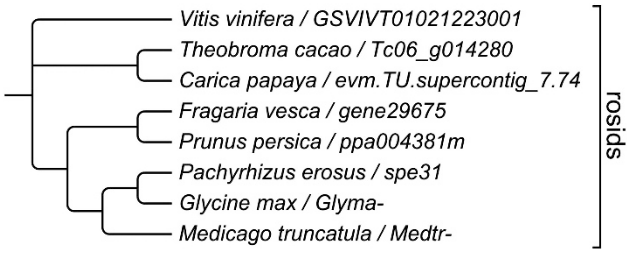

Vitis vinifera / GSVIVT01021223001
Theobroma cacao / Tc06_g014280
Carica papaya / evm.TU.supercontig_7.74
Fragaria vesca / gene29675
Prunus persica / ppa004381m
Pachyrhizus erosus / spe31
Glycine max / Glyma-
Medicago truncatula / Medtr-
rosids

Figure 3 Species tree and corresponding gene nomenclatures.

node IV; and the genes in group B were further divided into groups C and D with node III (Figure 2).

According to MEROPS, a database that classifies peptidases (Rawlings *et al.*, 2010), all proteins encoded by the above 10 genes belong to the papain family. Papain-like proteins are initially synthesized as inactive or less active precursors, and then the inhibitory N-terminal amino-acid sequences are cleaved (if a C-terminal granulin domain is present, it is also cleaved), generating mature proteins (Yamada e*t al*., 2001). As for the gene and protein domain structure, the genes in group A contain five exons and three domains (peptidase inhibitor, cysteine peptidase and extended granulin domains). However, the genes in group B have lost part of the fourth exon and the complete fifth exon and therefore lack the extended granulin domain (Figure 2). These data indicate that exon shuffling and following domain mutation occurred during the process of the functional transition. Therefore, P34 originates from a cysteine protease with an extended granulin domain that was lost by dismissing portions of exons during the early legume WGD.

Previous analyses of the SPE31 crystal structure (Figure 4a) revealed a series of sites responsible for the functional transition (Zhang *et al.*, 2006). There are two sites responsible for the catalytic activity loss: site 26 (referring to the alignment in Figure 4b), the replacement of catalytic cysteine (red in Figure 4) with a glycine, and site 173, the emergence of phenylalanine (blue in Figure 4), whose longer side chain stretches into the substrate-binding cleft and prevents SPE31 and P34 from exhibiting normal peptidase activity (Zhang *et al.*, 2006). According to the results in Figure 4b, reconstructed ancestral amino-acid sequences between nodes IV and I are different at sites 26 and 173, indicating that the catalytic activity could be affected after the legume-specific WGD. On the other hand, there are sites adapted for the new function as syringolide receptor. The asparagine at site 162 (purple in Figure 4), which was found to be glycosylated, could serve a role in recognizing the syringolide elicitor (Zhang *et al.*, 2006). The four residues of SPE31 at sites 22Q, 65Y, 151H and 171N (yellow in Figure 4) can bind directly to a natural peptide via hydrogen bonding, which is suggested as part of second messenger for SPE31 to transmit syringolide signal (Zhang *et al.*, 2006). Similarly, five residues at sites 22, 65, 151, 162 and 171, relative to new function, are different between nodes IV and I (Figure 4b). Therefore, almost all of the sites responsible for the loss of peptidase activity and the new function as a receptor have undergone nonsynonymous substitutions and have been fixed during the time period from node IV to node I. In other words, the transition of *P34* was largely accomplished in this time period because the new function could be further enhanced along the soybean lineage leading to *P34*.

In summary, gene duplications, exon shuffling and following granulin domain loss and some critical substitutions are associated with the evolution of the functional transition of *P34*.

