Theor Appl Genet (1995) 90: 991--999

T. Zhu · L. Shi · J. J. Doyle · P. Keim

A single nuclear locus phylogeny of soybean based on DNA sequence

Received: 20 May 1994 / Accepted: 30 September 1994

Abstract Soybean [*Glycine max* (L.) Merr.] evolution was examined by sequencing portions of the restriction fragment length polymorphism (RFLP) locus A-199a of 21 taxa from the Glycininae and 1 from the Phaseoleae. Four hundred nucleotides were determined in each, aligned, and then compared for these taxa. Within the annual soybean subgenus (*Soja*), the four accessions differed at as many as 2.2% of the nucleotides. Among 13 perennial soybean species (subgenus *Glycine*), nucleotide variation ranged from 1.7% to 8.4%. The nucleotide difference between the two soybean subgenera was 3.0–7.0%. Nucleotide variation between the genus *Glycine* and the related genera of *Neonotonia*, *Amphicarpa*, *Teramnus*, and *Phaseolus* ranged from 8.2% to 16.4%. In addition to nucleotide substitutions, insertions/deletions (indels) differences were also observed and were consistent with nucleotide-based analysis. Cladistic analysis of the A-199a sequences was performed using Wagner parsimony to construct a soybean phylogeny. Sixteen equally parsimonious trees were produced from these data. The trees were 246 steps in length with a consistency index of 0.78. Indels distribution upon the consensus topology revealed a pattern congruent with the nucleotide-based phylogeny. The current taxonomic status of the soybean subgenera and the related genera of *Neonotonia*, *Amphicarpa*, and *Teramnus* were well-supported and appear monophyletic in this analysis. Homoplasy within the subgenus *Glycine* led to a lack of resolved topology for many of these 13 taxa. However, the *Glycine* clade topology was consistent with phylogenies proposed using crossing experiments and cpDNA RFLPs. These genera were arranged from ancestral to derived as: *Teramnus*, *Amphicarpa*, *Neonotonia*, and *Glycine* when *Phaseolus vulgaris* was used as an outgroup.

Key words Glycininae · Soybean · Phylogeny · DNA sequence · Genomic donors · *Glycine* · Phaseoleae

Communicated by A. L. Kohler

T. Zhu · L. Shi · P. Keim (✉)
Department of Biological Sciences, Box 5640,
Northern Arizona University, Flagstaff, AZ 86011, USA

J. J. Doyle
L.H. Bailey Hortorium, 462 Mann Library Building,
Cornell University, Ithaca, NY 14853, USA

Introduction

Polyploid species may have complex origins deriving from a single- or multiple-genomic donors. Identification of genomic donors provides information about the evolutionary origins and may also help in designing systematic programs for the introduction of foreign germ plasm into cultigens (Kenworthy 1989). Notable successes have been achieved in polyploids such as cotton (Wendel et al. 1989) and *Brassica* (Song et al. 1990) where probable genome donors have been identified. Soybean [*Glycine max* (L.) Merr.] represents a difficult taxonomic polyploid because it has significantly evolved towards diploidy (Hymowitz and Singh 1987; Zhu et al. 1994). The polyploid event may be sufficiently distant that the original genome donors are now extinct (Kumar and Hymowitz 1989). Distant evolutionary events are difficult to characterize, but molecular genetic data can be very powerful phylogenetic tools that may help in the understanding of soybean evolution.

In soybeans, phylogenetic studies have made considerable progress through the use of classical genetics, cytogenetics, and molecular genetics (Grant et al. 1984; Singh and Hymowitz 1985, 1988; Singh et al. 1988, 1992, 1993; Viviani et al. 1991; Kumar and Hymowitz 1989; Doyle et al. 1990; Doyle and Doyle 1993). Our current understanding of soybean taxonomy and phylogeny has been greatly improved in the decades since Hermann's 1962 revision (Hymowitz and Singh 1987; Doyle et al. 1990). Currently, the cultivated annual soybean (*G. max*) and its wild annual sister taxa, *G. soja* Sieb. & Zucc., are grouped in the subgenus *Soja*, while the 13 perennial soybean taxa represent the subgenus *Glycine* (see Table 1; Newell and Hymowitz 1980; Tindale 1984, 1986a, b; Tindale and Craven 1988).

Table 1 Species and genotypes of *Glycine* and relative genera[a]

Species	Abbreviation	2n	Accession[b]	Origin
Amphicarpa bracteata (L.) Fernald	ABR	22	Doyle 988, CU	Tompkins Co, N.Y.
Glycine max (L.) Merr.	BSR	40	BSR-101	Cultigen
G. max (L.) Merr.	A81	40	A81-356022	Cultigen
G. max (L.) Merr.	GPI	40	PI437.654	China
G. soja Seib. & Zucc.	GSJ	40	PI468.916	China
G. albicans Tind.	ALB	40	G2049	Australia
G. arenaria Tind.	ARE	40	G1931	Australia
G. argyrea Tind.	ARG	40	G1626	Australia
G. canescens F.J. Herm.	CAN	40	G1120	Australia
G. clandestina Wendl.	CLA	40	G1874	Australia
G. curvata Tind.	CUR	40	G1846	Australia
G. cyrtoloba Tind.	CYR	40	G1236	Australia
G. falcata Benth.	FAL	40	G2086	Australia
G. latifolia (Benth.) Newell & Hymowitz	LAT	40	G1497	Australia
G. latrobeana (Meissn.) Benth.	LTR	40	G1252	Australia
G. microphylla (Benth.) Tind.	MIC	40	G1901	Australia
G. tabacina (Labill) Benth	TAB	40	G2600	Australia
G. tomentella Hayata	TOM	40	G1366	Australia
Neonotonia verdcourtii Isly	NVE	22	*Peter* 43348, K	Africa
N. wightii (Arnott) Lackey	NWI	22	Doyle 1111, CU	US Virgin Islands
Teramnus labialis (L.f.) Spreng.	TLA	28	Doyle 1114, CU	US Virgin Islands
Phaseolus vulgaris L.	PVU	22	Big Sweet Baker	Cultigen

