Theor Appl Genet (2002) 104:1132–1140
DOI 10.1007/s00122-002-0884-6

V. Beilinson · Z. Chen · R.C. Shoemaker
R.L. Fischer · R.B. Goldberg · N.C. Nielsen

Genomic organization of glycinin genes in soybean

Received: 20 April 2001 / Accepted: 23 August 2001 / Published online: 30 March 2002

Abstract Glycinin is the predominant seed storage protein in most soybean varieties. Previously, five major genes (designated *Gy1* to *Gy5*) encoding glycinin subunits have been described. In this report two new genes are identified and mapped: a glycinin pseudogene, *gy6*, and a functional gene, *Gy7*. Messenger RNA for the *gy6* pseudogene is not detected in developing seeds. While *Gy7* mRNA was present at the midmaturation stage of seed development in the soybean variety Resnik, the steady state amount of this message was at least an order of magnitude less-prevalent than the mRNA encoding each of the other five glycinin subunits. Even though the amino-acid sequence of the glycinin subunit G7 is related to the other five soybean 11S subunits, it does not fit into either the Group-1 (G1, G2, G3) or the Group-2 (G4, G5) glycinin subunit families. The *Gy7* gene is tandemly linked 3′ to *Gy3* on Linkage Group L (chromosome 19) of the public molecular linkage map. By contrast, the *gy6* gene occupies a locus downstream from *Gy2* on Linkage Group N (chromosome 3) in a region that is related to the position where *Gy7* is located on chromosome 19.

Keywords Soybean · Glycinin · Plant genes · Storage proteins

Communicated by P. Langridge

V. Beilinson
Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA

Z. Chen
Departments of Agronomy and Zoology, Iowa State University, Ames, IA 50011, USA

R.C. Shoemaker
United States Department of Agriculture,
Agricultural Research Service,
Departments of Agronomy and Zoology, Iowa State University, Ames, IA 50011, USA

R.L. Fischer
Department of Plant and Microbial Biology,
University of California, Berkeley, CA 94720, USA

R.B. Goldberg
Department of Molecular, Cell, and Developmental Biology,
University of California, Los Angeles, CA 90095-1606, USA

N.C. Nielsen (✉)
United States Department of Agriculture,
Agricultural Research Service, Department of Agronomy,
Purdue University, West Lafayette, IN 47907, USA
e-mail: nnielsen@purdue.edu
Tel.: 765-494-8057, Fax: 765-494-6508

Introduction

Glycinin, an 11S globulin, is the predominant seed storage protein in soybean. It accounts for over 50% of the total seed protein in most varieties. Because glycinin makes an important contribution to the nutritional quality of this economically important crop species, it has been characterized extensively. Glycinin can be extracted from seeds as a hexamer of about 350,000 Da using dilute salt solutions. To-date, five glycinin genes have been described, and they encode each of the subunits known to comprise this protein. The five genes can be arranged into two families based on the extent of identity they exhibit to one another (Nielsen et al. 1997). Glycinin genes *Gy1*, *Gy2* and *Gy3* form one group (Group-1), while the second consists of *Gy4* and *Gy5* (Group-2). Although subunits derived from both groups of genes contain suboptimal amounts of the nutritionally important sulfur-containing amino acids methionine and cysteine, *Gy4* and *Gy5* encode proteins that contain substantially fewer sulfur amino acids than those derived from the other three genes.

There have been previous efforts to describe the genetic inheritance and genomic organization of glycinin genes. The three Group-1 genes are organized into two chromosomal domains, each about 45 kb in length (Nielsen et al. 1989). *Gy1* and *Gy2* are located in a direct repeat in one domain, and are separated by about 3 kb. *Gy3* is located at a position in the second domain that corresponds to that of *Gy1* and *Gy2*. The two domains each contain at least five pairs of functionally homolo-

gous genes that are expressed in either embryos or leaves of the mature plant (Nielsen et al. 1989). By making use of appropriate DNA polymorphisms, Cho et al. (1989) demonstrated that the two chromosomal domains segregated independently, both from one another and from genetic loci containing either *Gy4* or *Gy5*. However, on the basis of a mutant identified by Kaizuma et al. (1990), it has also been suggested that *Gy1/2* and *Gy3* might be genetically linked (Kitamura 1993).

The mapping of the glycinin genes should clarify their linkage relationships and provide additional information about glycinin gene organization and evolution. In this regard, the results reported by Cho et al. (1989) were confirmed in part by genetic-mapping experiments in which *Gy4* and *Gy5* were mapped to Linkage Groups O and F, respectively (Diers et al. 1993; Chen and Shoemaker 1998). The objective of the experiments described in this report was to genetically map the three Group-1 glycinin genes. The results demonstrated that *Gy3* is found on Linkage Group L. However, a probe for *Gy1* and *Gy2* mapped to two independently segregating genetic loci. One locus was tightly linked to *Gy3* on Linkage Group L and the other was located on Linkage Group N. To characterize these loci, the nucleotide sequences for the chromosomal regions downstream from the *Gy3* gene on Linkage Group L, and downstream from the *Gy1* and *Gy2* genes on Linkage Group N, were determined. These sequences revealed that there were remnants of a glycinin gene, *gy6*, downstream from *Gy2* on Linkage Group N, and another functional glycinin gene, *Gy7*, at an equivalent position in Linkage Group L. These data provide new insights about the complexity of the family of 11S genes that encode soybean glycinin and the changes that have occurred in this family during its evolution.

Materials and methods

Mapping of glycinin genes

Three populations derived from interspecific and intraspecific crosses were used to construct genetic maps with the glycinin genes. Crossing a *Glycine max* breeding line (A81-356022) and a wild *Glycine soja* accession (PI468916) generated one F2 population, which consisted of 60 individuals. This population has been used to develop the public USDA-ARS/ISU molecular linkage map (Shoemaker et al. 1996). The second F2 population, which consisted of 94 lines, was derived from a cross between a shriveled seed mutant, T311, and the cultivar Keburi. Keburi contains a null allele of *Cgy1* (Kitamura et al. 1984). The third F2 population contained 92 lines, and was developed from the cross of T311 and the cultivar Raiden. Raiden contains a null allele of *Gy4* (Kitamura et al. 1984).

Genomic DNA isolation, restriction-enzyme digestion, electrophoresis, blotting, probe preparation and labeling, hybridization and membrane washing were conducted as described by Keim et al. (1989), except for the washing conditions for the *Gy3*-specific probe ('gy3'). For 'gy3', membranes were washed in 2 × SSC, 0.5% SDS for 25 min, then 1 × SSC, 0.5% SDS for 25 min, and finally 0.5 × SSC, 0.1% SDS for 20 min at 60 °C. Simple sequence repeat (SSR) markers were processed as described by Akkaya et al. (1995).

