Theor Appl Genet (1985) 70:510–519

Identification and characterization of DNA clones encoding group-II glycinin subunits *

B. Scallon [1], V. H. Thanh [1], L. A. Floener [1] and N. C. Nielsen [2]
USDA [1]/ARS [2] and the Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

Received December 20, 1984; Accepted January 14, 1985
Communicated by D. von Wettstein

Summary. DNA clones that encode the group-II subunits of soybean glycinin were identified and compared with clones for group-I subunits. The group-I clones hybridize weakly to those from group-II at low stringency, but fail to hybridize with them at moderate or high stringency. The genes for the group-II subunits are contained in 13 and 9 kb EcoRI fragments of genomic DNA in cultivar CX635-1-1-1. These fragments contain genes for subunits $A_5A_4B_3$ and A_3B_4, respectively. The larger size of mature group-II subunits compared with group-I subunits is correlated with a larger sized mRNA. However, the gross arrangement of introns and exons within the group-II coding regions appears to be the same as for the genes which encode group-I subunits. Messenger RNA for both groups of glycinin subunits appear in the seed at the same developmental interval, and their appearance lags slightly behind that of mRNAs for the α/α' subunits of β-conglycinin. These data indicate that the glycinin gene family is more complex than previously thought.

Key words: Gene family – Exonuclease VII – Storage protein – Soybean

Introduction

A large number of plant genes have been shown to be members of multigene families. These include the genes for the small subunit of ribulose-1,5-bisphosphate carboxylase (Berry-Lowe et al. 1982), chlorophyll *a/b* binding protein (Dunsmuir et al. 1983), leghaemoglobin (Baulcombe et al. 1978) and storage proteins of various plants including maize (Pedersen et al. 1982; Burr et al. 1982), barley (Miflin et al. 1983; Rasmussen et al. 1983), and peas (Croy et al. 1982).

However, there is considerable variation in the sizes of these families. For example, Croy et al. (1982) estimated four legumin genes to be in the genome of pea, while Baulcombe et al. (1978) estimated forty genes to be in the leghaemoglobin gene family. We are investigating the relatively small family of genes in soybeans that encode the various subunits of the major seed storage protein, glycinin. By either taking advantage of natural variation in the soybean population or by introducing mutations in the DNA sequence that lead to structural changes in glycinin, the nutritional quality and functional properties of products made from soybean seeds may be improved.

Each glycinin gene encodes a precursor that consists of a signal sequence followed by an acidic polypeptide, a short peptide linker, and a basic polypeptide (Tumer et al. 1982; Marco et al. 1984). The signal sequence is removed co-translationally and the peptide linker post-translationally in a manner analogous to prohormone processing in mammals. A single disulfide bond connects the acidic polypeptide and basic polypeptide components in mature subunits (Kitamura et al. 1976; Badley et al. 1975; Staswick et al. 1981, 1984). The final glycinin complex of MW = 360,000 daltons is comprised of six such subunits (Badley et al. 1975; Catsimpoolas et al. 1967).

DNA clones of members of the glycinin multigene family were previously reported by Fischer and Goldberg (1982). Their experimental results showed that three different-sized EcoRI fragments of soybean DNA hybridized to a glycinin cDNA probe in Southern blot hybridizations. Each fragment originated from a different gene and each gene was estimated to be present once per haploid genome. However, characterization of purified glycinin subunits by NH_2-terminal sequence analysis implied that there should be a minimum of five different glycinin genes (Moreira et al. 1979; Staswick et al. 1981) and that the subunits could be separated into two major

Abbreviations: bp = base pairs; kb = kilobase pairs; SDS = sodium dodecyl sulfate

* Cooperative research between USDA/ARS and the Indiana Agric. Expt. Station. This work was supported in part by grants from the USDA Competitive Grants Program and the American Soybean Association Research Foundation. This is Journal Paper No. 10,078 from the Purdue Agricultural Experiment Station

groups based on differences in size and sequence homology (Nielsen 1984).

Since comparison of NH_2-terminal sequences among subunits from different groups showed only 50% homology, any given glycinin clone from one group might not cross-hybridize to members of the other group at moderate or high stringency. Consequently, the detection of glycinin genes as previously reported was likely incomplete. The objective of the present work was to determine if a second group of glycinin genes exists in the soybean genome.

Materials and methods

Materials

Restriction enzyme were purchased from either Bethesda Research Laboratories (BRL), New England Biolabs, or Boehringer Mannheim. Terminal transferase, 5-bromo-4-chloroindoyl-β-D-galactoside (X-gal) and DNA polymerase I were from Boehringer Mannheim. S1 nuclease, Exonuclease VII, nick-translation kits, T_4 polynucleotide kinase, isopropyl thio-β-D-galactoside (IPTG), low-melting-point (LMP) agarose and a nucleic acid chromatography system (NACS) were obtained from BRL. Oligo $(dT)_{12-18}$ and ribonuclease H (RNase H) were purchased from P. L. Biochemicals. The [α-^{32}P]-dNTP's and [γ-^{32}P]-ATP were from either Amersham or ICN Nutritional Biochemicals Inc. Avian myeloblastosis virus (AMV) reverse transcriptase was obtained from J. Beard, Life Sciences, Inc. Nitrocellulose was purchased from Schleicher and Schuell and the nylon membranes from Pall Biodyne. *E. coli* K12 strain TB1 was obtained from Dr. Tom Baldwin at Texas A & M University.

PolyA$^+$ RNA isolation

Mid-maturation stage embryos were harvested (about 21–25 days after flowering) from the soybean cultivar CX635-1-1-1 (Moreira et al. 1979) [*Glycine max* (L.) Merr.] which was grown at the Purdue University Agronomy Farm during the summer of 1982. Polyribosomes and polyA$^+$ RNA were isolated from embryos frozen in liquid nitrogen at harvest as described by Tumer et al. (1981).

For the developmental study, embryos were harvested 17, 19, 21, 23, 25, 28 and 31 days after flowering from CX635-1-1-1 grown at the farm in 1983. They were frozen in liquid nitrogen immediately after harvesting, then stored at –80 °C until used. Total RNA was extracted from these embryos by the method of Hall (1978) and stored in sterile water at –80 °C,until used.