[a] Chromosome number included in this study was cited from Hymowitz (1970), Hymowitz and Singh (1987), Kumar et al. (1989)
[b] Accession designations include herbarium vouchers (e.g., Doyle 988, CU), Australia CSIRO Division of Plant Industry accessions (G 2049), and cultivar names (e.g., 'BSR-101')

Although these annual and perennial soybeans have different distributions and origins (Hymowitz 1970; Kumar and Hymowitz 1989; Table 1), they share the same number of chromosomes and can be interspecifically crossed with some success (Ahmad et al. 1984; Grant et al. 1984; Singh and Hymowitz 1985, 1988; Singh et al. 1988, 1992, 1993), which suggests a close relationship. The interspecific crossability and chloroplast DNA (cpDNA) restriction patterns suggest four to seven natural groups within the genus *Glycine* (Doyle et al. 1990). Close *Glycine* relatives have also been identified, including *Neonotonia*, *Teramnus*, *Amphicarpa*, and *Dumasia* (Kumar and Hymowitz 1989; Viviani et al. 1991; Doyle and Doyle 1993). However, because these relatives have different chromosome numbers from the expectations for the soybean ancestors (Kumar and Hymowitz 1989), the genome donors of soybean are still unknown.

While DNA sequences are a promising source of characters for phylogenetic studies (Miyamoto and Cracraft 1991), they rarely have been used in legumes for phylogenetic comparisons (Doyle et al. 1992). Two strategies have been used in characterizing the evolution of soybean. First, cytological surveys have been performed to identify diploid relatives that might represent progenitor species with 2n=2x=20 or 22 chromosomes (Kumar and Hymowitz 1989). Secondly, studies of affinities have been conducted among those available diploid relatives by chemosystematic (Fuchman 1985; Keen et al. 1986) and restriction fragment length polymorphism (RFLP) analyses (Doyle et al. 1992). A further definition of taxonomic relationship might be obtained by the detailed DNA sequence analysis of selected genetic loci.

A-199a is one of the duplicated RFLP loci (A-199a and A-199b), that has been genetically mapped previously (Keim et al. 1990a) and which fits a homoeologous model. The homoeology nature of these duplicated loci has been approved by T. Zhu et al. (personal communication). Unlike multigene families, homoeologous loci are generated by polyploidization, which results in two copies in tetraploid species, but only one in diploids. The characterization and comparison of homoeologous sequences will provide phylogenetic details of the genomic constitution, origin, and evolution of polyploidy species (Doyle et al. 1992), such as soybean.

In this study, annual and perennial species from the genus *Glycine* and several related diploid taxa with 2n=22 or 28 chromosomes within Glycininae were chosen as materials on which to examine potential diploid progenitors. The A-199a RFLP locus was used to represent a model for the soybean genome that has evolved from tetraploidy to a "diploidized tetraploidy" status due to its homoeologous nature. Here, we report results from the single-locus phylogenetic analysis of soybean and its relatives, with a focus on the diploid taxa that potentially donated genomic material for this polyploidy species.

Materials and methods

Germ plasm

From the tribe Phaseoleae (Leguminosae), 22 representative species and cultivars of the subtribe Glycininae and 1 species of subtribe Phaseolinae were used in this study (Table 1). Annual soybean (*Gly-*

cine subgenus *Soja*) representatives from *G. max* include one unadapted plant introduction (PI437658), one cultivar ('BSR-101'), and one Iowa State University breeding line (A81-356022). One example of the wild annual soybean species *G. soja* (PI468916) was also included. Thirteen perennial soybean taxa (subgenus *Glycine*) were included as the closest relatives of the annual soybeans. In addition, five soybean diploid relatives were included due to their chromosome number and cpDNA relationships (Doyle and Doyle 1993). *Neonotonia verdcourtii*, *N. wightii*, *Amphicarpa bracteata*, and *Teramnus labialis* were selected from Glycininae, and *Phaseolus vulgaris* was selected from Phaseoleae.

Genomic DNA preparations

Plants were grown in the Northern Arizona University and Cornell University greenhouses. Vouchers for these species were deposited in both of the Bailey hortorium of Cornell University and the Deaver herbarium of Northern Arizona University. Genomic DNA from each genotype and species used was extracted from leaf materials as described by Keim et al. (1988) or by Doyle and Doyle (1987). Purified genomic DNAs were then used as templates for polymerase chain reaction (PCR) amplification.

DNA clone sequencing and oligonucleotide design

The plasmid clone pA-199 (pBS$^+$ vector) contains a *Pst*I fragment derived from hypomethylated DNA of A81-356022 (Keim and Shoemaker 1988). Two genomic regions will hybridize to the pA-199 probe and are located on linkage group K (A-199a) and linkage group J (A-199b) of the ISU-USDA genetic map (Keim et al. 1990a; Shoemaker et al. 1992). Moreover, this clone has been approved to be homologous to the A-199a locus and homoeologous to the A-199b locus (Keim et al. 1990a; T. Zhu et al. personal communication). The distal regions of the *Pst*I insert fragment were sequenced by pUC/M13 sequencing primers according to Sanger et al. (1977). Two oligonucleotides were then designed using the 'Oligo' primer analysis software (National Biosciences) as primers for PCR amplifying the A-199a region in diverse taxa. These oligonucleotides are 5′-CGCACAACAAAACTGAGCTGCAAAGCCCG-3′ (primer A199L-2) and 5′-ATGCTTCAAAGACATCAGAGACAATCAAGTCTGATGAAAG-3′ (primer A199R-2). Failure of the PCR amplification from soybean total genomic DNA using two sets of A-199b specific primers A199bL-4 and A199bL-5 with A199R-2 (T. Zhu et al. personal communication) suggests that the designed primer A199R-2 is specific to the A-199a locus (data not shown).