Divergent evolution after gene duplication

Two divergence events occurred, which arose after two cycles of gene duplication. As shown in Figure 2, the total branch length of each gene in group B is much greater than that of each gene in group A, indicating that the genes in group B evolved with accelerated rates after the ancient WGD in soybeans. Furthermore, the characteristics of group A are different from those of group B in terms of intron–

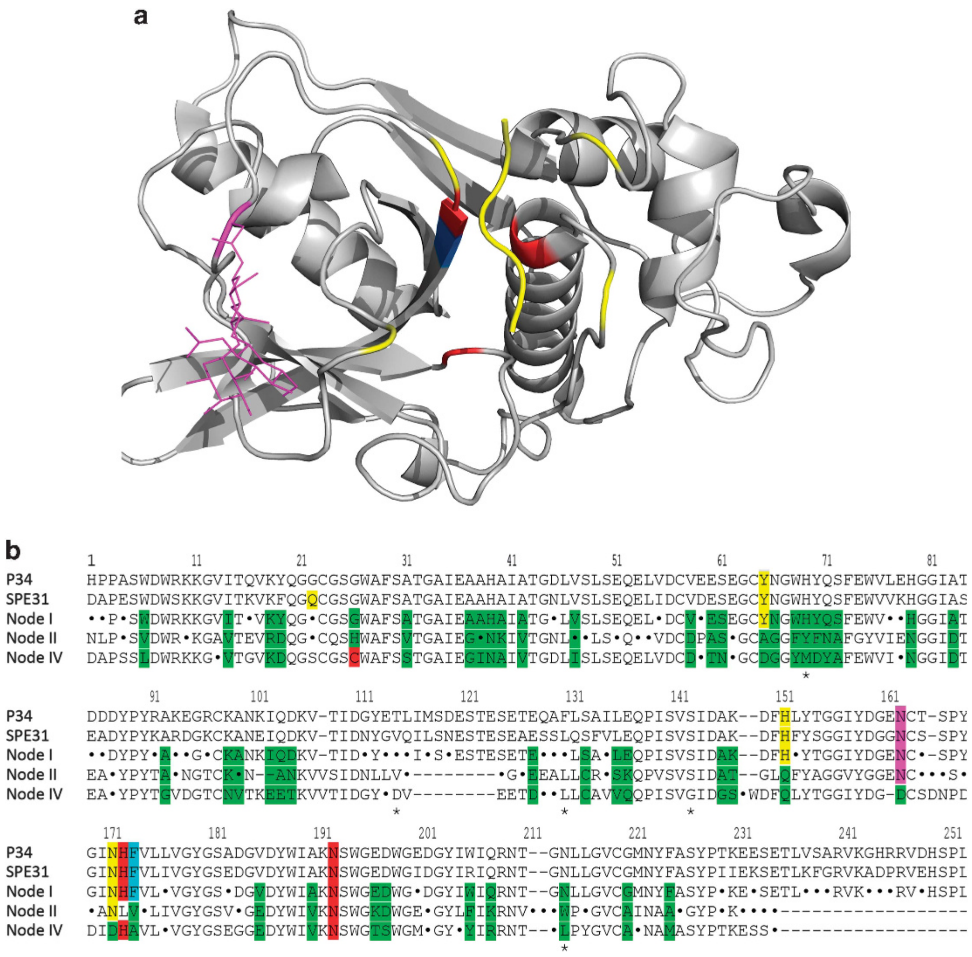

a
b
1 11 21 31 41 51 61 71 81
P34 HPPASWDWRKKGVITQVKYQGGCGSGWAFSATGAIEAAHAIATGDLVSLSEQELVDCVEESEGCYNGWHYQSFEWVLEHGGIAT
SPE31 DAPESWDWSKKGVITKVKFQGQCGSGWAFSATGAIEAAHAIATGNLVSLSEQELIDCVDESEGCYNGWHYQSFEWVVKHGGIAS
Node I ••P•SWDWRKKGVIT•VKYQG•CGSGWAFSATGAIEAAHAIATG•LVSLSEQEL•DCV•ESEGCYNGWHYQSFEWV••HGGIAT
Node II NLP•SVDWR•KGAVTEVRDQG•CQSHWAFSVTGAIEG•NKIVTGNL••LS•Q••VDCDPAS•GCAGGFYFNAFGYVIENGGIDT
Node IV DAPSSLDWRKKG•VTGVKDQGSCGSCWAFSSTGAIEGINAIVTGDLISLSEQELVDCD•TN•GCDGGYMDYAFEWVI•NGGIDT
*
91 101 111 121 131 141 151 161
P34 DDDYPYRAKEGRCKANKIQDKV-TIDGYETLIMSDESTESETEQAFLSAILEQPISVSIDAK--DFHLYTGGIYDGENCT-SPY
SPE31 EADYPYKARDGKCKANEIQDKV-TIDNYGVQILSNESTESEAESSLQSFVLEQPISVSIDAK--DFHFYSGGIYDGGNCS-SPY
Node I ••DYPY•A••G•CKANKIQDKV-TID•Y•••I•S•ESTESETE•••LSA•LEQPISVSIDAK--DFH•YTGGIYDGENC•-SPY
Node II EA•YPYTA•NGTCK•N--ANKVVSIDNLLV•---------•G•EEALLCR•SKQPVSVSIDAT--GLQFYAGGVYGGENC•••S•
Node IV EA•YPYTGVDGTCNVTKEETKVVTIDGY•DV--------EETD••LLCAVVQQPISVGIDGS•WDFQLYTGGIYDG-DCSDNPD
* * *
171 181 191 201 211 221 231 241 251
P34 GINHFVLLVGYGSADGVDYWIAKNSWGEDWGEDGYIWIQRNT--GNLLGVCGMNYFASYPTKEESETLVSARVKGHRRVDHSPI
SPE31 GINHFVLIVGYGSEDGVDYWIAKNSWGEDWGIDGYIRIQRNT--GNLLGVCGMNYFASYPIIEKSETLKFGRVKADPRVEHSPI
Node I GINHFVL•VGYGS•DGVDYWIAKNSWGEDWG•DGYIWIQRNT--GNLLGVCGMNYFASYP•KE•SETL•••RVK•••RV•HSPI