Preparation of cDNA libraries

Oligo (dC)-tailed double-stranded cDNA was made as described by Maniatis et al. (1982) with mid-maturation polyA$^+$ RNA as template, except that the double-stranded cDNA that resulted was size-fractionated on a 5–20% sucrose gradient. The molecules larger than 500 bp were annealed with PstI-cleaved, oligo (dG)-tailed pUC8 (Vieira and Messing 1982). *E. coli* K12 strain TB1 was transformed with the annealing mixture by the method of Mandel and Higa (1970), and transformants bearing recombinant plasmids were identified as white colonies on X-gal/IPTG indicator plates.

A second cDNA library, constructed to optimize the yield of full-length clones, was prepared by a modification of the method of Gubler and Hoffman (1983). First strand synthesis was carried out essentially as described in Maniatis et al. (1982), and was monitored by incorporation of ^{3}H-dCTP. The reaction was terminated by addition of 4 µl 250 mM EDTA and 1 µl 10% SDS. An equal volumn of chloroform/isoamyl-alcohol (v/v) was added and then the organic phase was re-extracted with 10 mM Tris · HCl (pH 8), 100 mM NaCl, 1 mM EDTA. The cDNA/mRNA hybrids were separated from unincorporated deoxynucleotides with a Sephadex G-100 column that had been equilibrated with Tris · HCl (pH 7.5), 5 mM NaCl and saturated with denatured salmon sperm DNA. The hybrids were ethanol precipitated and resuspended in deionized water. Second strand synthesis (100 µl final volume) was as described by Gubler and Hoffman (1983), except that DNA ligase was eliminated and the concentrations of the nonradioactive deoxynucleotides, RNAse H and DNA polymerase I were increased to 50 µM, 12.5 units/ml and 250 units/ml, respectively. Second strand synthesis was followed by incorporation of [α-^{32}P]dGTP. The reaction was stopped by addition of 8 µl 250 mM EDTA after sequential incubations at 12 °C for 60 min and 22 °C for 60 min. The ds-DNAs were extracted and separated from unincorporated deoxynucleotides as after first strand synthesis and had ^{32}P-second strand/^{3}H-first strand ratios of about 1.0. These DNAs were then inserted into the PstI site of pUC8 (Vieira and Messing 1982) without prior size selection.

DNA sequencing

DNA sequence was determined by the method of Maxam and Gilbert (1980). Fragments were labelled either by T_4 polynucleotide kinase and [γ-^{32}P]-ATP (7,000 Ci/mmole) or Klenow fragment and [α-^{32}P]-dNTP (3,000 Ci/mmole).

Nucleic acid hybridizations

Unless otherwise stated, all hybridizations were performed at 42 °C in 50% formamide, 5×SSC (0.75 M NaCl, 0.75 M Na citrate, pH 7.0), 5×Denhardt's (0.1% bovine serum albumin, 0.1% ficoll, 0.1% polyvinylpyrrolidone), 50 mM Na phosphate, pH 6.5, 1% SDS, 1 mM EDTA, 250 µg/ml denatured calf thymus DNA, and 1 µg/ml poly(dG) · poly(dC).

For Southern blots, 9 µg of DNA isolated from leaves of CX635-1-1-1 by a modification of the method of Heyn et al. (1974) were digested with excess EcoRI, electrophoresed in a 0.5% agarose gel, and transferred to Pall Biodyne nylon membrane by the method of Southern (1975). Modifications of the protocol provided by the manufacturer of the membrane included treatment of the gel with 0.25 N HCl for 20 min before alkali denaturation, as well as the following changes in the hybridization buffer: 1) the presence of 1 mM EDTA, 2) 1% SDS instead of 0.1%, 3) 1 µg/ml poly(dG) · poly(dC), 4) 4–6 h prehybridization instead of 1 h, and 5) 10–12 h of washing at 55 °C in 0.1×SSC, 0.1% SDS with several changes of wash buffer instead of a one-half-hour wash at 50 °C.

Northern blots were done following electrophoresis of 4 µg of polyA$^+$ RNA from CX635-1-1-1 embryos on 1% agarose gels that contained 10 mM methylmercury hydroxide (Bailey and Davidson 1976). RNA was transferred to Pall Biodyne nylon membrane without staining with ethidium bromide. Hybridizations were done as for Southern blots and the filters were washed for 2–4 h at 70 °C in 0.1×SSC and 0.1% SDS.

For RNA dot blots, total RNA (2 µg) from each developmental stage was denatured with 5 mM methylmercury hy-

droxide, heated to 37 °C for 2 min, and blotted onto nitrocellulose saturated in 10×SSC.

A genomic library of DNA from the variety 'Dare', constructed in the phage vector Charon 4A by R. B. Goldberg, was screened by plaque hybridization on nitrocellulose filters as described by Benton and Davis (1977). Colony hybridizations were done according to Grunstein and Hogness (1975).

Nucleic acid labelling

Plasmid inserts were isolated from low melting point agarose gels according to the method of Weislander (1979), and further purified by passage of the DNA through a NACS column according to the protocol supplied by the manufacturer. A nick-translation kit (BRL) was used to label either 10 ng (for genomic Southerns and Northerns) or about 100 ng of insert DNA with 60–100 μCi of [α-^{32}P]-dNTP (sp. act. = 3,000 Ci/mmole).

Exonuclease VII protection

Three micrograms of DNA from genomic clone λDG258 was either left intact or digested with HindIII and extracted with phenol/chloroform. The procedure of Maniatis (1982) was followed to denature and hybridize the DNA to 3 μg of total polyA$^+$ RNA. To the 15 μl of hybridization mix were added 300 μl of chilled Exonuclease VII buffer (67 mM K_2HPO_4 pH 7.9, 8 mM Na_2EDTA, 10 mM 2-mercaptoethanol) and one unit of Exonuclease VII. The digestion mixture was incubated at 37 °C for 30 min, chilled to 0 °C, extracted with phenol/chloroform, and then precipitated with ethanol. The protected fragments were separated by electrophoresis in a 1.5% alkaline agarose gel, blotted onto nitrocellulose, and then hybridized with either the nick-translated 7.2 kb EcoRI fragment, the 4.1 kb EcoRI-HindIII fragment (5′ half of the 7.2 kb EcoRI fragment), or the 3.1 kb HindIII-EcoRI fragment (3′ half of the 7.2 kb EcoRI fragment).