The public soybean molecular map derived from a cross between *G. max* and *G. soja* contains 810 markers and 25 linkage groups, and spans about 2,500 cM (Shoemaker et al. 1996). For the crosses between T311 and Keburi, and between T311 and Raiden, 248 markers covering all linkage groups at intervals of less than 20 cM were screened with five restriction enzymes (*Hin*dIII, *Eco*RI, *Eco*RV, *Dra*I and *Taq*I). Ninety and 87 polymorphic markers were mapped in the T311 × Keburi and T311 × Raiden populations, respectively (Chen and Shoemaker 1998). The Mapmaker program (Lander et al. 1987) was used to construct the linkage map. A LOD score of 3.0 was used as the lower limit for accepting linkage between two markers. The distances in cM between markers were calculated by the Haldane function.

Sequence determination of G* (*gy6*) and G*′ (*Gy7*)

DNA sequence analysis was carried out with an ALFexpress apparatus (Pharmacia) using an Amersham sequencing kit (Cat. no. RPN2538). Plasmid pG30H6.4 was used to obtain the sequence for *gy6*. This plasmid contains a *Hin*dIII–*Hin*dIII genomic DNA fragment of about 6.4 kb in length that includes the 3′ part of the *Gy2* gene on the 5′ end of the clone. The 3′ end of this plasmid downstream from *Gy2* contains about 1.5 kb of the *gy6* pseudogene. The 3′-end of plasmid pG6R3.2 (Nielsen et al. 1989) was used to sequence the chromosomal region downstream from the *Gy3* gene. Plasmid pG6R3.2 contains a 3.2-kb *Eco*RI–*Eco*RI genomic DNA fragment that had a 5′ portion of the *Gy7* gene. An oligo walking procedure was used to sequence the intergenic DNA 3′ from *Gy3*. When this analysis revealed that *Gy7* contained the 5′-end of a potential glycinin gene, the entire gene was obtained using a Promoter Finder Genomic Library (Clontech). The library was used for the polymerase chain reaction (PCR) with *Gy7* gene-specific and AP (adaptor–primer) oligonucleotides to obtain the 3′-part of the *Gy7* gene according to the manufacturer's instructions (user manual PT3042-1, Clontech). A full-length cDNA clone of *Gy7* was amplified from a Marathon (Clontech) cDNA library prepared from midmaturation Resnik soybeans. All amplified PCR DNA fragments were initially cloned into pCRII (Invitrogen).

Results

The Group-1 glycinin genes map to Linkage Groups L and N

DNA probes were generated to distinguish between the three Group-1 glycinin genes. This objective was complicated by the high degree of nucleotide sequence identity that exists among *Gy1*, *Gy2* and *Gy3* (Cho and Nielsen 1989; Nielsen et al. 1989; Sims and Goldberg 1989). As summarized in Fig. 1, a comparison of these nucleotide sequences revealed that *Gy3* regions 81–631 bp, 1,207–1,843 bp and 2,779–3,474 bp (Cho et al. 1989; Cho and Nielsen 1989) had 92, 85 and 86% identity, respectively, to the corresponding regions in *Gy1* (Sims and Goldberg 1989). Other regions had 40 to 45% identities. *Gy3* regions 187–631 bp and 2,766–3,326 bp were 90 and 94% identical to *Gy2*, respectively, although there were also other regions with less than 45% sequence identity. Based on these comparisons, the *Gy3* region between 2,417–2,695 bp, which was 41% identical to the corresponding regions in *Gy1* and *Gy2*, was amplified by PCR. This probe was designated 'gy3', because under highly stringent washing con-

ditions it hybridized specifically to *Gy3*, but not to *Gy1* and *Gy2* (data not shown). The *Gy3* region between 2,537–2,990 bp, which had 94 and 93% identity to *Gy1* and *Gy2*, was also amplified. Because this DNA probe recognized *Gy1*, *Gy2* and *Gy3* equally well (data not shown), it was designated 'gy123'. By comparing hybridization results obtained using the 'gy123' probe with those obtained with the 'gy3' probe, it was possible to determine which fragments corresponded to *Gy3* and which ones corresponded to *Gy1* plus *Gy2* (*Gy1/2*).

An F2 population was generated by crossing a *G. max* breeding line (A81-356022) with a wild *G. soja* accession (PI 468916), and progeny from this cross were used to develop the public USDA-ARS/ISU molecular linkage map (Shoemaker et al. 1996). To identify polymorphisms within this interspecific population, genomic DNAs from the two parents were digested with 14 restriction enzymes, blotted, and then hybridized sequentially using probes 'gy123' and 'gy3'. Figure 2 illustrates the DNA polymorphisms among *G. max* and *G. soja* that were identified using the restriction enzymes *Acc*I, *Bcl*I, *Eco*RI and *Ssp*I. These polymorphisms represented the *Gy3* gene, and were designated as markers *Gy3_Acc*, *Gy3_Bcl*, *Gy3_Eco*RI and *Gy3_Ssp*, respectively. By elimination, the polymorphic fragments produced by the restriction enzymes *Dra*I and *Eco*RV, corresponded to *Gy1* plus *Gy2*-related genes, and were called markers *Gy1/2_Dra* and *Gy1/2_Eco*RV, respectively.

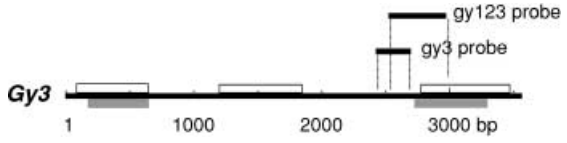

Fig. 1 Identity analysis of *Gy3* to *Gy1* and *Gy2*. The *long central bold line* in the figure designates the *Gy3* coding region. The *open boxes* above the central bold line identify highly related regions between *Gy3* and *Gy1*. The regions 81–631 bp, 1,207–1,843 bp and 2,779–3,474 bp have 92, 85 and 86% identity, respectively (Cho and Nielsen 1989; Nielsen et al. 1989; Sims and Goldberg 1989). The *shaded boxes* below the central bold line indicate highly homologous regions between *Gy3* and *Gy2*. The regions 187–631 bp and 2,766–3,326 bp have 90 and 94% identity, respectively. The 'gy123' probe was amplified by PCR from the *Gy3* 2,417–2,695-bp region that has 41% identity to *Gy1* and *Gy2* (Cho and Nielsen 1989; Nielsen et al. 1989; Sims and Goldberg 1989). The 'gy123' probe was amplified from the *Gy3* 2,537–2,990-bp region, and has 94 and 93% identity to *Gy1* and *Gy2* at their 3′ ends, respectively (Cho and Nielsen 1989; Nielsen et al. 1989; Sims and Goldberg 1989)