Node II •ANLV•LIVGYGSV•GEDYWIVKNSWGKDWGE•GYLFIKRNV•••WP•GVCAINAA•GYP•K••••-----------------
Node IV DIDHAVL•VGYGSEGGEDYWIVKNSWGTSWGM•GY•YIRRNT--•LPYGVCA•NAMASYPTKESS•-----------------
*

Figure 4 Structure of SPE31 (**a**) and alignment of P34, SPE31 and several significant ancestral nodes (**b**). The same color is assigned to each same residue. The color red is assigned to the conserved catalytic triad residues (Gly26 is also colored red); blue to Phe169 of SPE31, which disrupts catalytic activity; purple to the N-glycosylation site (Asn159 of SPE31) binding the three glycosyl residues (purple also); and yellow signifies the four sites that bind a natural peptide (yellow also) in SPE31. In addition, **a** was prepared using PyMOL (http://www.pymol.org/). In **b**, the sites in the ancestral nodes with maximum probabilities of <0.9 are indicated by dot. Green represents all of the different sites ranging from nodes I to VI. The asterisk indicates the positive selection sites identified by branch-site models.

exon structure, protein domain and the presence of a catalytic cysteine or other sites responsible for peptidase enzymatic activity. In other words, the ancient WGD of the soybean generated two different copies: one retained the original function as a cysteine peptidase; the other evolved a new function. This divergence was the first between groups A and B. Similarly, the total branch lengths between groups C and D are great, suggesting that groups C and D have lower sequence similarity, ~50%. According to MEROPS, groups A and C correspond to the C01.A13 and C01.987 categories, respectively. However, the appropriate class for group D has not been recorded. The characteristics of group D will be discussed below.

