Identification of presumed ancestral DNA sequences of phaseolin in Phaseolus vulgaris

(molecular evolution/seed protein/crop evolution/tandem repeat/polymerase chain reaction)

JAMES KAMIT,* VIVIANA BECERRA VELÁSQUEZ,** DANIEL G. DEBOUCK§, AND PAUL GEPTS*

*Department of Agronomy and Range Science, University of California, Davis, CA 95616-8515; and §Genetic Resources Unit, Centro Internacional de Agricultura Tropical, Apartado Aéro 6713, Cali, Colombia

Communicated by R. W. Allard, University of California, Davis, CA, November 17, 1994

ABSTRACT Common bean (Phaseolus vulgaris) consists of two major geographic gene pools, one distributed in Mexico, Central America, and Colombia and the other in the southern Andes (southern Peru, Bolivia, and Argentina). Amplification and sequencing of members of the multigene family coding for phaseolin, the major seed storage protein of the common bean, provide evidence for accumulation of tandem direct repeats in both introns and exons during evolution of the multigene family in this species. The presumed ancestral phaseolin sequences, without tandem repeats, were found in recently discovered but nearly extinct wild common bean populations of Ecuador and northern Peru that are intermediate between the two major gene pools of the species based on geographical and molecular arguments. Our results illustrate the usefulness of tandem direct repeats in establishing the polarity of DNA sequence divergence and therefore in proposing phylogenies.

Phaselin, the major seed storage protein of Phaseolus vulgaris, the common bean (1), is encoded by a small gene family of 6–10 tightly linked sequences on linkage group D7 (2–5). Extensive polyacrylamide gel electrophoretic studies of genetic variation at the phaselin locus (Phs) have allowed us to identify patterns of multiple domestication and a reduction in genetic diversity during domestication in the common bean (6–11). The two most prevalent phaselin protein types among cultivars are the S and T types (6, 12). The former is found among cultivars domesticated from wild beans in the Middle American center of origin (Mexico, Central America, and Colombia), whereas the latter is found among cultivars descended from wild beans in the Andean center of origin in southern Peru, Bolivia, Argentina, and Chile (6, 12, 13). Wild beans of the Middle American center show S phaselin type or other phaselin types, whereas Andean wild beans show T phaselin types or other types depending on the accession (6–8). Recently, wild bean populations have been found in northern Peru and Ecuador (14), which have an I (Inca) phaselin type that had not been described previously in wild or cultivated beans (8, 14). Allozyme analyses suggested that these populations were intermediate between the Middle American and Andean gene pools (14).

In the present study, we performed analyses of phaselin diversity at the DNA level in an attempt to account for the intermediate nature of the recently discovered wild populations. Prior to this study, DNA sequence analysis of genomic or cDNA clones of the S and T phaselin types had shown high levels of sequence similarity among members of the gene family within and between the S and T phaselin types (2, 15–17). The T phaselin gene family consists of two subfamilies—α and β. Members of the α subfamily have tandem direct repeat sequences of 15 bp in the fourth exon or 27 bp in the sixth exon, whereas members of the β subfamily do not have these repeats (17). The S phaselin gene family also consists of α and β subfamilies, with or without the 27-bp tandem direct repeat, respectively. No S phaselin gene has been identified thus far with a 15-bp tandem direct repeat, although this result may be due to insufficient sampling of S phaselin genes (15, 16). Further sequence analyses have also revealed that α phaselin genes of the S and T phaselin types show higher sequence similarity among each other than to the β phaselin genes of the S and T phaselin types, respectively, and conversely for the β phaselin genes. These comparisons suggest that the divergence of α and β phaselin genes predates the divergence of the S and T phaselin genes (16). No DNA sequence information was available for the I phaselin type prior to this study.

Specifically, we have performed PCR and DNA sequence analyses on the part of the phaselin sequence that contains the 15-bp sequence that is duplicated in certain members of the gene family. Our main findings suggest that the 15- and 27-bp tandem direct repeats, and a newly discovered 21-bp tandem direct repeat, have a distinct geographical distribution and that the presumed ancestral phaselin type, which lacks any of the tandem direct repeats, is present in the recently discovered and nearly extinct wild P. vulgaris of northern Peru and Ecuador.