PCR fragment amplification

PCR was used to prepare A-199a fragments for DNA sequencing. Amplification reactions were conducted in a total volume of 50 µl with a programmable thermal controller (MJ Research). The reaction mixture contained 50 m*M* KCl, 10 m*M* TRIS-HCl (pH 9.0), 1% Triton X-100, 1 m*M* $MgCl_2$, 0.2 m*M* dNTP mix, 10 pmol of each of the primers (A199L-2 and A199R-2), 10 ng of genomic DNA, and 5 units of *Taq* DNA polymerase (Promega). The reactions were performed for 40 cycles of 1 min at 94°C, 30 s at 60°C, and 30 s at 72°C. Amplifications were assayed by agarose gel electrophoresis at 80 V in 1 × TAE buffer on 0.7% agarose. PCR-amplified fragments were purified by Magic PCR columns (Promega) prior to DNA sequencing.

Sequencing of PCR fragments

Purified PCR fragments from each genotype and species were sequenced by using the oligonucleotides A199L-2 and A199R-2 labelled with [P^{32}]. Greater than 200 nucleotides of data were obtained from each side of the A-199a region for each accession (a total of 1260 bp in BSR-101, T. Zhu et al. personal communication). Only 200 nucleotides from each side were used in phylogenetic analysis to avoid sequencing errors that are more common distant from the primer and to create equal data sets for all taxa. All sequencing was triplicated to ensure the accuracy.

Sequence alignment and phylogenetic analysis

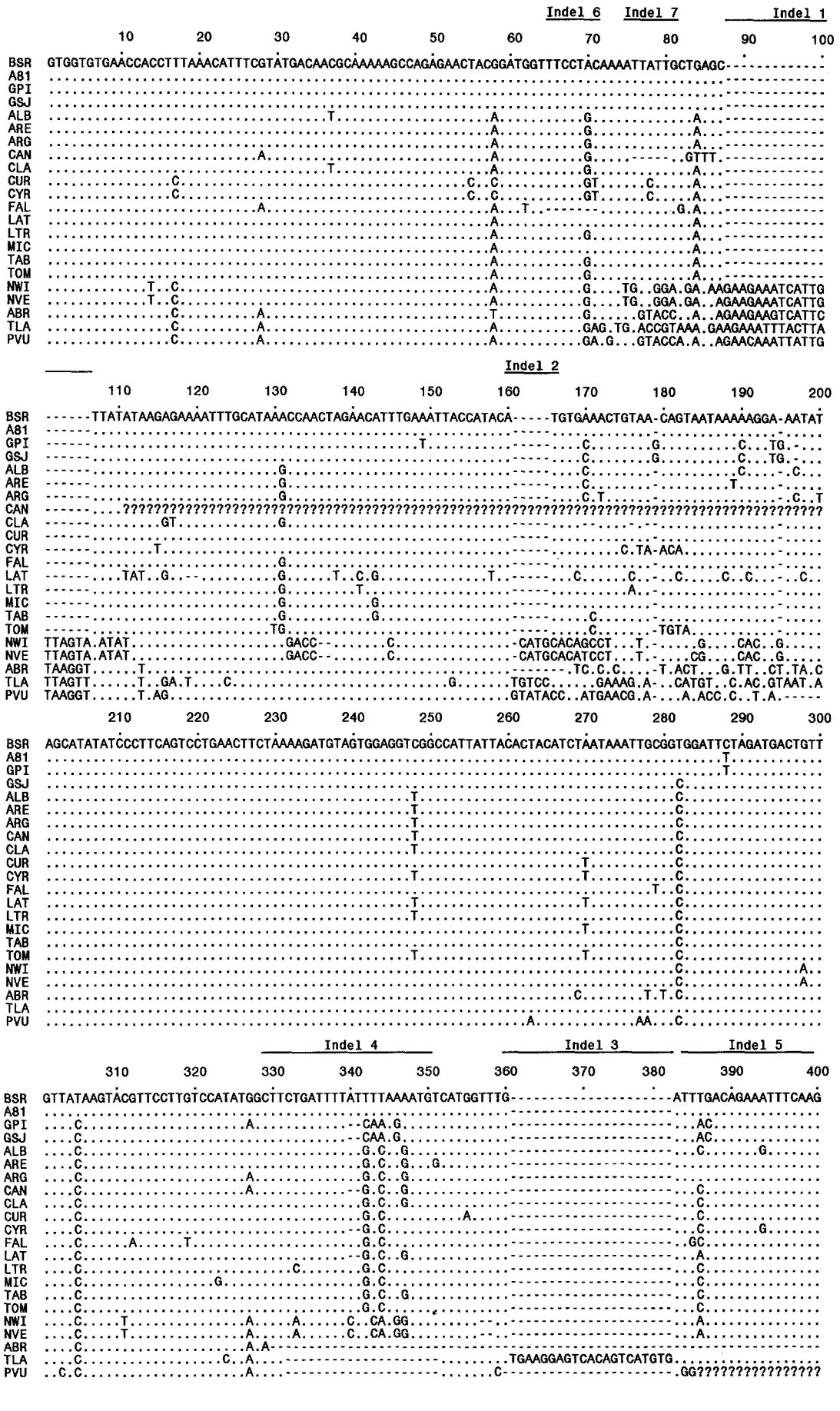
DNA sequences were aligned manually by sequential pairwise comparison or aligned by multiple sequential comparison using PC/GENE software (IntelliGenetics) with both open and unit gap cost at 10. Both methods generated the same alignment. Phylogenetic analyses were conducted using the computer program Phylogenetic Analysis Using Parsimony (PAUP 3.0s, Swofford 1991). In searching for the most parsimonious trees, a heuristic search was carried out with the following options: (1) the simple-addition sequence, (2) the closest-addition sequence, and (3) random-addition sequence, and then the tree bisection-reconnection (TBR) branch-swapping with MULPARS. Parsimony analysis was performed with and without insertion/deletion (indels) characters. When included, indels were treated as a single character with two states. The analysis presented in this study was performed on an unweighted data set. Different character weighting strategies were explored but did not alter their most parsimonious topologies and are not presented (see below). *Phaseolus vulgaris* sequences were used as an outgroup to root the phylogenetic trees.