Figure 3 shows polymorphic fragments identified in F2 populations from crosses between cultivars or varieties of *G. max*. One of these populations was derived from an intraspecific cross between T311 and Keburi, while the other derived from a cross between T311 and Raiden. Polymorphic fragments between T311 and Keburi were produced by the restriction enzymes *Hin*dIII, *Eco*RV and DraI, and were due to *Gy3* sequences. These were designated *Gy3_Hind*, *Gy3_Eco*RV and *Gy3_Dra*, respectively. The polymorphic fragment between T311 and Keburi with the enzyme *Taq*I identified *Gy1/2*, and was called *Gy1/2_Taq*. The polymorphic fragments between T311 and Raiden with *Eco*RV was also a marker for *Gy1* plus *Gy2*, and was designated *Gy1/2_Eco*RV.

The Group-1 genes were mapped initially using a segregating population derived from the interspecific cross described earlier in this communication. As indicated in Fig. 4A, each of the four *Gy3* markers, *Gy3_Acc*, *Gy3_Bcl*, *Gy3_Eco*RI and *Gy3_Ssp*, mapped to the same

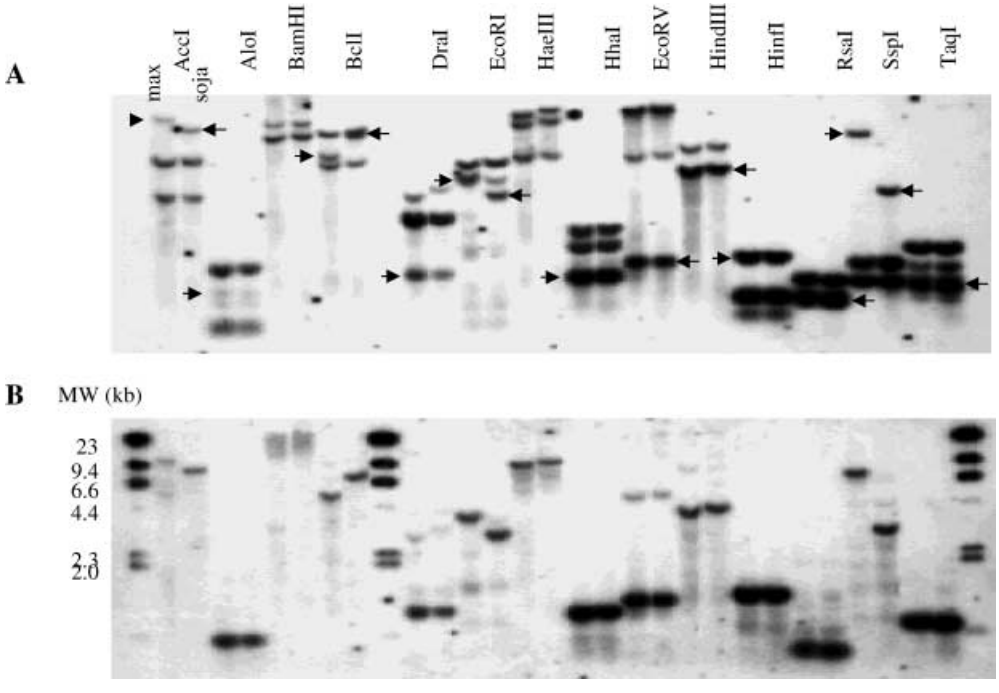

Fig. 2A, B Blot of *G. max* and *G. soja* genomic DNA. The same membrane was hybridized sequentially with the 'gy123' (**A**) and 'gy3' (**B**) probes. The restriction enzymes used are indicated at the top of Panel **A**. For each enzyme, the first lane was *G. max* and the second *G. soja*. The *arrows* in Panel **A** identify the *Gy3* fragments that correspond to the fragments on Panel **B**

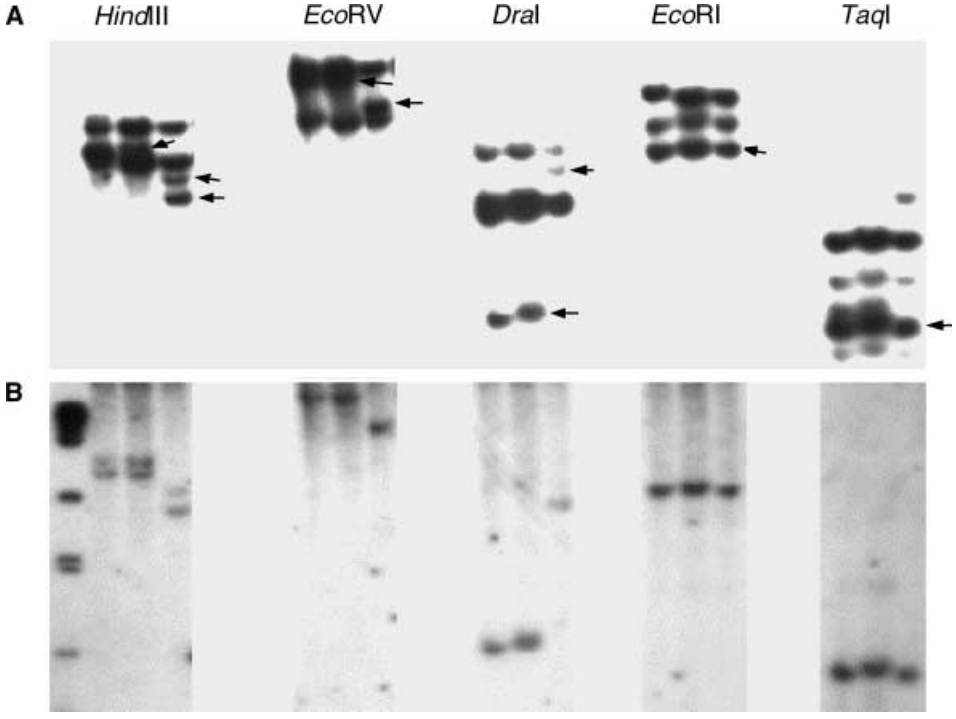

Fig. 3A, B DNA blot of T311, Keburi and Raiden. The same membrane was sequentially hybridized with the 'gy123' (**A**) and 'gy3' (**B**) probes. The five restriction enzymes used are indicated at the top of Panel **A**. For each enzyme, the order of lanes was T311, Raiden and Keburi. The arrows in Panel **A** identify *Gy3* fragments that correspond to the fragments on Panel **B**