S1 protection

The 7.2 kb EcoRI fragment from genomic clone λDG258 was subcloned into pUC8 and used to analyze gene structure. Two micrograms of the subclone DNA was linearized with BamHI, extracted with phenol/chloroform and chloroform, and then co-precipitated with five micrograms of total polyA$^+$ RNA and 100 μg of *E. coli* tRNA. The method of Maniatis et al. (1982) was followed to denature, hybridize, and treat with S1 nuclease (500 units/ml). The protected fragments were separated by electrophoresis in a 1.5% alkaline agarose gel, blotted onto nitrocellulose, and hybridized to the nick-translated 7.2 kb EcoRI fragment.

Results and discussion

Previous work established that there are five major types of glycinin subunits in seeds of soybean cultivar CX635-1-1-1 and they are denoted $A_{1a}B_2$, $A_{1b}B_{1b}$, A_2B_{1a}, A_3B_4 and $A_5A_4B_3$. Differences in their primary structure that permit them to be identified unambiguously are summarized elsewhere (Nielsen 1984). The five subunit types can be divided into two groups based on size and degree of sequence homology. Group-I subunits, which are all about 58,000 daltons, include $A_{1a}B_2$, $A_{1b}B_{1b}$ and A_2B_{1a}. The group-II subunits include A_3B_4 and $A_5A_4B_3$, and are more variable in size (62,000 and 69,000 daltons, respectively). Amino acid sequence homology among members of the same group exceeds 85%, whereas that between members of different groups is only about 50%.

Several partial cDNA clones that encode glycinin have been described (Goldberg et al. 1981; Marco et al. 1984). These clones hybridize to restriction fragments of about 5.6, 4.0 and 3.2 kb on Southern blots of EcoRI-digested genomic DNA (Fischer and Goldberg 1982). The EcoRI fragments originate from the 3′-end of three genes that encode group-I subunits (Nielsen and Goldberg, unpublished data). Thus, clones which encoded group-II subunits have until now not been described.

It was considered likely that group-I probes would not hybridize to group-II DNA under conditions of moderate or high stringency, and that this was responsible for the failure to detect group-II coding sequences in hybridization experiments that used group-I probes. To test this possibility, a cDNA library prepared from the polyA$^+$ RNA of mid-maturation stage embryos was screened by colony hybridization at low stringency (32 °C, 5×SSC, 50% formamide). The insert from clone pA06, which previously was shown to encode part of group-I subunit A_2B_{1a} (Marco et al. 1984) was used as probe. Two types of colonies were detected. One type hybridized strongly and was presumed to consist of plasmids with group-I inserts. Colonies of the other type hybridized weakly to the group-I probe and were characterized further. The colonies that hybridized weakly were amplified and their plasmid DNA was purified. When colony hybridization was repeated with DNA from one of the new clones as probe, most of the clones that hybridized faintly to pA06 at low stringency hybridized strongly to the new probe at moderate stringency (42 °C, 5×SSC, 50% formamide). This result suggested that the colonies which hybridized weakly to pA06 contained inserts related to, but distinct from, the group-I inserts.

The degree of cross-hybridization between a group-I clone and two representative members of the new group of clones, pG258 and pG23, was studied in more detail to define the relationships between them (Fig. 1). In these experiments, pG154 was used as the group-I probe because it was shown to contain a larger proportion of the A_2B_{1a} coding sequence than pA06 (e.g. COOH-terminal half of the acidic component, all of the basic component, and the poly A-tail). The group-I clone, pG154, hybridized to pG258 and pG23 at low stringency conditions (T_m-51 °C) but only very weakly at moderate stringencies (T_m-35 °C). On the other hand, pG258 and pG23 hybridized strongly to each other at moderate stringencies and only at high stringencies (T_m-17 °C) did they hybridize specifically to themselves. These data verify that pG258 and pG23 are closely related to each other and that both are related, although to a lesser extent, to pG154. Moreover, these data indicated that pG258 and pG23 had different inserts which could be resolved by their inability to cross-hybridize with one another at high stringencies.

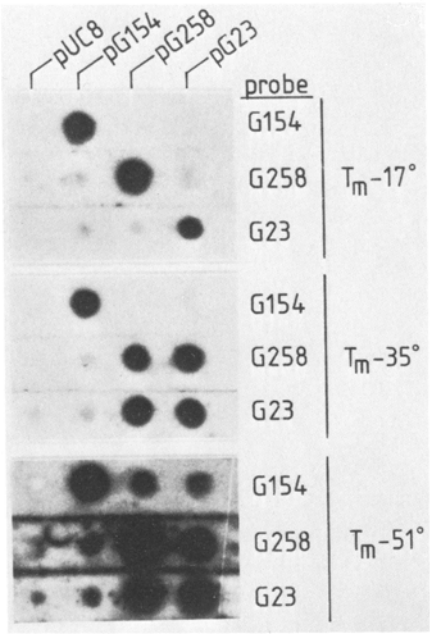

Fig. 1. Colony hybridizations to evaluate the relationship of pG258 and pG23 to a group-I glycinin clone, pG154. Nitrocellulose filters were pre-hybridized and hybridized in the buffer described in "Materials and methods" except 40% formamide was used instead of 50% formamide and 1 μg/ml of linearized, denatured pUC8 was included. Stringency conditions were the same for hybridization and washing and are indicated *at the right.* The T_m values were calculated as described by Casey and Davidson (1977) assuming a G+C value of 50%. All three probes were nick-translated inserts purified from pUC8

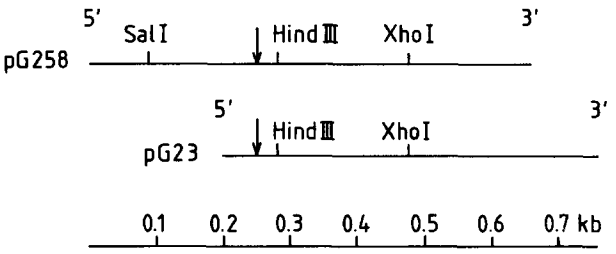

Fig. 2. Restriction maps of pG258 and pG23. The former encodes Gy_4 ($A_5A_4B_3$)and the latter Gy_5 (A_3B_4). *Arrows* mark the beginning of the regions that encode the NH_2-termini of the basic polypeptides, and correspond to a similarly marked point in the nucleotide sequence given in Fig. 3