Additionally, Glyma08g12340 and Medtr8g086470 of group D lost the C-terminal extension domain. As for the amino-acid sequences in group D, the conserved catalytic cysteine at site 26 in group A was replaced by a histidine in group D, different from a glycine in group C. In the same way, the alanine at site 173 in group A was replaced by a valine in group D, which is different from the phenylalanine in group C that occupies the substrate-binding cleft of SPE31 and P34. The side-chain isopropyl group in valine is longer than the methyl group in alanine, but whether it is long enough to extend into the active cleft and obstruct substrate binding remains unclear. However, another member of catalytic triad (Cys–His–Asn), histidine at site 172, was uniquely replaced in group D. Hence, the genes Glyma08g12340 and Medtr8g086470 may have lost original peptidase activity. For the four amino acids probably responsible for binding second messengers, all residues but site 171 are different between groups C and D. In addition, Glyma08g12340 and Medtr8g086470 lack the insertion of eight residues near site 119 and have shorter C-terminal sequences than that of P34 and SPE31. Regarding gene expression patterns, *P34* is expressed highly in soybean seed tissue and reaches a peak in later seed development. However, Glyma08g12340 is expressed at relatively low levels primarily in young leaves, and also in flowers, pods and seeds (Figure 5). Different expression patterns indicate different regulatory elements and functions. Clearly, groups C and D may have evolved independently and divergently after the tandem duplication following the legume-specific WGD, and acquired novel functions different from each other and from other members of the papain family. Therefore, we could also deduce that Glyma08g12340 and Medtr8g086470 belong to another new group within the papain family.

Positive selection test

The ω value variations along all branches were estimated using the GABranch method (Pond and Frost, 2005). An increase in the ω value in group B was found. There are two possible explanations for this phenomenon. One explanation is positive selection, leading to the gain of a new function; another explanation is the relaxation of purifying selection for losing the original molecular function (Ohta, 1973; Nozawa, 2010). To distinguish these two possibilities, a positive selection test was performed in PAML (Yang, 2007). We first used site modeling to fit the data. As a result, M1a and M2a have nearly the same log likelihood value; however, the likelihood ratio statistic ($2\Delta \ln L$) for M7 and M8 is 7.424 ($P=0.0238$, $df=2$), supporting the presence of positive selection. We then used branch-site modeling to test for positive selection in three branches of interest: branch IV-III following the ancient gene duplication, branch III-I following the tandem gene duplication in group B and branch I-P34 leading to P34. The results indicated that all three branches displayed evidence of positive selection (Table 3). Using Bayes Empirical Bayes, the M8 site model suggests that no sites in these genes exist under positive selection, with a posterior probability of >95%. The branch-site model A suggests that ~4%, 14% and 9% of sites in branches IV-III, III-I and I-P34, respectively, are under positive selection. Furthermore, five amino-acid sites, 69H on branch IV-III, 142S and 214N on branch III-I, and 114T and 130F (referring to the alignment in Figure 4b) on branch I-P34, were revealed to be under positive selection along the foreground lineages using a cutoff posterior probability of 95%. Specifically, all of the sites predicted to be related to the functional change and were not found to be under positive selection, having posterior probabilities of >95%.

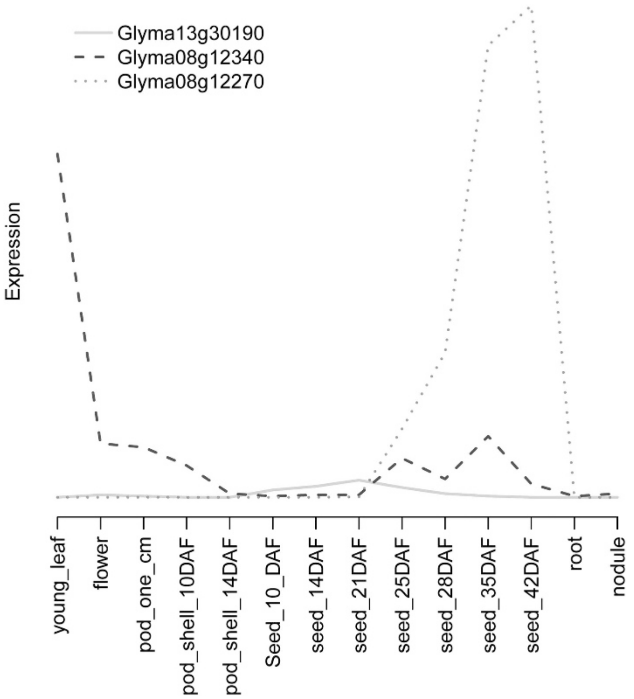


Figure 5 Expression patterns of P34 (Glyma08g12270), Glyma08g12340 and Glyma13g30190 in 14 tissues or phases. Expression data obtaining from next-generation sequencing (Severin *et al.*, 2010) was downloaded from Soybase (http://soybase.org/soyseq/). To compare expression pattern, the original expression of P34 was divided by 40. A full color version of this figure is available at the *Heredity* journal online.