Results

The PCR primers used in this study amplified only one fragment from each of the taxa used. In both annual and perennial soybeans, the amplified fragments represent only the A-199a locus since these primers can not PCR amplify the A-199b locus when it has been electrically separated from the A-199a locus. Subsequent experiments have shown that this is due to a lack of homology between the A199R-2 primer and the A-199b locus (T. Zhu et al. personal communication). Because there is only one A-199 locus in the soybean diploid relatives, the amplified fragments can only represent the single A-199 progenitor sequence presented in those taxa. Therefore, the sequences compared in this study were confirmed as the A-199a locus in soybeans, and the only A-199 locus in soybean's relatives.

Sequence analysis

Two hundred nucleotide sequences from each of the two distal A-199a regions were included in the study. One region (primer A199L-2) starts 110 nucleotides from the *Pst*I cloning site (110–310 bp), while the second (primer A199R-2) begins 120 nucleotides from the opposite *Pst*I site (120–320 bp) and proceeds towards the middle of the cloned insert. The 400 nucleotides included in this study represent greater than 30% of the total A-199a 1260 bp insert (T. Zhu et al. personal communication). The G+C content of these sequences averaged 36.3% with a range of 35.2% to 37.5% for the different taxa. Alignment of the DNA sequences was easily accomplished even though 27.3% of all nucleotides were present as an indel in at least 1 taxon. Among nucleotides varying due to indels, 47.7%


Indel 6
Indel 7
Indel 1
Indel 2
Indel 4
Indel 3
Indel 5
BSR
A81
GPI
GSJ
ALB
ARE
ARG
CAN
CLA
CUR
CYR
FAL
LAT
LTR
MIC
TAB
TOM
NWI
NVE
ABR
TLA
PVU

were in indel 5, 19.3% in indel 3, while 17.4% were in indel 1 (Figs. 1 and 2). Within the 400 aligned positions, 129 (32.3%) sites were variable as nucleotide substitutions in 1 or more taxa. There were 31% transitions, 31% transversions, and 38% with both transitions and transversions. Sixty-six percent of the variable sites (79) are informative in the Wagner parsimony analysis (see below).

The pairwise nucleotide dissimilarities between taxa indicated that there was only 0.3% variation between the annual cultivated soybean genotypes BSR-101 and A81-356022 (Table 2). However, there was a 2.2% nucleotide difference between the *G. max* cultivars (BSR-101 and A81-356022) and the unadapted *G. max* accession (PI437.654). This was the same observed dissimilarity between annual cultivars and wild annual soybean *G. soja*. The nucleotide variation among the 13 perennial soybean species ranged from 1.7% to 8.4%. The observed dissimilarity between annual soybeans and perennial soybeans ranged from 3.0% to 7.0%. There was 12.6–15.3% nucleotide differences between annual soybeans and taxa from outside of the genus *Glycine*, while 8.2–16.4% nucleotide differences existed between perennial soybeans and those taxa.

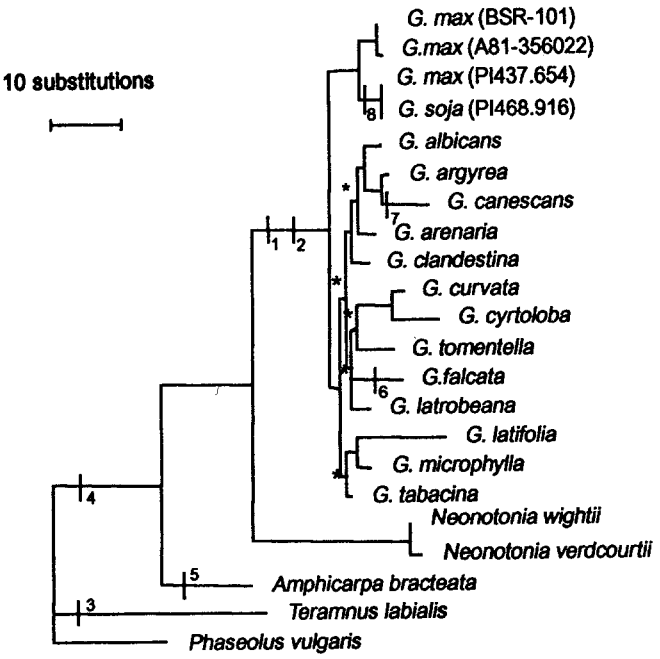


Fig. 2 One of the 16 most parsimonious trees derived from the equal weighted analysis by PAUP (length=246 steps, consistency index=0.78, consistency index for informative characters=0.69, retention index=0.71). Nodes that did not occur in all most parsimonious trees are marked by an *asterisk* (*). Indels position is superimposed on the tree with *numbers* indicating the locations of the indels within the sequences (Fig. 1). Length of the branches are proportional to the number of substitutions inferred

Phylogenetic analysis

Phaseolus vulgaris was used as the outgroup for rooting tree topology in these analyses because it is not considered to be a particularly close soybean relative based upon its morphology, anatomy, cytogenetics (Lackey 1981), and cpDNA phylogeny (Doyle and Doyle 1993). It is currently placed in a sister subtribe (Phaseolinae) to the other genera in this study. In addition, the A-199a nucleotide differences between annual soybean and *P. vulgaris* were among the highest observed in this study (Table 2). *Teramnus labialis* also was used as an outgroup in other analyses without changing the consensus tree topology (data not presented).

The heuristic analysis of unweighted nucleotide sequences without indels resulted in 16 equally parsimonious trees. The trees were 246 steps with a consistency index of 0.78 using all characters. Exclusion of autapomorphic characters reduced the consistency index to 0.69. A single representative topology from the 16 most parsimonious trees is presented in Fig. 2. A strict consensus tree was calculated from the 16 trees (Fig. 3). Examination of these figures reveals that five genera have been well resolved by multiple synapomorphic nucleotide changes. *Neonotonia* is the closest genus to *Glycine*, followed by *Amphicarpa*, then *Teramnus*. These data support the monophyly of the genus *Glycine*, as well as that of the two individual *Glycine* subgenera. However, the four accessions from the subgenus *Soja* are resolved into two clades that do not represent the 2 recognized species. For example, the *G. max* (PI437.654) and *G. soja* (PI468916) accessions form a single, well-supported clade. This is not surprising given the high fertility and sympatry of these 2 species. Previous studies have observed only few fixed RFLP alleles between these taxa (Keim et al. 1989), which is consistent with the idea that they are conspecific.