MATERIALS AND METHODS

Plant Material. Seventy-nine accessions of P. vulgaris were analyzed, consisting of 31 wild and 48 cultivated accessions. They were representative of the diversity included in the species based on morphology (18, 19), phaselin and isozyme diversity (6–8, 13, 20), restriction fragment length polymorphisms for single-copy nuclear sequences and M13-related sequences (11, 21), and reproductive isolation (22, 23). The sample included 10 wild and 28 cultivated accessions from the Middle American gene pool (Mexico, Central America, and Colombia), 11 wild and 8 cultivated entries from the Andean gene pool (southern Peru, Chile, and Argentina), and 10 wild and 12 cultivated accessions from the intermediate area in Ecuador and northern Peru. A segregation analysis was conducted in an F2 population of the cross BAT93 × Jalo EEEP58 described earlier (4).

Genomic DNA Isolation. Isolation of genomic DNA from leaves was as described (24). Genomic DNA from seeds was extracted as follows. Five to 30 mg of the seed flour was incubated at 65°C for 1 h in 350 μl of extraction buffer [65 mM Tris-HCl, pH 7.5/500 mM NaCl/10 mM EDTA/2% SDS/1% 2-mercaptoethanol/10% (vol/vol) glycerol]. One-tenth volume of 3 M NaOAc (pH 5.5) was added and the mixture was centrifuged for 5 min at 12,000 × g. The supernatant was mixed with an equal volume of isopropanol and held at room

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

1Present address: Section of Molecular and Cell Biology, Division of Biological Sciences, University of California, Davis, CA 95616.
2Present address: Universidad Adventista de Chile, Chilán, Chile.
temperature for 5 min to overnight. The precipitated material was collected by centrifugation for 5 min, washed with 70% ethanol, and allowed to air dry. The pellet was resuspended in 300 µl of TE buffer for 30–60 min. Undissolved material was removed by centrifugation for 5 min and the supernatant was precipitated with the addition of 1/5th vol of 5 M NaCl and 2 vol of 95% ethanol. After storage overnight at −20°C, the precipitate was collected by centrifugation for 10 min, washed with 70% ethanol, air-dried, and redissolved in 100 µl of TE buffer.

PCR Assay and DNA Sequencing. The PCR primers were selected from regions of complete identity between the T and S phaseolin sequences. The sequence for the upstream primer is 5'-AGCATATTCAGGCTCCTC-3', and the sequence for the downstream primer is 5'-GCTCATTTGTTTCTACATC-3'. Optimal PCR conditions were 2 mM MgCl₂, 200 µM NTP, 0.2 µM primers, and 40 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s, with 1–25 ng of DNA. Amplification products were visualized in 10% polyacrylamide gels in 10× TBE. PCR fragments were isolated by the method of Danner (25). Sequencing reactions were performed according to the manufacturer's instructions for Sequenase version 2.0 (United States Biochemical) except that the labeling mixture was diluted 1:8 and 1 µl of Mn buffer (10 mM MnCl₂/50 mM NaCl) was included.

SDS/PAGE and Southern Blot of Phaseolin. Tissue from the same seeds analyzed by PCR was also subjected to SDS/PAGE as described (24). Southern hybridization with a phaseolin probe was performed as described (4, 24).