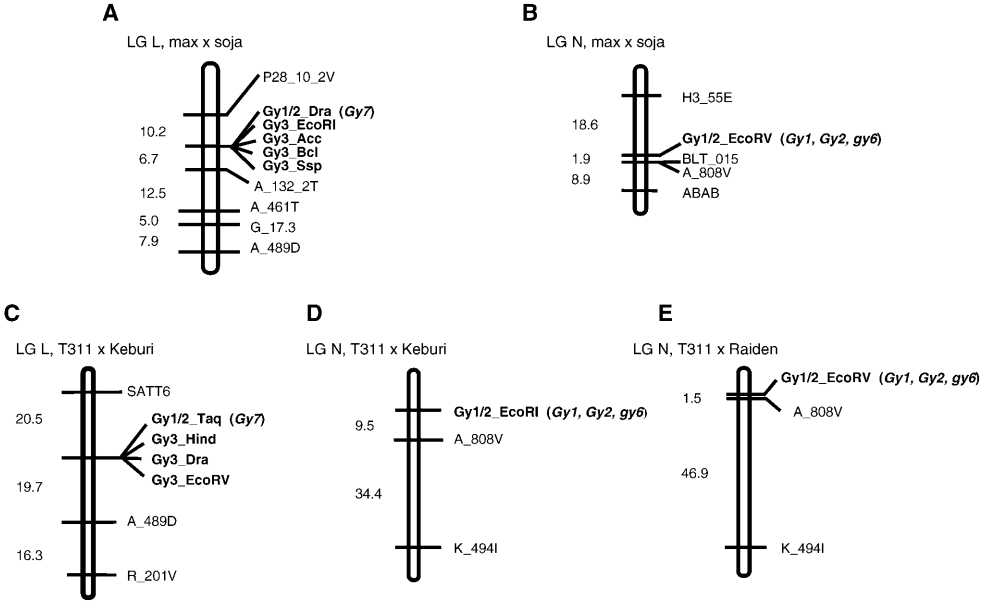

Fig. 4A–E Linkage maps of *Gy1*/*Gy2* and *Gy3*. The distances in centiMorgan (cM) were computed using the Mapmaker program (Lander et al. 1987). Panels **A** and **B** were constructed for the *G. max* × *G. soja* population. Panels **C** and **D** were developed for the T311 × Keburi population. E is for the T311 × Raiden population. DNA sequence analysis revealed that *Gy1*/*2_Dra* and *Gy1*/*2_Taq* were due to *Gy7*, while markers *Gy1*/*2_Eco*RV and *Gy1*/*2_Eco*RI were due to *gy6*. *LG* refers to linkage group

position on Linkage Group L. This locus was tightly linked to markers A_132_2T and A_461T, with LOD values of 8.7 and 10.1, respectively. The *Gy1* plus *Gy2* marker, *Gy1*/*2_Dra*, co-segregated with the four *Gy3* markers on Linkage Group L. By contrast, as indicated in Fig. 4B, the *Gy1* plus *Gy2* marker, *Gy1*/*2_Eco*RV, mapped to Linkage Group N. It was tightly linked to markers BLT_015 and A_808V with LOD values of 11.4 and 11.5, respectively. Thus, DNA from both linkage groups hybridized with the 'gy123' probe.

These map positions were confirmed using segregating populations from the T311 × Keburi and T311 × Raiden crosses. As indicated in Fig. 4C, the three *Gy3* markers, *Gy3_Hind*, *Gy3_Eco*RV and *Gy3_Dra*, all mapped to Linkage Group L in the T311 × Keburi population. They were linked to RFLP markers A_489D, R_210V, and to the SSR marker SATT6, with LOD values of 6.0, 4.0 and 4.1, respectively. Therefore, the map position of *Gy3* identified using the T311 × Keburi population coincided with the one mapped using segregants from the *G. max* × *G. soja* cross (i.e., compare Fig. 4A and C). *Gy1*/*2_Taq*, a *Gy1*/*2* marker used for the T311 × Keburi population, cosegregated with *Gy3*, and mapped to Linkage Group L (Fig. 4C). *Gy1*/*2_Eco*RI, the other