DISCUSSION

Implementation of the functional transition

Combining previous studies and evolutionary inferences in this study, we attempt to understand the functional transition of *P34*, including the below critical issues: losing the original function, recognizing syringolide signal, interacting with second messengers, transmitting signal and the roles of the ancestral characteristics. The replacement of C and A at sites 26 and 173 in the alignment destroyed original peptidase activity of P34 (Zhang *et al.*, 2006), and also prevent their second messengers from being hydrolyzed. The emergence of the glycosylated residue at site 162 in the alignment (purple in Figure 4) probably enables P34 to obtain the ability to recognize the syringolide signal (Zhang *et al.*, 2006). As amino acid of ancestral node IV at site 162 is different and other papain-like proteins were not extracted through syringolide affinity column (Ji *et al.*, 1998), the ability to recognize syringolide may be a novel function of SPE31/P34, although amino acids of ancestral nodes II and I at site 162 are the same. As for the ability to interact with second messengers for P34, it may be a remnant of the role had by its ancestral protein, that is, these second messengers may be substrates of the ancestral peptidase. It should be noted that several residues responsible for the substrate specificity of the individual proteins were found to be located in the cleft (Choi *et al.*, 1999; Thakurta *et al.*, 2004; Wenig *et al.*, 2004) and the four sites probably responsible for binding second messenger (Zhang *et al.*, 2006) experienced individual amino-acid replacements. Therefore, the interaction may be novel and caused by changing specificity of P34. As for how the signal is transmitted, Okinaka *et al.* (2002) identified HPR as a potential second messenger of P34, and suggested that HPR binding with the complex of P34/syringolide induces hypersensitive response by inhibiting the activity of HPR in soybean. The location of glycosylated residue at site 162 is close to the cleft (Figure 4). Thus, we suggest that only when P34 is bound to both HPR and syringolide, the complex enters a proper conformation, and the interaction makes syringolide exactly stretch into the active location of HPR and inhibits its activity, eventually inducing hypersensitive response. In addition, some of the other mutations not identified by previous studies may be significant for the transition. Beside these mutations, the ancestral

Table 3 The parameters and statistical significances of branch-site tests

Foreground branch	*Parameters ($\omega_1=1$; $0<\omega_0<1$; ω_2 to be estimated)*					*LRT*		*Positive selected sites*
	e	p_1	p_2	ω_0	ω_2	$2\Delta \ln L$	P-*value*	
Branch IV-III	0.78	0.17	0.05	0.14	Infinity	5.48	0.019	69H*
Branch III-I	0.70	0.16	0.14	0.14	4.75	6.36	0.012	142S*, 214N*
Branch I-P34	0.75	0.15	0.10	0.14	Infinity	7.38	0.007	114T*, 130F*

Abbreviation: LRT, likelihood ratio test.
*$P>0.95$, where P is the Bayes Empirical Bayes posterior probability.

characteristic may be important for the novel function of P34. For example, P34, like its ancestral peptidase, is in the precursor form (Herman *et al.*, 1990; Kalinski *et al.*, 1990, 1992), being benefit for its normal biological function, because the N-terminal inhibitor region can obstruct the active cleft.