RESULTS

PCR Amplification. A PCR assay was designed specifically to amplify a region surrounding the 15-bp tandem direct repeat. This region included 78 bp of the fourth exon in which the 15-bp sequence motif is located, as well as the entire third intron (124 bp) and 6 bp of the third exon to provide sufficient fragment size for improved resolution by PAGE. If none of the phaseolin genes in a genotype contained the 15-bp tandem direct repeat, we expected to detect a single band of 249 bp. If different members of the family either lacked or contained the 15-bp repeat, we expected to detect two bands of 249 and 264 bp, respectively. However, the PCR assay actually resulted in a more complex banding pattern, especially in nonadenaturing PAGE compared to agarose gel electrophoresis (Fig. 1). In particular, the apparent molecular weight of some amplified fragments was much higher than expected. Repeated electrophoreses with the same or different genomic DNA extracts from cultivars Sanilac and Tendergreen showed reproducible patterns. In addition, the progeny of an F₂ cross between BAT93, an S-type phaseolin cultivar, and Jalo EEP558, a T-type phaseolin cultivar (4), was screened by the phaseolin 15-bp PCR assay, one-dimensional SDS/PAGE of total seed protein, and Southern hybridization to a phaseolin probe. Results demonstrated a perfect correlation among the results of the three assays. Among the 43 F₂ individuals tested, 9 showed a T phaseolin, 7 showed an S phaseolin, and 27 were heterozygous for phaseolin type.

Sequencing of PCR Products. Sequencing of individual PCR products in polyacrylamide gels was performed after electroblotting of the PCR products onto DEAE membranes. The DEAE electroblopting method (25) proved to be highly reproducible, with sufficient yields to permit direct sequencing of the isolated fragments.

Sequence analysis of the isolated fragments showed that the smaller molecular size bands (240–300 bp) were true amplification products and the larger molecular size bands (>300 bp) were heteroduplexes formed through various combinations of the lower bands during the denaturing and annealing reactions of the PCR. The S phaseolin genomic DNA yielded two homoduplex PCR products, one that contained no repeated sequences (PCR-S1) and one that contained a 21-bp tandem direct repeat within the third intron (PCR-S2) (Figs. 1A and 2). The T phaseolin genomic DNA produced three homoduplex amplification products. The smallest product (PCR-T1) was identical in sequence to the smallest amplification product of the S phaseolin (PCR-S1). The second largest product, PCR-T2, was identical to PCR-T1/PCR-S1 except that it contained the 15-bp repeat in the fourth exon and a G → T transversion at position 84 of the amplified sequence (Fig. 2). The third and largest of the T-type homoduplex bands, PCR-T3, contained both the 15-bp repeat in the fourth exon and the same 21-bp tandem direct repeat in the third intron as PCR-S2 (Fig. 2).

![Fig. 1. Results from phaseolin 15-bp PCR assay. (A) A 10% polyacrylamide gel. (B) A 1.6% agarose gel. Lanes: T, Tendergreen genomic DNA; S, Sanilac genomic DNA. Size standards, 123-bp ladder.](image-url)

![Fig. 2. Sequence comparison of PCR bands from phaseolin 15-bp PCR assay. T1, T2, and T3 represent homoduplex bands from the T-type phaseolin; S1 and S2 represent homoduplex bands from the S-type phaseolin; and I represents the band from the I-type phaseolin. Double underline indicates tandemly duplicated repeat elements; single underline indicates unduplicated repeat elements; asterisk indicates a DNA mismatch.](image-url)
The existence of the two sets of direct duplications may explain the appearance of the additional bands of larger apparent molecular size, particularly in PAGE. The unpaired, single-stranded loop in the heteroduplexes may be responsible for the altered mobility of these fragments in polyacrylamide gels (26), because it would induce a conformation change that is revealed in particular by the smaller pore size of the polyacrylamide matrix compared to that of the agarose matrix. In the case of the S phaseolin type, the two homoduplex bands could reanneal to produce two additional heteroduplex fragments. For the T phaseolin type, the three homoduplex bands could give rise to a total of six possible heteroduplexes; however, two of these would result in similar heteroduplex structures that would comigrate, thus explaining the five heteroduplex bands observed.

**P. vulgaris Germplasm Survey.** To determine the geographical distribution of PCR patterns and specifically of the 15- and 21-bp tandem direct repeats, 27 accessions were assayed that are representative of the *P. vulgaris* germplasm. These same accessions had been characterized previously for phaseolin protein, isozymes, and single copy restriction fragment length polymorphisms (6, 8, 13, 19, 21).