Identification of two cDNA clones

To identify the new glycinin clones, restriction maps of pG258 (700 bp insert) and pG23 (630 bp insert) were made by conventional methods (Fig. 2) and used to obtain complete DNA sequences of the glycinin inserts (Fig. 3). The nucleotide sequence immediately upstream from the unique HindIII site in both clones permitted their unambiguous identification. These regions had reading frames that translated perfectly into the NH_2-terminal amino acid sequence of a basic polypeptide component (Moreira et al. 1979). Since earlier work by Staswick et al. (1981) established the specific pairings between acidic and basic polypeptide components for soybean line CX635-1-1-1, it was determined that pG258 encoded the $A_5A_4B_3$ group-II subunit, while pG23 encoded the other member of this group, A_3B_4. After identifying pG258 as a group-II clone by DNA sequencing, we observed that the sequence was matched almost exactly by a clone, reported by Schuler et al. (1982), whose identity remained unknown to them.

Comparison of the DNA sequence in pG258 and pG23 with sequence in a group-I gene (Marco et al. 1984) revealed that there was about 60% nucleotide homology between the two groups. The two group-II clones, on the other hand, showed 85–90% homology with each other.

Southern hybridizations

Because colony hybridizations at moderate stringency showed very little cross-hybridization between members of the two groups of clones, it was important to identify the genomic DNA fragments that hybridized to them in Southern blots. When EcoRI-digested leaf DNA from cultivar CX635-1-1-1 was probed with nick-translated inserts from either pG258 or pG23, two fragments hybridized which were about 13 and 9 kb (Fig. 4). Both were larger than the three EcoRI fragments shown previously to hybridize to group-I probes (Fischer and Goldberg 1982). A mixed probe with both group-I and group-II DNA, however, hybridized with all five restriction fragments in the DNA preparations. These results indicated that it was unlikely that the 13 and 9 kb EcoRI fragments that hybridized to the group-II probes were due to incomplete digestion of the DNA. Like the group-I EcoRI fragments, the two group-II fragments appeared to be present at a frequency of once per haploid genome as shown by comparison to genomic reconstructions (Fig. 4). Thus, none of the glycinin subunit genes appear to be present in substantially higher copy number than the others.

The two group-II EcoRI fragments hybridized differentially to the inserts from pG23 and pG258 (Fig. 4). The 13 kb fragment interacted more strongly with pG258 than pG23, whereas the 9 kb fragment hybridized more strongly to pG23. These data suggested that the 13 kb fragment contained a gene that encoded the $A_5A_4B_3$ glycinin subunit, while the 9 kb fragment contained a gene for the A_3B_4 subunit. The glycinin genes on the 13 and 9 kb EcoRI fragments have been referred to as Gy_4 and Gy_5, respectively, in keeping with the nomenclature initiated by Kitamura et al. (1984).

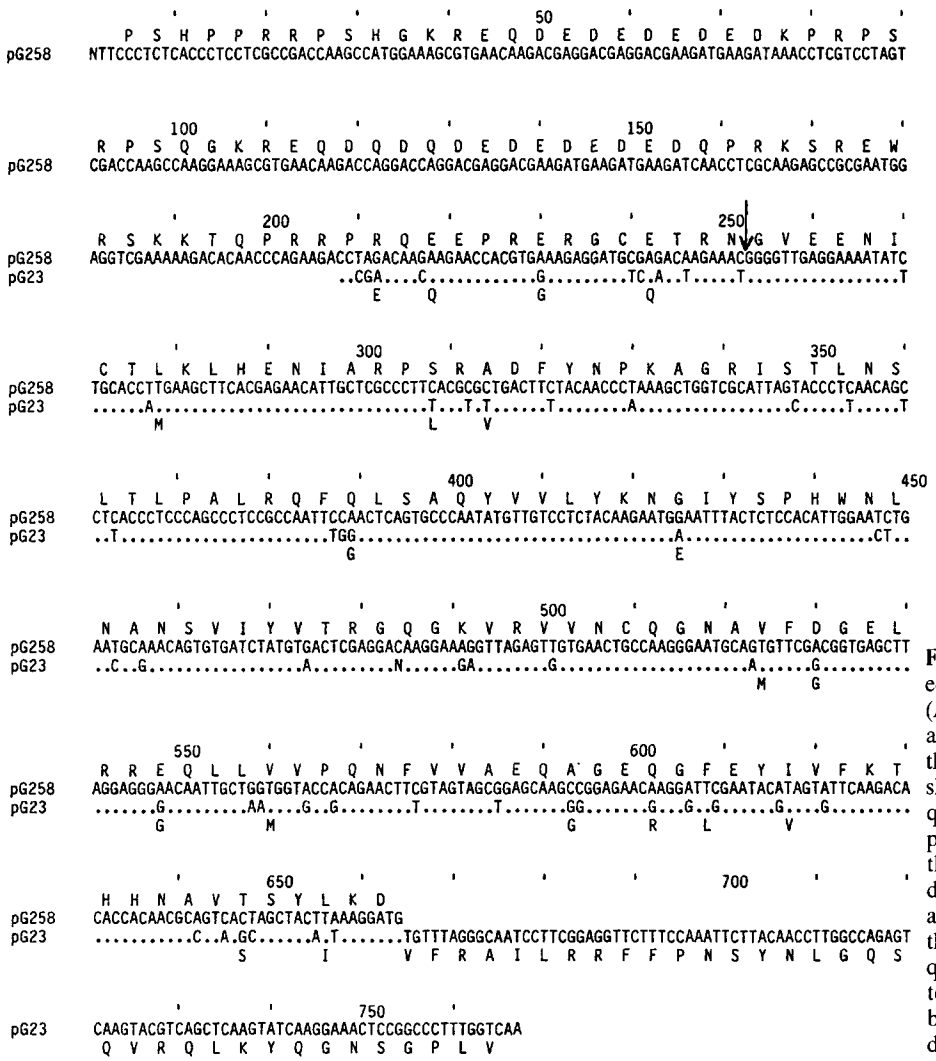

Fig. 3. DNA sequences and deduced amino acid sequences of pG258 ($A_5A_4B_3$) and pG23 (A_3B_4). The amino acid sequence for $A_5A_4B_3$ that was deduced from pG258 is shown above its nucleotide sequence. Dots in the sequence of pG23 indicate that the nucleotide is the same as in pG258. Amino acids deduced from the pG23 sequence are shown only if they differ from those deduced from the pG258 sequence. The *arrow* marks the NH_2-terminal of the basic polypeptide for both subunits. N indicates an undetermined nucleotide