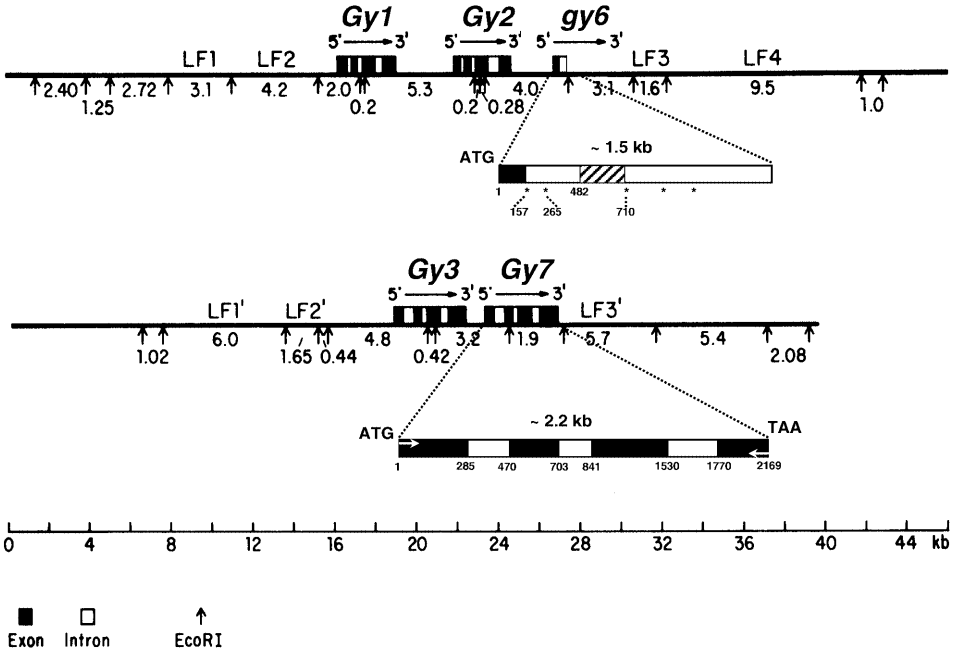

Fig. 5 Molecular maps of two related chromosomal domains that contain glycinin genes. The domain that contains *Gy1* and *Gy2* is on Linkage Group N (Shoemaker et al. 1996) of the public molecular-linkage map. The glycinin genes are flanked on each side by genes expressed in leaves (*LF*). A pseudogene, designated *gy6*, is tandemly linked to *Gy2*, and begins with a putative start codon that is located 2,271-bp 3′ from the TAG stop codon of *Gy2*. About 1.5 kb of DNA following the putative start codon was sequenced. The predicted translation product from this gene terminates after 52 codons (*solid black part* of coding region). While an open reading frame was identified further into the sequence (*cross-hatched region* of insert), it was not homologous with other glycinin sequences. Multiple stop codons were encountered downstream from the large ORF. *gy6* occupies the position previously identified as G* (Nielsen et al. 1989). The second chromosomal domain contains *Gy3* and is on Linkage Group L of the public molecular linkage map. A new functional glycinin gene, denoted *Gy7*, is tandemly linked to *Gy3*. Its putative start codon is 1,370-bp 3′ from the TAG stop codon of *Gy3*. The four exons in *Gy7* are found in the same relative positions in the new glycinin gene as the equivalent structures of the other glycinin genes (indicated by *solid black lines*). *Gy7* occupies a region previously referred to as G*′ (Nielsen et al. 1989). *White arrows* denote DNA sequences used as oligonucleotide primers to amplify Gy7 cDNA by PCR. *Vertical arrows* indicate *Eco*RI cleavage sites, and *numbers* between these arrows identify the size of the fragment that is generated. The scale at the bottom of figure indicates distance in kb. *Open boxes* are introns and *closed boxes* exons. Expanded inserts below chromosomal fragments summarize structures of the *gy6* pseudogene and *Gy7* functional gene, respectively, between ATG start and chain-termination codons. *Numbers* below inserts refer to nucleotides, where numbering starts with the ATG start-codon. *Asterisks* signify chain-termination codons embedded in the nucleotide sequences

Gy1/2 polymorphism marker used for T311 × Keburi population, was tightly linked to A_808V and K_494I on Linkage Group N with LOD values of 8.2 and 3.0 respectively (Fig. 4D). Similarly, in the T311 × Raiden population, *Gy1/2_Eco*RV was linked to A_808V and K_494I on Linkage Group N with LOD values of 16.0 and 3.2, respectively (Fig. 4E).

In summary, these data identify two regions on the molecular linkage map for the soybean genome, one on Linkage Group L and the other on N. The map position of *Gy3* on Linkage Group L was confirmed with seven markers in two independent soybean populations. A *Gy1/2* region on Linkage Group L was confirmed with two markers in two populations, and the position of the *Gy1/2* region on Linkage Group N was confirmed with three markers in three populations. Although the map distances between *Gy1/2* and the flanking markers on Linkage Group L and between *Gy3* and flanking markers on Linkage Group N varied slightly, the order of the genes with the flanking markers was invariant among all three segregating soybean populations.

Why do two chromosomal regions hybridize with the ‘gy123’ probe?

In a previous report (Nielsen et al. 1989), the three Group-1 glycinin genes were found to be located in two related chromosomal domains that each encompass about 45 kb of DNA. The basic features of the molecular map describing this organization are presented in Fig. 5. *Gy1* and *Gy2* are arranged in a direct tandem repeat, and are flanked by several leaf genes. *Gy3* occupies a position analogous to *Gy1* and *Gy2* in the second chromosomal domain. Immediately 3′ from both *Gy2* and *Gy3*, DNA regions were identified that hybridized with glycinin-specific probes and appeared to produce mRNA of an appropriate size to encode a glycinin subunit. The two regions were designated G* and G*′ in the original report (Nielsen et al. 1989). Because DNA within G* and G*′ could potentially hybridize with the ‘gy123’ probe, and thereby account for the results from the genetic studies described earlier in this paper (Fig. 4), the DNA sequences 3′ from both *Gy2* and *Gy3* were determined.

Table 1 Characteristics of glycinin subunits

Glycinin subunit	Database accession number	Calculated molecular weight	Calculated isoelectric point (IEP)	Acidic polypeptide		Basic polypeptide		Number of Met per subunit	Number of Cys per subunit
				Molecular weight	IEP	Molecular weight	IEP		
G1	CAA33215	53,624	5.78	33,161	5.37	20,481	7.99	6	8
G2	CAA33216	52,445	5.37	32,136	5.03	20,327	7.97	7	8
G3	AAB23211	52,311	5.22	31,718	5.09	20,611	5.69	5	8
G4	AAB23212	61,294	5.33	40,687	4.72	20,626	9.90	2	6
G5	FWSYG3	55,424	5.53	36,392	4.97	19,050	9.51	4	6
G7	AF319777	57,677	6.37	37,335	5.90	20,359	9.41	6	11

gy6 encodes a glycinin pseudogene, but *Gy7* encodes a new glycinin subunit

To understand the relationship between G* and G*′ and the 'gy123' probe, the nucleotide sequences of the chromosomal regions 3′ from *Gy2* and *Gy3* were both determined. In the case of the G* nucleotide sequence, nearly 4 kb of DNA after the chain termination codon of *Gy2* was analyzed. While the entire sequence can be accessed in GenBank (AF319775), the essential features of this sequence are outlined in Fig. 5. A putative start codon was found 2,271 bp after the TAG stop codon of *Gy2*. It was followed by 52 codons that encoded a protein homologous to other glycinin subunits, and then a stop codon was encountered. The first exons from other glycinin genes encode more than 90 amino acids, so it appeared that the first exon in G* was truncated. As indicated by the hatched box in Fig. 5, another open reading frame was located in the sequence of G* that followed. However, the protein sequence predicted from this open reading frame did not have appreciable homology with any of the glycinin subunits. Following this extended open reading frame, multiple stop codons occurred. Attempts to generate an amplification product by PCR using a cDNA library prepared from Resnik seeds at midmaturation were unsuccessful (data not shown). Based on these data, we concluded that G* harbors the remnants of a glycinin gene that is apparently unable to produce a message that can be amplified by PCR. To facilitate discussion, we designated the G* pseudogene as *gy6*.