This analysis separated only 8 of the 13 perennial soybean taxa into completely resolved clades (Figs. 2 and 3): Group A includes *G. albicans*, *G. arenaria*, *G. argyrea*, and *G. canescens*; Group B includes *G. latifolia* and *G. microphylla*; and Group C includes *G. curvata*, and *G. cyrtoloba*. The taxa *G. clandestina*, *G. falceta*, *G. latrobeana*, *G. tabacina*, and *G. tomentella* were placed in unique positions in each of the 16 different trees and are part of the large polytomy present in the consensus tree for the subgenus *Glycine* (Fig. 3).

The use of indels as characters in parsimony is logical, but their weighting relative to nucleotide substitutions is problematic. We have only presented results from analyses where indels were coded as missing data. However,

Fig. 1 Aligned nucleotide sequences of A-199a locus from soybeans and relatives. *Vertical columns* represent nucleotide positions. *Dots* in the *lines* indicate sequence identity with *G. max* (BSR-101), and *dashes* are gaps required for alignment of the sequences. *?* indicates nucleotides of unknown identity, and *blanks* represent sequence not determined. *Italic numbers above* sequences show the location of gaps corresponding to inferred deletions/insertions superimposed on the phylogenetic tree in Fig. 2. Nucleotides 1–200 were determined using the A199L-2 primer, while 201–400 represents the opposite side of the pA-199a region and was sequenced with the A199R-2 primer

Table 2 Pairwise distances between taxa generated from PAUP. Below diagonal: absolute distance; above diagonal: mean distances (adjusted for missing data)

	BSR	AB1	GPI	GSJ	ALB	ARE	ARG	CAN	CLA	CUR	CYR	FAL	LAT	LTR	MIC	TAB	TOM	NWI	NVE	ABR	TLA	PVU
BSR	–	0.003	0.022	0.022	0.047	0.039	0.044	0.070	0.039	0.039	0.061	0.045	0.070	0.036	0.030	0.033	0.047	0.134	0.137	0.126	0.149	0.150
A81	1	–	0.025	0.025	0.050	0.041	0.047	0.074	0.042	0.041	0.064	0.048	0.073	0.039	0.033	0.036	0.050	0.137	0.140	0.129	0.152	0.153
GPI	8	9	–	0.000	0.038	0.045	0.034	0.053	0.040	0.045	0.068	0.051	0.068	0.042	0.036	0.034	0.053	0.125	0.127	0.123	0.152	0.153
GSJ	8	9	0	–	0.036	0.045	0.034	0.053	0.040	0.045	0.068	0.051	0.068	0.042	0.036	0.034	0.053	0.125	0.127	0.123	0.152	0.153
ALB	17	18	13	13	–	0.019	0.019	0.044	0.017	0.044	0.050	0.045	0.059	0.028	0.033	0.025	0.036	0.137	0.140	0.119	0.164	0.156
ARE	14	15	16	16	7	–	0.017	0.044	0.020	0.041	0.053	0.042	0.056	0.025	0.030	0.022	0.033	0.134	0.137	0.119	0.155	0.147
ARG	16	17	12	12	7	6	–	0.037	0.025	0.047	0.058	0.048	0.059	0.030	0.036	0.028	0.039	0.128	0.131	0.113	0.146	0.147
CAN	19	20	14	14	12	12	10	–	0.041	0.068	0.063	0.068	0.048	0.044	0.055	0.041	0.044	0.082	0.082	0.094	0.099	0.114
CLA	14	15	14	14	6	7	9	11	–	0.036	0.051	0.037	0.048	0.017	0.025	0.014	0.028	0.130	0.133	0.127	0.145	0.136
CUR	14	15	16	16	16	15	17	18	13	–	0.028	0.042	0.070	0.033	0.028	0.030	0.039	0.137	0.140	0.119	0.149	0.144
CYR	22	23	24	24	18	19	21	17	18	10	–	0.065	0.064	0.050	0.050	0.053	0.047	0.152	0.154	0.135	0.161	0.150
FAL	16	17	18	18	16	15	17	18	13	15	23	–	0.074	0.034	0.031	0.031	0.045	0.151	0.154	0.135	0.161	0.147
LAT	25	26	24	24	21	20	21	13	17	25	30	26	–	0.063	0.047	0.050	0.061	0.155	0.158	0.156	0.165	0.164
LTR	13	14	15	15	10	9	11	12	6	12	18	12	19	–	0.022	0.019	0.028	0.134	0.137	0.126	0.152	0.144
MIC	11	12	13	13	12	11	13	15	9	10	18	11	17	8	–	0.014	0.028	0.137	0.140	0.129	0.158	0.150
TAB	12	13	12	12	9	8	10	11	5	11	19	11	18	7	5	–	0.025	0.123	0.126	0.123	0.152	0.141
TOM	17	18	19	19	13	12	14	12	10	14	17	16	22	10	10	9	–	0.142	0.145	0.129	0.161	0.153
NWI	48	49	44	44	49	48	46	22	46	49	54	53	55	48	49	44	51	–	0.005	0.168	0.199	0.192
NVE	49	50	45	45	50	49	47	22	47	50	55	54	56	49	50	45	52	2	–	0.168	0.199	0.195
ABR	39	40	38	38	37	37	35	21	39	37	42	41	48	39	50	38	40	55	55	–	0.176	0.130
TLA	51	52	52	52	56	53	50	25	49	51	55	54	56	52	54	52	55	72	72	58	–	0.154
PVU	48	49	49	49	50	47	47	27	43	46	48	46	52	46	48	45	49	65	66	42	53	–

their inclusion in the analysis as single characters each did not change the topologies generated (data not presented). Their relationship with nucleotide characters is presented in Fig. 2, where the phylogenetic distribution of indels is mapped onto the nucleotide substitution based phylogeny. All eight indels represent synapomorphic or autapomorphic characters on this tree. Many of the indels are deeply rooted, but three (6, 7, and 8) are in terminal branches. The synapomorphic indel 8 supports the clade containing both annual soybean species. Indels 1 and 2 support the monophyly of the genus *Glycine*. The indel data are consistent with the proposed phylogeny and support the relevance of indels in understanding evolution at a molecular level.