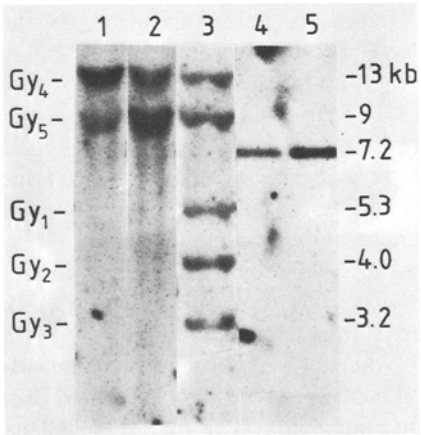

Fig. 4. Southern hybridization of EcoRI-digested leaf DNA. *Lane 1* pG258 insert probe; *Lane 2* pG23 insert probe; *Lane 3* mixed probe with inserts from pG258, pG23, and pA06; *Lanes 4 and 5* one and two copy-equivalents of Gy_4 for comparison to *lane 3* only (e.g. 50 and 100 pg of pDG258R7.2 digested with EcoRI)

Isolation and characterization of a genomic clone

The results of the Southern hybridizations clearly showed that more glycinin genes exist in the soybean genome than were previously detected. Experiments to isolate and characterize these genes began by screening an *Alu*I-*Hae*III genomic library by plaque hybridization using as probe a mixture of the inserts from cDNA clones pG258 and pG23. Of the 300,000 plaques that were screened, several were identified as clones containing Gy_4. This conclusion was drawn when it was shown that these genomic clones hybridized more strongly with pG258 than pG23. One clone, λDG258 (14 kb insert), was amplified for further study.

To locate the glycinin sequence within the 14 kb insert, Southern transfers were done after digestion of λDG258 DNA with EcoRI alone, or after digestion with both EcoRI and HindIII. The fragments were probed with nick-translated pG258 insert. The results

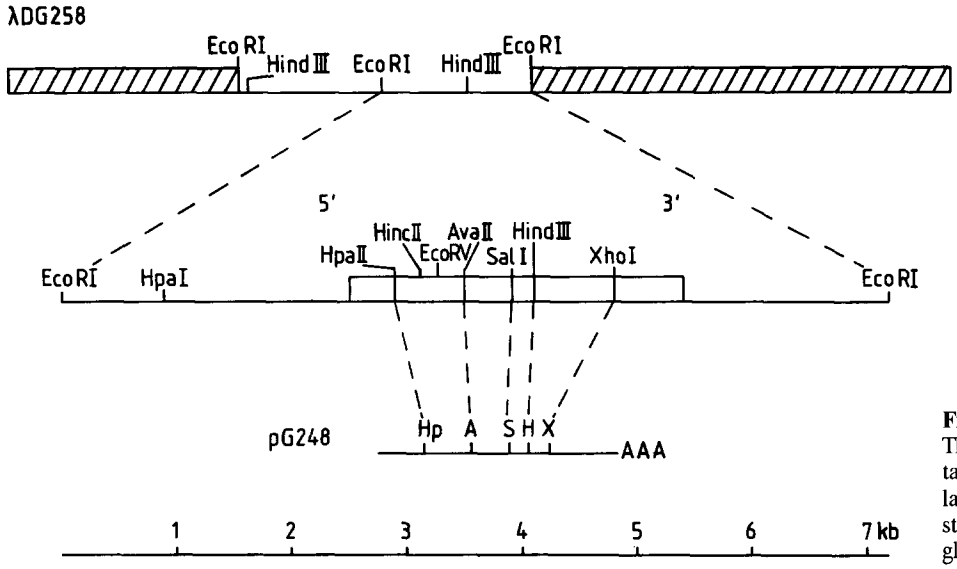

Fig. 5. Restriction map of λDG258. The 7.2 kb EcoRI fragment containing the glycinin gene is enlarged for comparison to the restriction map of a full-length glycinin cDNA clone, pG248

indicated that the entire glycinin sequence in λDG258 resided on a 7.2 kb EcoRI fragment which was cleaved by HindIII into a 4.1 kb fragment and a 3.1 kb fragment (Fig. 5). This HindIII site corresponded to the one in cDNA clone pG258 that was shown by nucleotide sequence analysis to be in the basic polypeptide 30 bp from the linker sequence. There were five other EcoRI fragments in the 14 kb insert of λDG258 that ranged from 0.4 to 2.0 kb. However, they were not studied further because they did not contain coding sequence for Gy_4. The position of the 7.2 kb EcoRI fragment, the orientation of the gene, and restriction sites of interest are shown in Fig. 5.

Several mRNA protection experiments were performed to determine if the size and structure of Gy_4 was different from that determined for the three group-I genes (Nielsen 1984; Nielsen et al., unpublished). Exonuclease VII digestion (Chase and Richardson 1974) was used for one set of experiments. Since this enzyme cleaves single-stranded DNA from free ends, it can be used to define the transcribed region of the gene including introns. DNA of λDG258 was hybridized to total polyA$^+$ RNA, the mixture was digested with Exonuclease VII, and then protected fragments were separated in alkaline agarose gels as described by McDonell et al. (1977). After being blotted onto nitrocellulose, the protected fragments were identified using one of three probes. A 2.9 kb protected fragment was observed when undigested λDG258 DNA was hybridized to polyA$^+$ RNA and identified using the nick-translated 7.2 kb EcoRI fragment (Fig. 6A). This result indicated the overall length of the transcribed region in Gy_4. To specifically map the 5′ and 3′ ends of the transcribed region, DNA of λDG258 was digested with HindIII prior to hybridization with RNA and the Exonuclease VII-protected sample divided into two equal portions. One portion was hybridized to the nick-translated 4.1 kb EcoRI-HindIII fragment (5′ half of the original 7.2 kb EcoRI fragment). A 1.6 kb fragment hybridized strongly and a 1.3 kb fragment hybridized weakly (Fig. 6A). The other half of the sample was hybridized to the 3.1 kb HindIII-EcoRI fragment (3′ half of the original 7.2 kb EcoRI fragment). This hybridized to a 1.3 kb fragment. The 1.3 kb band observed when the 4.1 kb EcoRI-HindIII probe was used was undoubtedly caused by contamination of the 4.1 kb probe by the 3.1 kb EcoRI-HindIII fragment upon purification from the original 7.2 kb EcoRI piece. Together the data show that the transcribed region of Gy_4 is 2.9 kb in length and that it extends 1.6 kb upstream and 1.3 kb downstream from the HindIII site.