To characterize the G*′ chromosomal region that was 3′ from glycinin *Gy3*, approximately 4 kb of nucleotide sequence was determined. The important features of this sequence are summarized in Fig. 5, and the entire sequence has been deposited in a public database (GenBank AF319776). The nucleotide-sequence data permitted identification of a functional glycinin gene that has not previously been described. The ATG start codon of the new gene, designated as *Gy7*, was located 1,370 bp downstream from the TAG stop codon of *Gy3*. Part of the *Gy7* nucleotide sequence had 88% identity with the 'gy123' probe. This observation accounts for the apparent anomaly in the genetic linkage data described earlier in this paper in which the Gy1/2 probe was mapped to both Linkage Groups L and N (Fig. 4).

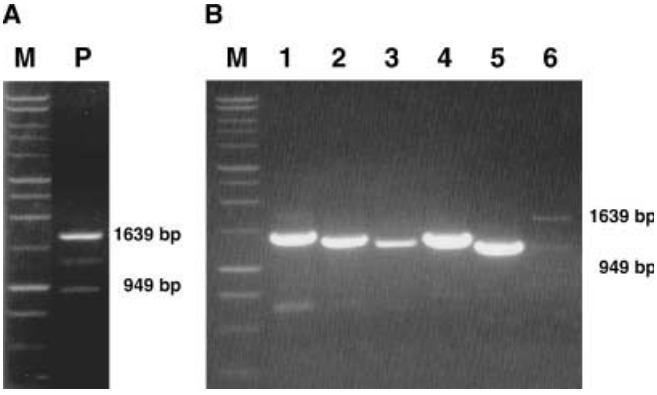

Fig. 6A, B PCR amplification products from glycinin cDNAs. Panel **A**: *M* DNA size marker, *P* PCR using Marathon cDNA as a template and *Gy7* specific oligonucleotides (5′-caccatgtttaaccattctgcgctccat and 5′-ttacatggtgacaatgaggggatttggag). Panel **B**: *M* DNA size marker, *1–6* Products from PCR carried out using soybean cDNA from developing seeds as a template and oligonucleotide pairs that specifically amplify each glycinin cDNA (see Materials and methods for details). *Lanes 1 to 5* Gy1 to Gy5 cDNAs, respectively. *Lane 6* Gy7 cDNA

Thus, genetic markers *Gy1/2_Dra* and *Gy1/2_Taq* corresponded to *Gy7* on Linkage Group L, while the rest of the markers identified by the gy1/2 probe are associated with the glycinin genes on Linkage Group N (Fig. 4).

Examination of *Gy7* revealed that the positions of the three introns and four exons in this new gene were similar to those found in the other glycinin genes (Nielsen et al. 1989). The size of the preproprotein encoded by the gene was predicted to be 57,677 daltons, and had a calculated isoelectric point of 6.37 (Table 1). After co- and post-translational processing, the preproprotein encoded by the new glycinin gene would yield an acidic chain of about 37 kDa and a basic chain of about 20 kDa. The NH_2-terminal amino acid of the basic chain was valine rather than glycine.

To confirm that *Gy7* was expressed, PCR was performed using a Marathon cDNA library prepared using mRNA isolated from seeds at the midmaturation stage of development. The positions of the DNA primers used for this amplification are identified in Fig. 5. As shown in Fig. 6A, three amplification products were obtained. After it was cloned, the nucleotide sequence of the largest and most-prevalent amplification product was deter-

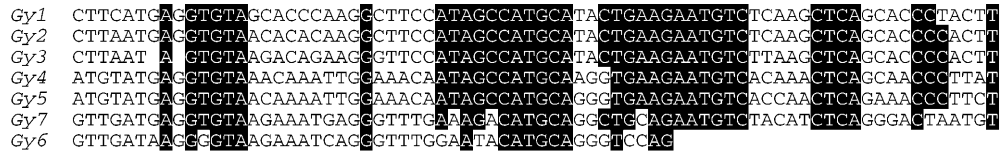

Fig. 7 Alignment of legumin boxes in the promoter regions of the seven glycinin genes from soybean. The conserved nucleotides are *highlighted*. Those from *gy6* and *Gy7* are the most divergent

mined and shown to correspond to a full-length transcript derived from *Gy7* (GenBank AF319777). A second product, which accounted for an estimated 30% of the total amplified DNA, yielded a nucleotide sequence in which DNA encoding the third exon was absent. This result implied that an appreciable portion of mRNA from *Gy7* was misspliced. The third minor PCR product produced a nucleotide sequence of unknown origin, that was unrelated to *Gy7* (data not shown). Together these data demonstrate that *Gy7* is indeed expressed.

We addressed the question about the prevalence of *Gy7* mRNA at the midmaturation stage of seed development. For this purpose, PCR was performed using Marathon cDNAs prepared from the soybean variety Resnik and primers specific for each of the glycinin genes. Results from these experiments are shown in Fig. 6B. The results revealed that *Gy7* transcripts were present in amounts that were several orders of magnitude lower than that of transcripts for the other five glycinin genes. Thus, while *Gy7* mRNA is accumulated, it accounts for a small proportion of the total glycinin messages at seed midmaturation, at least in the case of the soybean variety Resnik. This conclusion is consistent with results of experiments designed to detect the G7 protein. When mixtures of either the acidic or basic chains from glycinin were purified (Staswick et al. 1984), and the mixtures analyzed by Edman degradation, an amino-acid sequence due to G7 could not be distinguished from the other glycinin sequences in the mixtures. Thus, if present, the G7 protein must account for less than 5% of the total protein in the samples analyzed.

Discussion

Gy1 and *Gy2* segregate independently from *Gy3*

The primary goal of this research was to locate the glycinin genes in the soybean genetic linkage map. When these experiments were initiated, five glycinin genes had been identified (Nielsen et al. 1989). As a result of the experiments reported here, another functional glycinin gene, *Gy7*, and a pseudogene, *gy6*, were discovered. Prior to the start of these experiments, two of the genes, *Gy4* and *Gy5*, had already been assigned to Linkage Groups O and F, respectively (Diers et al. 1993; Chen and Shoemaker 1998). Linkage Group F is on chromosome 13 (Cregan et al. 2001), while Linkage Group O has not yet been assigned to a chromosome. Based on the present experiments, the remaining five glycinin genes can be assigned to genetic loci. *Gy1*, *Gy2* and *gy6*, which are tandemly linked one after another respectively, are on Linkage Group N, while *Gy3* and *Gy7* are found linked to one another on Linkage Group L. The identification of two linkage groups that contain *Gy1* plus *Gy2* and *Gy3* confirm studies by Cho et al. (1989), who demonstrated that RFLPs associated with these genes underwent independent genetic segregation.