Discussion

Although single-locus phylogenies and species trees may be incongruent due to lineage sorting or hybridization (Pamilo and Nei 1988; Doyle 1992), the phylogenetic analysis of DNA sequences can provide a large number of multistate characters. These types of data can be used to provide detailed and accurate phylogenies for the loci studied and important inferences for the species. In this report, we have demonstrated the potential of nuclear sequences for phylogeny construction in the Glycininae, and in the genus *Glycine* in particular. Our results correlate well with recent phylogenetic studies on subtribe Glycininae using crossing compatibility and cpDNA RFLP analyses (Kumar and Hymowitz 1989; Doyle and Doyle 1993). This congruence argues that for these taxa, this single-locus analysis is representative of the species phylogeny and provides detailed relationships among cultivated soybean, wild soybean, and related genera based upon nuclear DNA sequences.

Some of the relationships among Glycininae genera included in this study were poorly understood prior to this study. According to Lackey (1981), *Glycine* and *Teramnus* belong to the *Glycine* group, while *Amphicarpa* and *Neonotonia* belong to the *Shuteria* group. However, a recent cpDNA phylogenetic analysis positioned these four genera in a single Glycininae group without further resolution (Doyle and Doyle 1993). By using genomic DNA sequence data, we were able to resolve the relationships among the closely related genera to the soybeans. The genera most closely related to *Glycine* are *Neonotonia*, followed by *Amphicarpa* and *Teramnus*. *Neonotonia* was considered to be one of the *Glycine* members until two decades ago (Lackey 1977), mainly due to the differences in chromosome size, number, and morphology. *Neonotonia* has a chromosome number of 2n=22 or 44, while the *Glycine* species have 2n=38, 40, 78, or 80 (Kumar and Hymowitz 1989). Lackey (1980) proposed that the unique chromosome number of *Glycine* maybe derived from diploid ancestors with a base number of X=11, followed by aneuploid reduction to a base number of X=10. *Glycine* then arose in a subsequent tetraploid event. Our previous studies suggested that the soybean genome is 25% smaller than expected for a "full" tet-

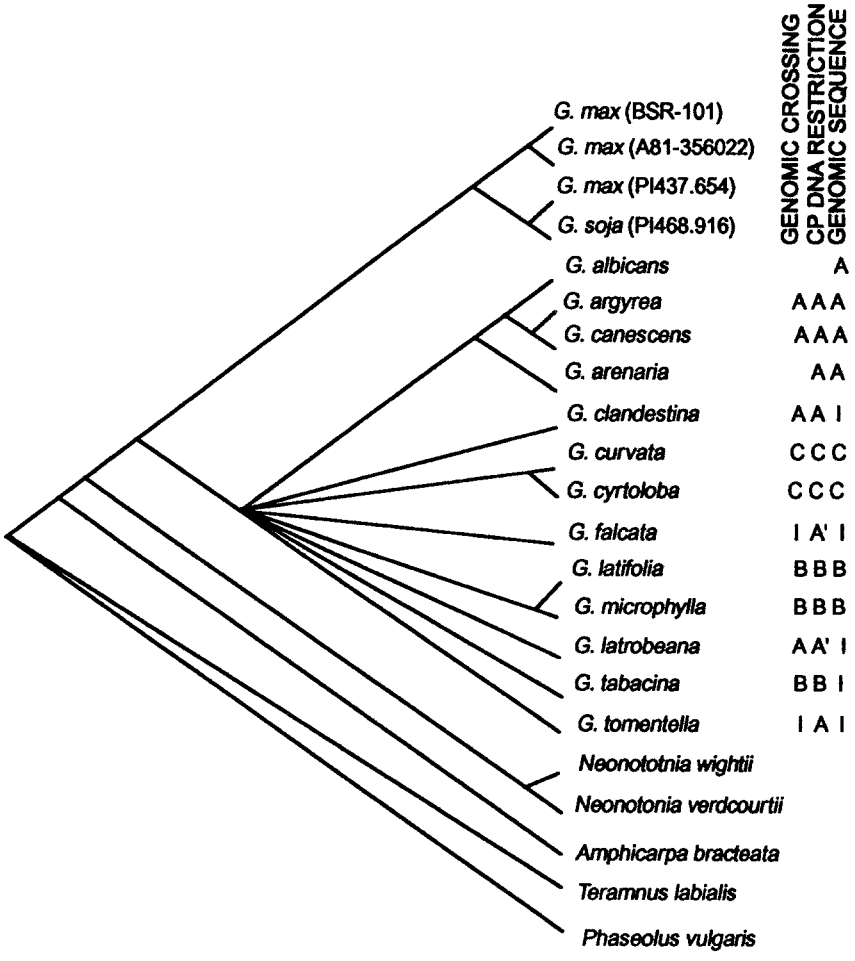


Fig. 3 Strict consensus tree of the 16 most parsimonious trees derived from equally weighted analysis by PAUP. *Phaseolus* was designated as a co-outgroup. Comparison to crossing experiments and cpDNA restriction data within perennial soybeans was made following the taxa name. *A* represents taxa that belong to the genome type A or plastome type A, *A'* indicates those taxa with a loose relationship with genotype A or plastome type A, *B* represents the taxa of the genome type B or plastome type B. Taxa within the genome type C and plastome type C group is indicated by *C*. *I* represents the taxa that are independent to types A, B, and C. Data from crossing experiment and cpDNA restriction are cited from Doyle et al. 1990

raploid in terms of duplicated sequences, and that this 25% could account for 2 large chromosomes (Zhu et al. 1994). The evidence of 25% sequence reduction of the duplicated genome, along with the current study, suggests that *Neonotonia* may be the closest reasonable candidate for a soybean genomic donor. This model requires an autotetraploid intermediate. However, a second model involving an allotetraploid ancestor seems just as likely.