Since the Exonuclease VII data described above indicated that the site for initiation of transcription is about 1.6 kb upstream from the central HindIII site, it was anticipated that this region would be approximately 300 bp upstream from the HpaII site shown in Fig. 5. When the nucleotide sequence upstream from the HpaII site was determined by the method of Maxam and Gilbert (1980), a region was found between −69 and −129 bp of the HpaII site that predicted the exact NH_2-terminal sequence determined for A_5 (Moreira et al. 1979). Moreover, sequences downstream from the HpaII site predicted the primary structure at the NH_2-terminal of A_4 that was reported earlier by Moreira et al. These data together with the amino acid sequence deduced from the studies with pG258 (Fig. 3)

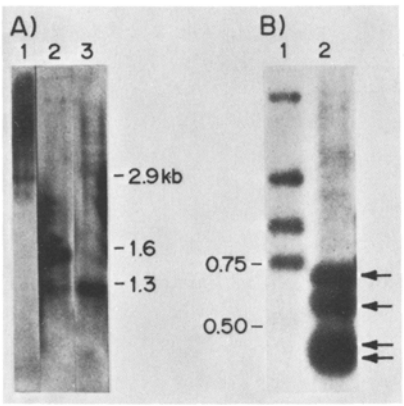

Fig. 6 A, B. Characterization of the glycinin gene by mRNA protection experiments. **A** Exonuclease VII: *Lane 1* Undigested DNA of λDG258 was hybridized to total mRNA and the RNA/DNA hybrids were treated with Exonuclease VII as described in "Materials and methods". The protected fragments were electrophoresed in alkaline agarose gels, blotted to nitrocellulose and hybridized to the nick-translated 7.2 kb EcoRI fragment. *Lanes 2 and 3* λDG258 DNA was digested with HindIII before hybridizing to mRNA then treated the same as the sample in *lane 1* except the protected fragments were hybridized to either the nick-translated 4.1 kb EcoRI-HindIII fragment from the 7.2 kb fragment (*lane 2*) or the 3.1 kb hindIII-EcoRI fragment from the 7.2 kb fragment (*lane 3*). **B** Linearized DNA of the 7.2 EcoRI subclone, pDG258R7.2, was hybridized to total mRNA then treated with S1 nuclease as described in "Materials and methods". The protected fragments (*lane 2*), after electrophoresis in an alkaline agarose gel and transfer to nitrocellulose, were hybridized to the nick-translated 7.2 kb EcoRI probe

confirmed that the polypeptide composition of Gy_4 was $A_5A_4B_3$ as predicted from earlier biochemical (Staswick et al. 1983, 1984) and genetic data (Kitamura et al. 1984).

The coding region of the Gy_2 gene for the A_2B_{1a} group-I subunit is interrupted three times by introns (Nielsen 1984). It was of interest to determine if the Gy_4 gene of group-II had the same general structure. To investigate this, the 7.2 kb EcoRI fragment from λDG258 was subcloned into pUC8 and the linearized subclone (pDG258R7.2) was used to perform an S1 protection experiment as described in "Materials and methods". Four fragments that represent exons of approximately 700, 600, 350 and 300 bp were protected from S1 digestion by seed mRNA (Fig. 6 B). This result suggests that the Gy_4 gene, like Gy_2, consists of four exons and three introns.

Recently we isolated a full-length cDNA clone for Gy_4, pG248 (2,000 bp insert), from a cDNA library that we constructed by the method of Gubler and Hoffman (1983). Comparison of its restriction map (Fig. 5) with that of λDG258 revealed additional detail about introns in Gy_4. Only five of the seven restriction sites shown in the 7.2 kb EcoRI fragment are present in pG248, presumably because the sites for HincII and EcoRV lie within an intron. In addition, the distances from AvaII to SalI and from HindIII to XhoI are less in the cDNA clone than they are in the genomic clone. The simplest explanation for the differences is that an intron between these restriction sites has been removed. Thus, the map data also suggest that the Gy_4 coding region is interrupted three times by introns.

The sizes of the introns and exons in Gy_4 can be estimated by comparison of the restriction maps for pG248 and λDG258. The distance from HpaII to AvaII is about 600 bp in the genomic clone but only about 300 bp in the cDNA clone. This suggests that the first intron is approximately 300 bp in length. Similar reasoning gives estimates of 100 bp for the intron between AvaII and SalI, and 600 bp for the one between HindIII and XhoI. The corresponding introns in Gy_2 are 238, 292 and 624 bp, respectively (Nielsen 1984). Moreover, the four exons of Gy_2 are 322, 254, 537 and 630 bp, respectively. These exon sizes also correspond reasonably well with the sizes of the four exons in Gy_4 that were estimated on the basis of S1 protection experiments (e.g., 300, 350, 600, 700 bp). Thus, at a gross level the group-I and group-II genes appear to have the same intron-exon structure.

Size of the group-II gene products

The group-I subunits have smaller apparent molecular weights than those in group-II (Nielsen 1984). Hybridization of group-I and group-II specific probes to Northern blots of soybean mRNA provided additional detail about these size differences (Fig. 7). Consistent with earlier reports (Tumer et al. 1981; Goldberg et al. 1981) the group-I probe, pG154, hybridized to mRNA about 1,950 bases in length. However, pG258, the group-II probe which contains coding sequence for the largest glycinin subunit, $A_5A_4B_3$, hybridized to slower migrating mRNA. When the two probes were mixed, two bands were evident with the slower migrating mRNA estimated to be about 200 bp larger than the faster migrating mRNA. The group-II probe that contains sequence for A_3B_4, pG23, hybridized to an mRNA intermediate in size to those that hybridized to pG154 and pG258. Thus, the increased sizes of the mature group-II subunits are correlated with the increased size of the mRNAs which encode them.