Several methods of extracting genomic DNA from seeds were attempted in order to avoid the delays associated with DNA extraction from leaves. The method that gave the most consistent results used as extraction buffer a simple modification of the Laemmli (27) protein sample buffer followed by a series of alcohol precipitations (see Materials and Methods).

Yields averaged \( \sim 0.1 \mu g \) of DNA per mg of seed flour and varied within 25% across samples from different genotypes.

PCR patterns were scored for the presence of the 15- and 21-bp repeats (Table 1). Several other patterns in addition to the T and S patterns were identified (Fig. 3). One of the most significant was that of a wild accession, G21245, from northern Peru, which showed a single band (lane 8). The SDS/PAGE pattern of this material has been designated as I-type phaseolin (8). Sequence analysis indicated that this amplified product was identical to that of the PCR-S1 and PCR-T1 bands and, consequently, that I-type phaseolin genes do not contain either the 15- or the 21-bp tandem direct repeats (Fig. 2).

The survey of cultivated *P. vulgaris* showed that none of the accessions displayed the single-band amplification pattern characteristic of the I-type phaseolin. This result confirmed previous results based on SDS/PAGE that the I phaseolin was not represented in the cultivated gene pool (8). Among wild beans, only wild accessions from Ecuador and northern Peru showed the single-band pattern diagnostic of the absence of the 15- and 21-bp tandem direct repeats and characteristic of the I phaseolin (Table 1). The 21-bp tandem direct repeat was present in all wild beans tested from the two other geographic regions, Middle America and the southern Andes. The 15-bp tandem direct repeat was generally absent from the Middle American wild beans but was present in all of the southern Andean wild beans. The distribution of the latter tandem direct repeat correlates, although not absolutely, with the distribution of the S and T phaseolin in the Middle American and southern Andean regions, respectively.

<table>
<thead>
<tr>
<th>Region</th>
<th>15-bp tandem repeat</th>
<th>21-bp tandem repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico, Central America, and Colombia</td>
<td>Present 3 Absent 6</td>
<td>Present 9 Absent 0</td>
</tr>
<tr>
<td>Ecuador, northern Peru</td>
<td>0 10</td>
<td>0 10</td>
</tr>
<tr>
<td>Southern Andes: southern Peru, Bolivia, Argentina</td>
<td>12 0</td>
<td>12 0</td>
</tr>
</tbody>
</table>

**DISCUSSION**

The PCR test of phaseolin diversity developed here has proved to be highly reproducible as the same banding pattern was observed in multiple experiments. The complete correlation between the PCR, one-dimensional SDS/PAGE, and Southern hybridization patterns in the F2 generation of a cross between an S and a T phaseolin cultivar further confirmed the reproducibility of the assay. Although a larger than expected number of bands were observed, our subsequent investigations allowed us to propose an explanation for the origin of the supernumerary bands. Bands of larger apparent molecular size in polyacrylamide gels turned out to be heteroduplex bands. Although artefactual, these bands were diagnostic of specific phaseolin types and helped recognize heterozygotes. Other cases of heteroduplex bands with altered mobility have recently been reported (e.g., see ref. 28) and have been used to study polymorphism of human immunodeficiency virus 1 env genes (29).

PCR-based analyses of specific regions of phaseolin genes provide a more powerful means than SDS/PAGE to analyze phaseolin gene diversity. Variation in both exons and introns can be identified. More detailed analyses using restriction digestion or sequencing of amplified fragments further increase the power of this approach. Analysis of amplified fragments by PAGE provides a more accurate estimate of molecular mass (within a few base pairs) than one-dimensional SDS/PAGE (several hundred daltons). Minor size variations due to insertions/deletions could easily be overlooked by the latter method. We have also been able to extend comparisons to other *Phaseolus* (*P. coccineus* and *P. lunatus*) or *Vigna* (*V. unguiculata*) species with the same primers used in *P. vulgaris* (J.K., V. Llica, and P.G., unpublished results). Because of its superior discriminating power, this PCR-based analysis could supersede SDS/PAGE-based analysis for screening of phaseolin diversity in *Phaseolus* germplasm.