Kitamura (1993) reached the opposite conclusion; that these three Group-1 glycinin genes were linked. His conclusion was based on genetic characteristics of a mutant soybean line induced by γ-ray irradiation. The mutant line lacked each of the Group-1 subunits as determined by SDS-PAGE, and this trait segregated genetically as if controlled by a single recessive allele (Kaizuma et al. 1990). Thus, the data supporting the conclusion that the Group-1 glycinin genes are linked relies on the accumulation of a gene product rather than on identification of the glycinin genes themselves. The RFLPs associated with Linkage Group L reside on chromosome 19 (Garner et al. 2001), while those on Linkage Group N are on chromosome 3 (Lee et al. 2000). Because the RFLPs marking glycinin genes *Gy1*, *Gy2*, *Gy3*, *gy6* and *Gy7* are distributed among two different chromosomes, they cannot all be linked genetically. A more-likely possibility to explain the results of Kaizuma et al. (1990) is that the γ-irradiation-induced mutation uncovered a trans-acting modifier gene. Modifier genes that modulate seed-storage protein accumulation have been described previously (Lopes et al. 1995).

Gy7 is poorly expressed

Our data reveal that the steady state amount of mRNA encoding *Gy7* at seed midmaturation is an order of magnitude less than the mRNAs encoding the five other glycinin subunits (Fig. 6B), at least in the soybean variety Resnik. Purified acidic chains of glycinin subunits from Resnik are unavailable, but we were unable to detect protein due to this subunit in preparations of acidic subunits purified from the breeding line CX635-1-1-1. If G7 was present among purified glycinin acidic polypeptides in the preparations from CX635-1-1-1 that were analyzed, it must have accounted for less than 5% of the total protein. As indicated in Fig. 7, a potential explanation for this observation was discovered when the legumin box in the *Gy7* nucleotide sequence was compared with the equivalent regions in the promoters from the other 11S storage protein genes. This region has an important regulatory function during the expression of seed-storage protein genes (Bäumlein et al. 1986; Dickinson et al.

1988). It is evident that the two central highly conserved regions of the legumin box are aberrant in both *gy6* and *Gy7* (Fig. 7). Whereas these legumin box regions are highly mutated in *gy6*, those in *Gy7* are less disturbed (Fig. 7). Most changes in the legumin box lead to a severe down-regulation in gene expression and the accumulation of glycinin gene products in seeds of transgenic plants (Bäumlein et al. 1992; Lelievre et al. 1992). Perhaps lines can be identified within the extensive soybean germplasm in which *Gy7* is more highly expressed than in Resnik.

Gy7 encodes a new type of glycinin subunit

Glycinins are members of the ancient cupin superfamily of proteins (Dunwell et al. 2000). The term cupin refers to a β-barrel structural motif found in proteins from organisms that range evolutionarily from primitive prokaryotes to eukaryotes. Both single-domain cupins and two-domain bicupins exist, the latter presumably arising as a result of gene duplication. The 11S and 7S storage proteins, together with the sucrose binding protein, are examples of bicupins found in seeds of higher plants, while germins and germin-like proteins contain single cupin motifs. Shutov and Bäumlein (1999) have presented detailed analysis of the evolutionary relationships among these proteins.

Structural properties of the six known glycinin subunits are compared in Table 1. G7 is the second largest of the glycinin subunits that have been identified. It is slightly larger than G5, but smaller than G4, and is 3 to 5 kDa larger than the Group-1 subunits. As with the two Group-2 subunits, the increased size of G7 compared to the three Group-1 subunits is due to a larger acidic chain. The increase in size is due to a large insertion toward its COOH-terminal just prior to the conserved post-translational processing site (NV in G7). This region is highly variable among all 11S subunits. That the conserved post-translational cleavage site is NV rather than NG has precedent. Aberrant cleavage sites have been described previously in subunits from pea (March et al. 1988), gingko (Arahira and Fukazawa 1994) and pine (Allona et al. 1992). Deviations from the NG cleavage site usually present among the 11S subunits in angiosperms are widespread in the Taxodiaceae, Cupressaceae and Taxaceae (Häger and Fischer 1999). The asparaginyl endopeptidase that carries out post-translational modification has an absolute requirement for asparagine on the N-terminal side of the peptide bond that is cleaved, but is far-less discriminating in its choice of amino acids on the C-terminal side (Jung et al. 1998).

Figure 8 presents a distance-matrix diagram of the glycinin subunits that was generated using the method of Kimura (1983). Genetic distances between glycinin subunits are presented as the number of nucleotide substitutions per 100 codons that would be required to convert one protein subunit to the other. The data indicate that G1, G2 and G3 comprise one family, G4 and G5 a second, and G7 a third (i.e., Group-3). Whereas G4 and G5 are nearly 90% identical with one another when the hypervariable region is excluded from the comparison, and less than 50% identical with the three Group-1 subunits, G7 is only about 45% identical to the Group-2 proglycinin subunits and less than 40% identical with the Group-1 subunits. This result suggests that *Gy7* followed a different evolutionary pathway than either the Group-I (*Gy1*, *Gy2*, *Gy3*) or Group-II (*Gy4*, *Gy5*) glycinin genes.

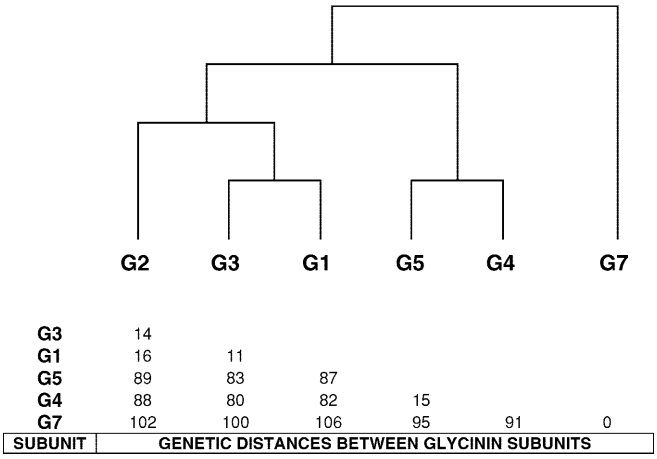

Fig. 8 Phylogenic tree of glycinin subunits. Pairwise distances shown at the bottom of the figure are the calculated number of nucleotide substitutions per 100 codons

The organization of glycinin genes in the soybean genome reflects the complex evolutionary history of this species. Whereas *Gy4* and *Gy5* are each found as single genes, probably on different chromosomes, the other five glycinin genes are found in two large related regions on chromosomes 3 and 19. These regions are either homoeologous as a result of an ancient tetraploidization event (i.e., bringing chromosomes 3 and 19 into the same nucleus) or result from a large segmental duplication. The extent of relatedness between these two chromosomes can be viewed at SoyBase in Linkage Groups L and N (http://129.186.26.94).