RNA dot blot hybridization

The acidic polypeptide component of A_3B_4 has a slightly higher apparent molecular weight than those from the other glycinin subunits and can easily be resolved from the others in SDS-polyacrylamide gels (Moreira et al. 1979). Meinke et al. (1981) made use of

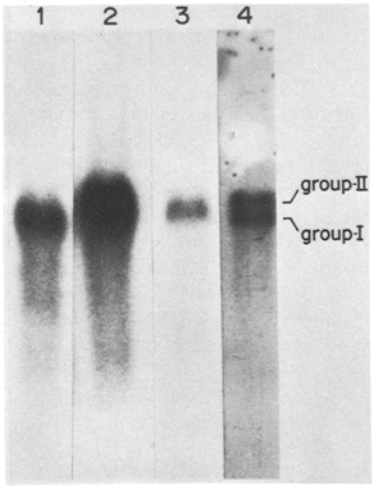

Fig. 7. Northern blot analysis of polyA$^+$ RNA isolated from mid-maturation stage embryos. The probes were the nick-translated inserts of: *lane 1* pG154 (group-I), *lane 2* pG258 (group-II), *lane 3* pG23 (group-II), and *lane 4* a mixture of pG154 and pG258. The mRNAs hybridizing to pG154 and pG258 are estimated to be about 1,950 and 2,150 bases in length, respectively

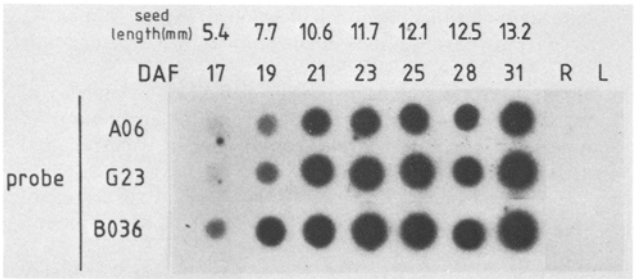

Fig. 8. Dot blot hybridization to total RNA isolated from embryos at seven stages of development. The stages are described by days after flowering (DAF) and average seed length in millimeters (mm). R and L indicate total RNA from roots and leaves, respectively. The probes were the nick-translated inserts of pA06, pG23, and pB036. The latter is a clone for the α/α' subunit of β-conglycinin (Nielsen, unpublished data). Hybridization was at 52 °C, 50% formamide, 5× SSC. Washing was done in 0.075× SSC, 0.1% SDS at 52 °C (same stringency as hybridization)

this difference in a developmental study to show that A_3 appeared one to two weeks later than the rest of the seed 11S polypeptides in the cultivar they studied. Their report suggested that the A_3B_4 subunit, or for that matter both group-II genes, were expressed at a later developmental stage than the group-I genes. To test this possibility, we isolated total RNA from CX635-1-1-1 at seven early stages of seed development. The RNA from each stage was denatured with methylmercury hydroxide, blotted onto nitrocellulose, and then hybridized to probes for either a group-I glycinin subunit, a group-II glycinin subunit, or the α/α' subunit of β-conglycinin (Fig. 8). In each case the same amount of RNA was blotted onto the nitrocellulose and the specific activities of the nick-translated probes were about equal. No apparent difference was evident between the time of appearance of message for the group-I and group-II subunits. Identical results were obtained when the polyA$^+$ RNA fraction of the total RNA was used, or when the hybridizations were repeated at stringencies high enough to ensure minimal cross-hybridization between A_3B_4 and $A_5A_4B_3$ mRNAs (e.g., 62 °C, 50% formamide, 5×SSC). While no difference in the time of appearance among the group-I and group-II mRNAs was observed, evidence was obtained which confirmed the report by Meinke et al. (1981) that message for the α/α' subunits of β-conglycinin appeared slightly earlier than those for glycinin. Moreover, as also reported by Fischer and Goldberg (1982), the seed protein mRNAs were tissue-specific (Fig. 8).

Goldberg et al. (1981) estimated that glycinin mRNAs comprise 10% of the mRNA mass in mid-maturation stage soybean embryos. This value was obtained by DNA-excess DNA/RNA hybridization experiments which utilized a group-I glycinin probe. Since cross-hybridization of the probe to group-II mRNAs would not have occurred under the conditions they employed, their data undoubtedly underestimated the total glycinin mRNA content. Our experiments will not permit accurate quantitation of group-II mRNAs. However, the results of the RNA dot blot experiments suggest that the group-II mRNAs are as abundant as group-I mRNAs during the early and mid-maturation stages of embryogenesis. This suggests that glycinin mRNAs account for a minimum of 20% of the total seed mRNA.

The data provide clear evidence that there are a minimum of two groups of genes in soybeans that contribute messages used during seed development for the synthesis of glycinin. Since one long-range objective is to correct the nutritional deficiency that total soybean seed protein has for the sulfur amino acids, methionine and cysteine, the existence of the group-II glycinin genes has practical importance. The group-II subunits contain fewer methionine residues than the group-I subunits. Therefore, one way that has been proposed for improving seed nutritional quality (Nielsen 1984) is to eliminate the group-II subunits from glycinin by somehow preventing the expression of the group-II genes.

This report also shows that the group-II glycinin genes contain the same number of exons and introns as

the group-I genes and that these appear to be located at analogous positions. This conservation of structure implies that all of the glycinin genes has a common origin. Moreover, the gross structure of the soybean glycinin genes closely resembles that for pea legumin as reported by Boulter's group (Lycett et al. 1984), and this implies that all legume 11S genes evolved from a common ancestor. The genes for analogous 11S globulins in *Brassica,* the cucurbits and even cereals may be found to have related structures and their gene products may undergo similar processes of post-transcriptional and post-translational modification.

Acknowledgements. We are grateful to Robert Goldberg for his *Alu*I-*Hae*III soybean genomic library and to Brian Larkins for supplying us with some oligo (dG)-tailed pUC8 vector. We thank Drs. Jack Dixon, Larry Dunkle, and Brian Larkins for reviewing the manuscript and Crescentia Motzi for typing it.