Aside from serving as the syringolide receptor in leaves (Ji *et al.*, 1998), P34 acts as a seed storage protein and an allergen (Herman *et al.*, 1990; Ogawa *et al.*, 1993). The reasons for this array of functions are as follows. First, the granulin domain loss may allow the mature protein to accumulate more quickly because granulin domain can slow the maturation of precursor (Yamada et *al.*, 2001). Second, when comparing the sequences upstream of *P34* and Glyma13g30190, *P34* has one more RY motif than Glyma13g30190 (motif search of promoter was performed in http://bioinformatics.cau.edu.cn/SFGD/). As the number of repeated RY motifs is essential for high seed-specific expression (Bäumlein *et al.*, 1992; Reidt, 2000), the additional RY motif may be one of the significant reasons that both the expressions of *P34* and Glyma13g30190 reach peaks in seeds; however, *P34* is expressed at higher levels and slightly later than Glyma13g30190 (Figure 5). Therefore, both losing the granulin domain and changing the promoter could lead to an abundant accumulation of P34 during seed development. The increase in protein content could be a key reason that P34 is an allergen, as the dosage of an allergen is an important factor in triggering allergic reactions.

Therefore, multiple gene duplications, exon shuffling and point mutation contribute to the functional transition of P34 from a cysteine peptidase to a syringolide receptor, a storage protein or an allergen together, and thus the evolution of P34 represents a typical and complex case of functional transition caused by combined mechanism.

What drives molecular evolution?

In previous sections, we have provided hypotheses about when and how P34 accomplished its functional transition. However, we cannot help but ask a classic question: what drives molecular evolution? In this study, we performed positive selection tests using site and branch-site models in PAML. Although the presence of positive selection is supported by likelihood ratio tests of individual models, there are some issues that remain to be considered. First, when site models M1a and M2a, but not M7 and M8, were used to test the positive selection, the presence of positive selection was rejected. The reasons for this difference are described below. Models M1a and M2a do not account for the variation of ω among sites, leading to inaccurate results and poor power. Although the variation of ω among sites is considered in models M7 and M8 by assuming a β distribution of ω, this assumption may result in false positives. Second, branch-site models assume positive selection acting on specific lineages and specific sites; this assumption seems reasonable, but the branch-site model can produce significant false-positive results even when there is no selection (Nozawa *et al.*, 2009). Third, we constructed ancestral coding sequences at interior nodes of the tree using the ANC-GENE program (Zhang *et al.*, 1998), but did not identify any branch with positive selection using the positive selection test in MEGA5. However, it may be inappropriate to use sequences with considerable divergence to construct ancestral coding sequences and test for selection along the branches. Fourth, the positive selection sites predicted by the site model are few and do not contain the sites predicted by previous experiments to relate to functional transition. Therefore, it is doubtful about the role of positive selection.

Indeed, we have confirmed the accelerated rate of evolution and higher ω values since the early duplication. In addition to positive selection, relaxation of purifying selection due to loss or diminishment of protein function can also increase ω values (Ohta, 1973; Nozawa, 2010). There is some evidence that supports this viewpoint. In the alignment (Figure 4b), we marked 50 sites that are different in nodes IV and I with green shades. These differences represent major amino-acid changes that clearly occurred and were fixed during the period between the evolution of nodes IV and I. However, the branch-site models only identified three positive selection sites in branches IV-III and III-I with $>95\%$ posterior probability; that is, the number of positive selection sites is relatively low. More importantly, the sites predicted to be responsible for the functional transition, losing original function and gaining new function, are not included in the computed positive selection sites, that is, these important sites could be selectively neutral and randomly fixed. Therefore, P34 might have evolved neutrally under the relaxation of purifying selection, with mutations occurring in coding regions and noncoding regions being fixed randomly. Eventually, these mutations caused the formation of a novel function for *P34* when some environments or the genetic backgrounds were altered.

The arguments presented above do not negate the role of gene duplication in the functional transition of *P34*; indeed, gene duplication is an important evolutionary force in many organisms, especially plants (Lynch and Conery, 2003; Jiao *et al.*, 2011). The evolutionary history of *P34* is highly associated with gene duplications, which may not only provide raw material for novel functions, but also lead to changes in gene regulatory regions and expression patterns through tandem duplication. For soybean, the two recent cycles of WGD correspond to the emergence of the legume and *Glycine* genus, respectively (Bertioli *et al.*, 2009; Gill *et al.*, 2009; Schmutz *et al.*, 2010). Thus, mining genes with considerable variation after the two cycles of WGD is important for understanding the characteristics specific to legumes. The evolution of P34 also represents such a typical case to study the contribution of gene duplication to the evolution of traits in legumes.