It has been proposed that soybean arose from an allotetraploid ancestor and is currently evolving towards diploidy through the progressive loss of genes (Hymowitz and Singh 1987; Zhu et al. 1994). Such evolution might have involved an initial hybridization of two different species (or even genera) with 11 haploid chromosomes, then chromosome doubling to 1n=22 followed by the loss of 2 chromosomes. This invokes the well-documented model of allotetraploid evolution coupled with chromosome loss. Because the duplicated genomes of soybean are very different at the molecular level (T. Zhu et al. personal communication), we suggest the either the ploidy event was very ancient or that the taxa hybridizing were very distinct. We are currently studying the A-199a genomic duplicate sequence (A-199b) to construct phylogenies for both homoeologous genomes. The preliminary data from A-199b phylogenetic studies suggested that *Teramnus* may have close evolutionary relationships to the soybean secondary genome donors (T. Zhu et al. personal communication). Such data could be useful in resolving relationships within the Glycininae.

The genus *Glycine* appears as a monophyletic group in our study, supporting the current taxonomic relationships in this genus (Hymowitz and Singh 1987). The two subgenera, *Glycine* and *Soja*, are well-defined (Fig. 3) and consistent with results from previous studies (Hymowitz and Singh 1987). In the subgenus *Glycine*, the perennial soybeans are loosely united into three groups: Group A consists of *G. albicans*, *G. arenaria*, and *G. canescens*. Group B includes *G. latifolia* and *G. microphylla*. Group C includes *G. curvata* and *G. cyrtoloba*. This three-group tree is congruent with the tree generated from cpDNA analysis, though less resolute. The subgenus *Glycine* topology is dominated by an unresolved polytomy (Fig. 3) that is due to homoplasious characters. Five species, *G. clandestina*, *G. falcata*, *G. latrobeana*, *G. tabacina* and *G. tomentella*, are not resolved in the consensus tree. Among these five species, *G. falcata*, *G. latrobeana*, and *G. tomentella* are independent or loosely associated with 1 of the clades in previous genomic crossing studies (see Doyle et al. 1990). The lack of synapomorphic characters to resolve this group might be due to a recent radi-

ation or slower evolutionary rates than observed in the annual soybean. Slower evolutionary rates have been observed in longer-lived species (e.g., Bousquet et al. 1992), and this may account for the lack of informative characters in this group. In contrast, synapomorphic characters, autapomorphic characters, and an indel resolved three of the four annual soybean accessions in this study (Fig. 2).

Although organellar genomes had low diversity between the wild and cultivated annual soybean (Sission et al. 1978; Shoemaker et al. 1986; Close et al. 1989; Keim et al. 1989), nuclear genome analysis indicated that the two species are almost identical in terms of number and size of chromosomes (Singh and Hymowitz 1988) and their 5S and 18-25S ribosomal genes (Doyle and Beachy 1985; Doyle 1988). Even the cDNA sequence analysis of glycinin only shows 0.4% nucleotide changes between these two species (Zakharova et al. 1989). Our study, again, suggests that differences between these two species is not greater than the difference between the cultivar and the wild accession within *G. max*. The morphological differences between these taxa are due primarily to human selection during domestication and may not always be detected at the molecular level because this may involve relatively few genetic loci (Keim et al. 1990a, b).

Acknowledgments We thank A.H.D. Brown for making available accessions from the CSIRO Division of Plant Industry (Canberra, Australia) perennial *Glycine* germ plasm collection, J.F. Wendel for *N. vercourdtii*, and T. Ayers and R. Scott for critical reading for the manuscript.This work was supported by USDA grant no. 92-37300-7523 to PK, and NSF grant BSR 9107480 to JJD.

References

Ahmad QN, Britten EJ, Byth DE (1984) The karyotype of *Glycine soja* and its relationship to that of the soybean, *Glycine max*. Cytologia 49:645–658

Bousquet J, Strauss SH, Doerksen AH, Price RA (1992) Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. Proc Natl Acad Sci USA 89:7844–7848

Close PS, Shoemaker RC, Keim P (1989) Distribution of restriction site polymorphism within the chloroplast genome of the genus *Glycine*, subgenus *Soja*. Theor Appl Genet 77:768–776

Doyle JJ (1988) 5S ribosomal gene variation in the soybean and its progenitor. Theor Appl Genet 75:621–624

Doyle JJ (1992) Gene trees and species trees: molecular systematics as one-character taxonomy. Syst Bot 17:144–163

Doyle JJ, Beachy RN (1985) Ribosomal gene variation in soybean and it relatives. Theor Appl Genet 70:369–376

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15

Doyle JJ, Doyle JL (1993) Chloroplast DNA phylogeny of the papilionoid legume tribe Phaseoleae. Syst Bot 18:309–327

Doyle JJ, Doyle JL, Brown AHD (1990) A chloroplast DNA phylogeny of the wild perennial relatives of soybean (*Glycine* subgenus *Glycine*): congruence with morphological and crossing groups. Evolution 44:371–389

Doyle JJ, Lavin M, Bruneau A (1992) Contribution of molecular data to Papilionoid legume systematics. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Chapman & Hall, New York London, pp 223–251

Fuchman WH (1985) Discrepancies among published amino acid sequences of soybean leghemoglobins: experimental evidence against cultivar differences as the sources of the discrepancies. Arch Biochem Biophys 243:454–460

Grant JE, Grace JP, Brown AHD, Putievsky (1984) Interspecific hybridization in *Glycine* Willd. subgenus *Glycine* (Legumeinosae). Aust J Bot 32:655–663