References

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

Bailey JM, Davidson N (1976) Methylmercury as a reversible denaturing agent for agarose gel electrophoresis. Anal Biochem 70:75–85

Baulcombe D, Verma DPS (1978) Preparation of a complementary DNA for leghaemoglobin and direct demonstration that leghaemoglobin is encoded by the soybean genome. Nucleic Acids Res 5:4141–4153

Benton WD, Davis RW (1977) Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196:180–182

Berry-Lowe SL, McKnight TD, Shah DM, Meagher RB (1982) The nucleotide sequence, expression, and evolution of one member of a multigene family encoding the small subunit of ribulose-1,5-bisphosphate carboxylase in soybean. J Mol Appl Genet 1:483–498

Burr B, Burr FA, St John TP, Thomas M, Davis RW (1982) Zein storage protein gene family of maize. J Mol Biol 154: 33–49

Casey J, Davidson N (1977) Rates of formation and thermal stabilities of RNA:DNA and DNA:DNA duplexes at high concentrations of formamide. Nucleic Acids Res 4: 1539–1552

Catsimpoolas N, Rogers DA, Circle SJ, Meyer EW (1967) Purification and structural studies of the 11S component of soybean proteins. Cereal Chem 44:631–637

Chase JW, Richardson CC (1974) Exonuclease VII of *Escherichia coli.* J Biol Chem 249:4545–4552

Croy RRD, Lycett GW, Gatehouse JA, Yarwood JN, Boulter D (1982) Cloning and analysis of cDNAs encoding plant storage protein precursors. Nature 295:76–79

Densmuir P, Smith SM, Bedbrook J (1983) The major chlorophyll *a/b* binding protein of petunia is composed of several polypeptides encoded by a number of distinct nuclear genes. J Mol Appl Genet 2:285–300

Fischer RL, Goldberg RB (1982) Structure and flanking regions of soybean seed protein genes. Cell 29:651–660

Goldberg RB, Hoschek G, Ditta GS, Breidenbach RW (1981) Developmental regulation of cloned superabundant embryo mRNAs in soybean. Dev Biol 83:218–231

Grunstein M, Hogness D (1975) Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc Natl Acad Sci USA 72:3961–3965

Gubler U, Hoffman B (1983) A simple and very efficient method for generating cDNA libraries. Gene 25:263–269

Hall TC, Ma Y, Buchbinder BU, Pyne JW, Dun SM, Bliss FA (1978) Messenger RNA for G1 protein of French bean seeds: cell-free translation and product characterization. Proc Natl Acad Sci USA 75:3196–3200

Heyn RF, Hermans AK, Schilperoort RA (1974) Rapid and efficient isolation of highly polymerized plant DNA. Plant Sci Lett 2:73–78

Kitamura K, Davies CS, Nielsen NC (1984) Inheritance of alleles for Gy_1 and Gy_4 storage protein genes in soybean. Theor Appl Genet 68:253–257

Kitamura K, Takagi T, Shibasaki K (1976) Subunit structure of soybean 11S globulin. Agric Biol Chem 40:1837–1844

Lycett GW, Croy RRD, Shirsat AH, Boulter D (1984) The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.) Nucleic Acids Res 12:4493–4506

Mandel M, Higa A (1970) Calcium dependent bacteriophage DNA infection. J Mol Biol 53:159–162

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Marco YA, Thanh VH, Tumer NE, Scallon BJ, Nielsen NC (1984) Cloning and structural analysis of DNA encoding an A_2B_{1a} subunit of glycinin. J Biol Chem 259:13436–13441

Maxam AM, Gilbert W (1980) Sequencing end-labelled DNA with base-specific chemical cleavages. In: Grossman L, Moldare K (eds) Methods in enzymology, vol 65. Academic Press, New York, pp 499–560

McDonell MW, Simon MN, Studer FW (1977) Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline gels. J Mol Biol 110:119–146

Meinke DW, Chen J, Beachy RN (1981) Expression of storage-protein genes during soybean seed development. Planta 153:130–139

Miflin BJ, Rahman S, Kreis M, Forde BG, Blanco L, Shewry PR (1983) The hordeins of barley: developmentally and nutritionally regulated multigene families of storage proteins. In: Ciferri O, Dure LS (eds) Structure and function of plant genomes. Plenum Press, New York, pp 21–90

Moreira MA, Hermodson MA, Larkins BA, Nielsen NC (1979) Partial characterization of the acidic and basic polypeptides of glycinin. J Biol Chem 254:9921–9926

Nielsen NC (1984) The chemistry of legume storage proteins. Philos Trans R Soc London, Ser B 304:287–296

Pedersen K, Devereux J, Wilson DR, Sheldon E, Larkins BA (1982) Cloning and sequence analysis reveal structural variation among related zein genes in maize. Cell 29: 1015–1026

Rasmussen SK, Hopp HE, Brandt A (1983) Nucleotide sequences of cDNA clones for B1 hordein polypeptides. Carlsberg Res Commun 48:187–199

Schuler MA, Ladin BF, Pollaco JC, Freyer G, Beachy RN (1982) Structural sequences are conserved in the genes coding for the α,α' and β-subunits of the soybean 7S seed storage protein. Nucleic Acids Res 10:8245–8261

Southern E (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Staswick PE, Hermodson MA, Nielsen NC (1981) Identification of the acidic and basic subunit complexes of glycinin. J Biol Chem 256:8752–8755

Staswick PE, Nielsen NC (1983) Characterization of a soybean cultivar lacking certain glycinin subunits. Arch Biochem Biophys 223:1–8

Staswick PE, Hermodson MA, Nielsen NC (1984) Identification of the cystines which link the acidic and basic components of the glycinin subunits. J Biol Chem 259: 13431–13435

Tumer N, Thanh VH, Nielsen NC (1981) Purification and characterization of mRNA from soybean seeds. J Biol Chem 256:8756–8760

Tumer N, Richter JD, Nielsen NC (1982) Structural characterization of the glycinin precursors. J Biol Chem 257: 4016–4018

Vieira J, Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19:259–268

Weislander L (1979) A simple method to recover intact high molecular weight RNA and DNA after electrophoretic separation in low gelling temperature agarose gels. Anal Biochem 98:305–309