It seems likely that the evolution of the three main groups of glycinin genes predates the appearance of soybean as a species. Indeed, as discussed elsewhere (Nielsen et al. 1997), distinct groups of genes that correspond to the Group-1 and -2 genes are evident in other legumes. The origin of the relatedness among the various Group-1 and -3 genes remains an open question and could predate the origin of soybean as a species. It could also reflect either global or segmental duplication as part of the presumed tetraploidization event that occurred during soybean evolution (Shoemaker et al. 1996). Events that involved unequal crossing-over could also have played a role during the evolution of the related regions that contain glycinin genes. This phenomenon could be responsible for the mutations accumulated in *gy6* and *Gy7* noted in this study. It might also have resulted in the appearance of *Gy1* and *Gy2* in tandem linkage, although gene duplication is another possibility.

Answers to these questions should eventually result from studies that compare the organization of soybean chromosomes with those from other legumes.

Acknowledgement This work was supported in part by grants from the United Soybean Board to RCS and NCN.

References

Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci 35:1439–1445

Allona I, Casado R, Aragoncillo C (1992) Seed storage proteins from *Pinus pinaster* Ait: homology of major components with 11S proteins from angiosperms. Plant Sci 87:9–18

Arahira M, Fukazawa C (1994) Ginkgo 11S seed storage protein family mRNA: unusual Asn–Asn linkage as a post-translational cleavage site. Plant Mol Biol 25:597–605

Bäumlein H, Wobus U, Pustell J, Kafatos FC (1986) The legumin gene family: structure of a B-type gene of *Vicia faba* and a possible legumin gene-specific regulatory element. Nucleic Acids Res 14:2707–2720

Bäumlein H, Nagy I, Villarroel R, Inze 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

Chen Z, Shoemaker RC (1998) Four genes affecting seed traits in soybean map to linkage group F. J Hered 89:211–215

Cho TJ, Nielsen NC (1989) Glycinin *Gy3* gene from the soybean variety Dare. Nucleic Acids Res 17:4388

Cho TJ, Davies CS, Nielsen NC (1989) Inheritance and organization of glycinin genes in soybean. Plant Cell 1:329–337

Cregan PB, Kollipara KP, Xu SJ, Singh RJ, Fogarty SE, Hymowitz T (2001) Primary trisomics and SSR markers as tools to associate chromosomes with linkage groups in soybeans. Crop Sci 41:1262–1267

Dickinson CD, Evans RP, Nielsen NC (1988) RY-repeats are conserved in the 5′-flanking regions of legume seed-protein genes. Nucleic Acids Res 16:371

Diers BW, Beilinson V, Nielsen NC, Shoemaker RC (1993) Genetic mapping of the *Gy4* and *Gy5* glycinin genes in soybean and analysis of a variant of *Gy4*. Theor Appl Genet 89:297–304

Dunwell JM, Khuri S, Gane PJ (2000) Microbial relatives of the seed storage proteins of higher plants: conservation of structure and diversification of function during evolution of the cupin superfamily. Microbiol Mol Biol Rev 64:153–179

Garner ME, Hymowitz T, Xu SJ, Hartman GL (2001) Physical map location of the Rps1-k allele in soybean. Crop Sci (in press)

Häger K-P, Fischer H (1999) Molecular phylogenies and structural diversification of gymnosperm and angiosperm storage globulins. In: Shewry P, Casey R (eds) Seed proteins. Kluwer Academic Publishers, The Netherlands, pp 499–516

Jung R, Scott MP, Nam YW, Beaman TW, Bassüner R, Saalbach I, Müntz K, Nielsen NC (1998) The role of proteolysis in the processing and assembly of 11S seed globulins. Plant Cell 10:343–357

Kaizuma H, Odanaka H, Sato H, Kowata H (1990) Mutants of soybean storage proteins induced with gamma-ray irradiation. III. Crude protein and oil contents of a mutant strain on 11S globulin subunits. Jpn J Breed 40 Suppl 1:504–505

Keim P, Shoemaker RC, Palmer RG (1989) Restriction fragment length polymorphism diversity in soybean. Theor Appl Genet 68:253–257

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK

Kitamura K (1993) Breeding trials for improving the food-processing quality of soybeans. Trends Food Sci Technol 4:64–67

Kitamura K, Davies CS, Nielsen NC (1984) Inheritance of alleles for *Cgy1* and *Gy4* storage protein genes in soybean. Theor Appl Genet 68:253–257

Lander ES, Green P, Abranhamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) Mapmaker, an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Lee J, Zou J, Xu S, Kollipara KP, Cregan PB, Singh RJ, Hymowitz T (2000) Development of the universal soybean map. 8th Biennial Conf Cellular and Mol Biol of the Soybean. August 13–16, 2000, Lexington, Kentucky, Abstract PIII 06

Lelievre JM, Oliveira LO, Nielsen NC (1992) 5′-CATGCAT-3′elements modulate the expression of glycinin genes. Plant Physiol 98:387–391

Lopes MA, Takasaki K, Bostwick DE, Helentjaris T, Larkins BA (1995) Identification of two opaque2 modifier loci in Quality Protein Maize. Mol Gen Genet 247:603–613

March JF, Pappin DJC, Casey R (1988) Isolation and characterization of a minor legumin and its constituent polypeptides from *Pisum sativum* (pea). Biochem J 250:911–915

Nielsen NC, Dickinson CD, Cho TJ, Thanh VH, Scallon BJ, Fischer RL, Sims TL, Goldberg RB (1989) Characterization of the glycinin gene family. Plant Cell 1:313–328

Nielsen NC, Bassüner R, Beaman TW (1997) The biochemistry and cell biology of embryo storage proteins. In: Larkins BA, Vasil IK (eds) Cellular and molecular biology of plant seed development. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 151–220

Shoemaker RC, Polzin KM, Labate J, Specht JE, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, Kochert G, Boerma HR (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–338

Shutov AD, Bäumlein H (1999) Origin and evolution of seed globulins. In: Shewry P, Casey R (eds) Seed proteins. Kluwer Academic Publishers, The Netherlands, pp 543–562

Sims TL, Goldberg RB (1989) The glycinin *Gy1* gene from soybean. Nucleic Acids Res 17:4386

Staswick PE, Hermodson MA, Nielsen NC (1984) The amino-acid sequence of the A2B1a subunit of glycinin. J Biol Chem 259:13,424–13,430