Syntenic network analyses

Syntenic network analyses depend on the following notions. If gene A is identified as being syntenic with genes B and C, although syntenic relationship between genes B and C cannot be found, we can conclude that genes A, B and C are homologous, because any two genes with syntenic relationship mean they are homologous, or come from a common ancestor. Taking this study as an example, if querying a gene *P34* in simple syntenic search, four homologous genes (Glyma05g29130, Tc06_g014280, GSVIVT01021223001 and ppa004381m) were identified. This result could not tell us the story in this study. However, if constructing syntenic network of *P34*, 13 homologous genes were found in Figure 1, offering a more interesting story in this study. Therefore, the creation of syntenic network can more effectively use syntenic relationship database for collecting sequences than simple synteny search.

However, two issues should be considered in syntenic network analyses. First, the time span of evolutionary history that can be traced by synteny information is limited because gene order may be disrupted as time elapses. The time span from the basal of rosids to now in this study may be sufficient for most comparative studies of molecular evolution. If not, we need to merge multiple homologous syntenic networks to reconstruct a gene family with longer time span and more genes. Second, syntenic network analyses might fail to identify genes derived by tandem duplication that are not shared by other syntenic blocks. For example, Glyma08g12280, a neighbor of *P34* and derived from a tandem duplication event, which is not found in other syntenic blocks in this study, was not included in the syntenic network in Figure 1. However, this tandem duplication gene does not affect our study because Glyma08g12280 is a pseudogene for frameshift mutations and no expression. Although syntenic network provided genes sufficient for studying the functional transition of P34, in practice, choosing additional genes of interest from results of sequence similarity search is recommended.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are grateful to the associate editor and three anonymous reviewers for their constructive comments and suggestions that significantly improve the presentation of the manuscript, and to Dr Hugo Zheng at the McGill University for help with improvements to the English text. This work was supported by the National Basic Research Program of China grant 2011CB109300; the National Natural Science Foundation of China grant 30971848; the Fundamental Research Funds for the Central Universities grants KYT201002 and KJ2011003; a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions and the Specialized Research Fund for the Doctoral Program of Higher Education grant 20100097110035.

Bäumlein H, Nagy I, Villarroel R, Inzé D, Wobus U (1992). Cis-analysis of a seed protein gene promoter: the conservative RY repeat CATGCATG within the legumin box is essential for tissue-specific expression of a legumin gene. *Plant J* **2**: 233–239.

Bertioli DJ, Moretzsohn MC, Madsen LH, Sandal N, Leal-Bertioli SCM, Guimarães PM *et al.* (2009). An analysis of synteny of Arachis with Lotus and Medicago sheds new light on the structure, stability and evolution of legume genomes. *BMC Genomics* **10**: 45.

Choi KH, Laursen RA, Allen KN (1999). The 2.1 Å structure of a cysteine protease with proline specificity from ginger rhizome, Zingiber officinale. *Biochemistry* **38**: 11624–11633.

Csardi G, Nepusz T (2006). The igraph software package for complex network research. *Inter Journal, Complex Systems* 1695. http://igraph.sf.net.

Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R *et al.* (2010). A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. *PLoS Genet* **6**: e1001255.

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

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

Herman EM, Melroy DL, Buckhout TJ (1990). Apparent processing of a soybean oil body protein accompanies the onset of oil mobilization. *Plant Physiol* **94**: 341–349.

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

Ji C, Boyd C, Slaymaker D, Okinaka Y, Takeuchi Y, Midland SL *et al.* (1998). Characterization of a 34-kDa soybean binding protein for the syringolide elicitors. *Proc Natl Acad Sci USA* **95**: 3306–3311.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Kalinski A, Weisemann JM, Matthews BF, Herman EM (1990). Molecular cloning of a protein associated with soybean seed oil bodies that is similar to thiol proteases of the papain family. *J Biol Chem* **265**: 13843–13848.