One of the most significant features of this research is the discovery that the I phaseolin genes present in wild beans from Ecuador and northern Peru (8, 14) did not carry the 15- and 21-bp tandem direct repeats. Duplications that generated the tandem direct repeats would be more likely than deletions that specifically eliminated a member of a tandem direct repeat because the former could occur in many locations along the phaseolin sequence, whereas the latter would occur only at the sites of the tandem direct repeats. This argument suggests that the I phaseolin sequences are probably ancestral to the other phaseolin sequences found in *P. vulgaris*. It is interesting to note in this respect that accessions of *P. coccineus coccineus* and *P. coccineus darwinianus* (or *Phaseolus polyanthus*), the taxa most closely related to *P. vulgaris*, also show a pattern...
consisting of a single band pattern of similar size as the I phaseolin band (V. Llaca, J.K., and P.G., unpublished results). In addition, I phaseolin genes also appear to lack the 27-bp tandem direct repeat in their sixth exon, whereas S and T phaseolin genes contain this repeat (V. Llaca, J.K., R. Freyre, and P.G., unpublished results).

It is interesting to speculate that wild beans from the mountainous areas of Ecuador and northern Peru, an area that is geographically intermediate between that of the Middle American and Andean gene pools, may therefore constitute the ancestral wild P. vulgaris type. Arguments in favor of this hypothesis include isozyme data that show that wild P. vulgaris from this area are distinct from and intermediate between wild P. vulgaris of Middle America and the southern Andes (13, 14). The isozyme loci analyzed were unlinked among each other and to the phaseolin locus. They therefore characterized different regions of the nuclear genome.

Khairallah et al. (30) demonstrated that accession G21245, used in this study as representative of the I phaseolin, was more variable in its mitochondrial DNA than other wild bean populations in the Middle American and Andean gene pools. An alternative possibility is that the I phaseolin type originated in P. coccineus and was introduced by introgression into P. vulgaris. This is an unlikely possibility given that the environments of P. coccineus and wild P. vulgaris do not coincide in Ecuador and northern Peru and there is no morphological or isozyme evidence for such an introgression in these wild P. vulgaris populations (14, 30). Similar analyses of other regions of the phaseolin sequences or of other genes could provide additional information on the potential ancestral status of these populations.

Under the assumption that the I phaseolin constitutes the ancestral phaseolin type, the following events in the evolution of this multigene family in P. vulgaris can be proposed. In a first stage, prior to the divergence of the two geographic gene pools, 21- and 27-bp tandem duplications would have taken place in the third intron and the sixth exon, respectively, of members of the phaseolin gene family. Subsequently, the 15-bp repeat was probably introduced in the southern dissemination branch and, thus, came to predominate in the Andean gene pool of P. vulgaris. Analyses by pulsed-field gel electrophoresis of the phaseolin locus suggest that, in addition to the increase in size of the genes, there was also an increase in the number of genes in the multigene family (V. Llaca and P.G., unpublished results).

A striking observation is the progressive accumulation in the phaseolin sequences of repeats that maintain the reading frame, whereas no deletions or insertions that perturb the reading frame have been observed in these sequences during the same evolutionary time frame. This raises the possibility that selection for adaptation during dissemination into new environments and domestication may have led to the fixation of phaseolin types conferring some advantage during seed or seedling development. Whereas the nature of this selective effect is unknown so far, it has been observed that different phaseolin types are associated with variation for amount of phaseolin and seed size, although the absence of phaseolin does not reduce overall seed protein levels or, apparently, fitness (31–33). The precise nature of the relationship, if any exists, between phaseolin sequence variation and fitness of wild and cultivated beans remains to be determined.

We thank S. Singh and M. Iwanaga for providing seeds of accessions studied here, J. Slightom for providing phaseolin clones, and L. Gottlieb and C. Rick for comments on the manuscript. This research was supported in part by the International Board for Plant Genetic Resources, Rome, Italy, the Agency for International Development, Washington, DC, Program in Science and Technology Cooperation, and the United States Department of Agriculture, Washington, DC.

1. Osborne, T. B. (1907) _The Vegetable Proteins_ (Longmans and Green, London).