Hermann FJ (1962) Revision of the genus *Glycine* and its immediate allies. USDA Tech Bull 1268:1–82

Hymowitz T (1970) On domestication of the soybean. Econ Bot 24:408–421

Hymowitz T, Singh RJ (1987) Taxonomy and speciation. In: Wilcox JR (ed) Soybeans: improvement, production, and uses, 2nd edn. (Agron Ser 16.) ASA, CSSA, SSSA, Madison, Wis., pp 23–48

Keen NT, Lyne RL, Hymowitz T (1986) Phytoalexin production as a chemosystematic parameter within the genus *Glycine*. Biochem System Ecol 14:481–496

Keim P, Shoemaker RC (1988) Construction of a random recombinant DNA library that is primarily single copy sequence. Soybean Genet Newsl 15:147–148

Keim P, Olson TC, Shoemaker RC (1988) A rapid protocol for isolating soybean DNA. Soybean Genet Newsl 15:150–152

Keim P, Shoemaker RC, Palmer RG (1989) RFLP diversity in soybean. Theor Appl Genet 77:786–792

Keim P, Diers BW, Olson T, Shoemaker RC (1990a) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Keim P, Biers BW, Shoemaker RC (1990b) Genetic analysis of soybean hard seededness with molecular markers. Theor Appl Genet 79:465–469

Kenworthy WJ (1989) Potential genetic contributions of wild relatives to soybean improvement. In: Pascale AJ (ed) Proc World Soybean Res Conf IV. Assoc Argentina de La *Soja*, Buenos Aires, pp 883–888

Kumar PS, Hymowitz T (1989) Where are the diploid (2n=2x=20) genome donors of *Glycine* Willd. (Leguminosae, Papilionoideae)? Euphytica 40:221–226

Lackey JA (1977) Neonotonia, a new generic name to include *Glycine wightii* (Arnott) Verdcourt (Leguminosae, Papilionoideae). Phytologia 37:209–212

Lackey JA (1980) Chromosome numbers in the Phaseoleae (Fabaceae: Faboideae) and their relation to taxonomy. Am J Bot 67:595–602

Lackey JA (1981) Phaseoleae. In: Polhill RN, Raven PH (eds) Advances in legume systematics, part 1. Royal Botanical Gardens, Kew, England, pp 301–328

Miyamoto MM, Cracraft J (1991) Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. In: Miyamoto MM, Cracraft J (eds) Phylogenetic analysis of DNA sequences. Oxford University Press, New York Oxford, pp 3–17

Newell CA, Hymowitz T (1980) Taxonomic revision in the genus *Glycine* subgenus *Glycine* (Leguminosae). Brittonia 32:63–69

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Shoemaker RC, Hatfield PM, Palmer RG, Atherly AG (1986) Chloroplast DNA variation in the genus *Glycine* subgenus *Soja*. J Hered 77:26–30

Shoemaker RC, Guffy RD, Lorenzen LL, Specht JE (1992) Molecular genetic mapping of soybean: map utilization. Crop Sci 32: 1091-1098

Singh RJ, Hymowitz T (1985) The genomic relationships among six wild perennial species of the genus *Glycine* subgenus *Glycine* Willd. Theor Appl Genet 71:221–230

Singh RJ, Hymowitz T (1988) The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor Appl Genet 76:705–711

Singh RJ, Kollipara KP, Hymowitz T (1988) Further data on the genomic relationships among wild perennial species (2n=40) of the genus *Glycine* Willd. Genome 30:166–176

Singh RJ, Kollipara KP, Hymowitz T (1992) Genomic relationships among diploid wild perennial species of the genus *Glycine* Willd.

subgenus *Glycine* revealed by meiotic chromosome pairing and seed protein electrophoresis. Theor Appl Genet 85:276–282

Singh RJ, Kollipara KP, Hymowitz T (1993) Backcross (BC_2-BC_4)-derived fertile plants from *Glycine max* (L.) Merr. and *G. tomentella* Hyata intersubgeneric hybrids. Crop Sci 33:1002–1007

Sission VA, Brim CA, Levings CS III (1978) Characterization of cytoplasmic diversity in soybeans by restriction endonuclease analysis. Crop Sci 18:991–996

Song K, Osborn TC, Williams PH (1990) *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). 3. Genome relationships in Brassica and related genera and the origin of *B. oleracea* and *B. rapa* (syn. campestris). Theor Appl Genet 79:497–506

Swofford DL (1990) PAUP: Phylogenetic analysis using parsimony, version 3.0. Illinois Natural History Survey, Champaigne, Ill.

Tindale MD (1984) Two new eastern Australian species of *Glycine* Willd. (Fabaceae). Brunonia 7:207–213

Tindale MD (1986a) A new North Queensland species of *Glycine* Willd. (Fabaceae). Brunonia 9:99–103

Tindale MD (1986b) Taxonomic notes on three Australian and Norfolk Island species of *Glycine* Willd. (Fabaceae: Phaseoleae) including the choice of a neotype for *G. clandestina* Wend. Brunonia 9:179–191

Tindale MD, Craven LA (1988) Three new species of *Glycine* (Fabaceae: Phaseoleae) from north-western Australia, with notes on amphicarpy in the genus. Aust Syst Bot 1:399–410

Viviani T, Conte L, Cristofolini G, Speranza M (1991) Sero-systematic and taximetric studies on the Phaseoleae (Fabaceae) and related tribes. Bot J Linn Soc 105:113–136

Wendel JF (1989) New world tetraploid cotton contain old world cytoplasm. Proc Natl Acad Sci USA 86:4132–4136

Zaknarova ES, Epishin SM, Vinetski YuP (1989) An attempt to elucidate the origin of cultivated soybean via comparison of nucleotide sequences encoding glycinin B_4 polypeptide of cultivated soybean, *Glycine max*, and its presumed wild progenitor, *Glycine soja*. Theor Appl Genet 78:852–856

Zhu T, Schupp JM, Oliphant A, Keim P (1994) Hypomethylated sequences: characterization of the duplicate soybean genome. Mol Gen Genet 244:638–645