Kalinski A, Melroy DL, Dwivedi RS, Herman EM (1992). A soybean vacuolar protein (P34) related to thiol proteases is synthesized as a glycoprotein precursor during seed maturation. *J Biol Chem* **267**: 12068–12076.

Kamphuis I, Drenth J, Baker E (1985). Thiol proteases comparative studies based on the high-resolution structures of papain and actinidin, and on amino acid sequence information for cathepsins B and H, and stem bromelain. *J Mol Biol* **182**: 317–329.

Keen N, Buzzell R (1991). New disease resistance genes in soybean against Pseudomonas syringae pv glycinea: evidence that one of them interacts with a bacterial elicitor. *Theor Appl Genet* **81**: 133–138.

Lavin M, Herendeen PS, Wojciechowski MF (2005). Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst Biol* **54**: 575–594.

Long M, Betrán E, Thornton K, Wang W (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875.

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

Nozawa M, Suzuki Y, Nei M (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci USA* **106**: 6700–6705.

Nozawa M, Suzuki Y, Nei M (2010). Is positive selection responsible for the evolution of a duplicate UV-sensitive opsin gene in Heliconius butterflies? *Proc Natl Acad Sci USA* **107**: E96.

Ogawa T, Tsuji H, Bando N, Kitamura K, Zhu YL, Hirano H *et al.* (1993). Identification of the soybean allergenic protein, Gly m Bd 30K, with the soybean seed 34-kDa oil-body-associated protein. *Biosci Biotech Bioch* **57**: 1030–1033.

Ohta T (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246**: 96–98.

Okinaka Y, Yang CH, Herman E, Kinney A, Keen NT (2002). The P34 syringolide elicitor receptor interacts with a soybean photorespiration enzyme, NADH-dependent hydroxypyruvate reductase. *Mol Plant Microbe Interact* **15**: 1213–1218.

Patthy L (2008). *Protein Evolution*. 2nd edn. Wiley-Blackwell: Oxford, UK, pp 127–166.

Pond SLK, Frost SDW (2005). A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* **22**: 478–485.

Reidt W, Wohlfarth T, Ellerström M, Czihal A, Tewes A, Ezcurra I *et al.* (2000). Gene regulation during late embryogenesis: the RY motif of maturation specific gene promoters is a direct target of the FUS3 gene product. *Plant J* **21**: 401–408.

Rawlings ND, Barrett AJ, Bateman A (2010). MEROPS: the peptidase database. *Nucleic Acids Res* **38**: D227–D233.

Ronquist F, Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

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.

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD *et al.* (2010). RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol* **10**: 160.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008a). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**: 1944–1954.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008b). Synteny and collinearity in plant genomes. *Science* **320**: 486–488.

Thakurta PG, Biswas S, Chakrabarti C, Sundd M, Jagannadham MV, Dattagupta JK (2004). Structural basis of the unusual stability and substrate specificity of ervatamin C, a plant cysteine protease from *Ervatamia coronaria*. *Biochemistry* **43**: 1532–1540.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X *et al.* (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**: e49.

Wenig K, Chatwell L, von Pawel-Rammingen U, Björck L, Huber R, Sondermann P (2004). Structure of the streptococcal endopeptidase IdeS, a cysteine proteinase with strict specificity for IgG. *Proc Natl Acad Sci USA* **101**: 17371–17376.

Yamada K, Matsushima R, Nishimura M, Hara-Nishimura I (2001). A slow maturation of a cysteine protease with a granulin domain in the vacuoles of senescing Arabidopsis leaves. *Plant Physiol* **127**: 1626–1634.

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

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

Zhang J, Rosenberg HF, Nei M (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* **95**: 3708–3713.

Zhang M, Wei Z, Chang S, Teng M, Gong W (2006). Crystal structure of a papain-fold protein without the catalytic residue: A novel member in the cysteine proteinase family. *J Mol Biol* **358**: 97–105.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)