Mesoamerican origin of the common bean (Phaseolus vulgaris L.) is revealed by sequence data

Elena Bitocchi1, Laura Nanni1, Elisa Bellucci2, Monica Rossi3, Alessandro Giardini4, Pierluigi Spagnolletti Zeuli5, Giuseppina Logozzo6, Jens Stougard7, Philip McClean8, Giovanni Attene9, and Roberto Papa1,4,1

1Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, 60131 Ancona, Italy; 2Dipartimento di Biologia Difesa e Biotecnologie Agro-Forestali, Università degli Studi della Basilicata, 85100 Potenza, Italy; 3Centre for Carbohydrate Recognition and Signalling, Department of Molecular Biology, Aarhus University, DK-8000 Aarhus C, Denmark; 4Department of Plant Sciences, North Dakota State University, Fargo, ND 58105; 5Dipartimento di Scienze Agronomiche e Genetica Vegetale Agraria, Università degli Studi di Sassari, 07100 Sassari, Italy; and 6Cereal Research Centre, Agricultural Research Council (CRA-CER), S.S. 16, Km 675, 71122 Foggia, Italy

Knowledge about the origins and evolution of crop species represents an important prerequisite for efficient conservation and use of existing plant materials. This study was designed to solve the ongoing debate on the origins of the common bean by investigating the nucleotide diversity at five gene loci of a large sample that represents the entire geographical distribution of the wild forms of this species. Our data clearly indicate a Mesoamerican origin of the common bean. They also strongly support the occurrence of a bottleneck during the formation of the Andean gene pool that predates the domestication, which was suggested by recent studies based on multilocus molecular markers. Furthermore, a remarkable result was the genetic structure that was seen for the Mesoamerican accessions, with the identification of four different genetic groups that have different relationships with the sets of wild accessions from the Andes and northern Peru–Ecuador. This finding implies that both of the gene pools from South America originated through different migration events from the Mesoamerican populations that were characteristic of central Mexico.

crop evolution | Phaseoleae | population structure | mutation rate

The common bean (Phaseolus vulgaris L.) is the main grain legume for direct human consumption, and it represents a rich source of protein, vitamins, minerals, and fiber, especially for the poorer populations of Africa and Latin America (1). Investigations into the origins and evolution of this species would be expected to highlight the structure and organization of its genetic diversity and the role of the evolutionary forces that have been shaping this diversity. Such knowledge is a crucial prerequisite for efficient conservation and use of the existing germplasm for the development of new improved plant varieties.

The current distribution of the wild common bean encompasses a large geographical area: from northern Mexico to southwestern Argentina (2). In general, two major ecogeographical gene pools are recognized: Mesoamerica and the Andes. These two gene pools are characterized by partial reproductive isolation (3, 4), and they are seen in both wild and domesticated materials. They have been recognized in several studies based on morphology (5–7), agronomic traits (7), seed proteins (8), allozymes (9), and different types of molecular markers (10–15), which have given the overall indication of the occurrence of at least two independent domestication events in the two different hemispheres. The existence of these two geographically distinct and isolated evolutionary lineages that predate the domestication of the common bean represents a unique scenario among crops. This scenario differs from the occurrence over a smaller geographical range of multiple domestications that have been proposed for other species, such as barley (16) and wheat (17), and indeed, within the Mesoamerican gene pool of the common bean itself (18, but see 13, 19). In these cases, the lack of isolation between the populations prevented independent evolution of different lineages in both the wild and domesticated forms. Rice is the only crop that may show a scenario that is to some extent similar to the scenario of P. vulgaris (20, 21), although a recent study has suggested a single domestication event at the origin of the indica and japonica subgroups (22). While only these two major gene pools are recognized in the domesticated population, the geographical structure of the wild form of the common bean is more complex, with an additional third gene pool that is localized between Peru and Ecuador (23) and characterized by a specific storage seed protein, phaseolin type 1 (24). Moreover, wild populations from Colombia are often seen as intermediates, and a marked geographical structure is observed in wild beans from Mesoamerica (12). The population from northern Peru and Ecuador is usually considered the ancestral population from which P. vulgaris originated (the northern Peru–Ecuador hypothesis) (11, 24, 25). Indeed, the work by Kami et al. (24) analyzed a portion of the gene that codes for the storage seed protein phaseolin, including phaseolin type 1 from northern Peru–Ecuador accessions that was not present in the other gene pools, which indicates that type 1 phaseolin is ancestral to the other phaseolin sequences of P. vulgaris (24). Thus, the work by Kami et al. (24) suggested that, starting from the core area of the western slopes of the Andes in northern Peru and Ecuador, the wild bean was dispersed north (Colombia, Central America, and Mexico) and south (southern Peru, Bolivia, and Argentina), which resulted in the Mesoamerican and Andean gene pools, respectively.

The Mesoamerican origin of the common bean is an alternative and older hypothesis. This hypothesis is supported by the observations that the closest relatives of wild P. vulgaris in the Phaseolus genus are distributed throughout Mesoamerica (26–29). Additionally, the higher diversity found in the Mesoamerican compared with the Andean gene pool (phaseolin types, allozyme alleles, and molecular markers) (8–10, 30, 31) supports a Mesoamerican origin.

Recently, the work by Rossi et al. (13) compared amplified fragment length polymorphism (AFLP) data based on an analysis of wild and domesticated P. vulgaris accessions with the data from the work by Kwak and Gepts (14) based on simple sequence repeats (SSRs) obtained on a similar sample. This analysis suggested that, before domestication, there was a severe bottleneck in the Andean populations, and they reproposed the Mesoamerican origin of the common bean.


The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. JN796475–JN796922).

1To whom correspondence should be addressed. E-mail: r.papa@univpm.it.

See Author Summary on page 5148 (volume 109, number 14).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1108973109/-/DCSupplemental.
The main aim of the present study was to investigate the evolutionary history of *P. vulgaris* to solve the issue regarding the origin of this species. We have investigated the nucleotide diversity for five different genes from a wide sample of wild *P. vulgaris* that is representative of its geographical distribution. Our findings identify the origin of *P. vulgaris* as Mesoamerican, and they also reveal the level and structure of the genetic diversity that characterizes the wild accessions of this species. Our data are relevant for crop improvement and in particular, maximization of the efficient use of wild genetic diversity for breeding programs.

**Results**

**Nucleotide Variation in the Wild Common Bean.** We sequenced five loci of a large collection that included wild common bean accessions from Mesoamerican (*n* = 49) and Andean (*n* = 47) gene pools and genotypes from northern Peru–Ecuador (*n* = 6) that are characterized by the ancestral type I phaseolin (Table S1). These gene pools represent a cross-section of the entire geographical distribution of the wild form of *P. vulgaris*. The sequenced region for each gene is located within the transcriptional unit and encompasses between 500 and 900 bp, which includes both introns and exons. Altogether, we sequenced ~3.4 kb per accession.

We investigated the structure of the sequenced gene fragments in *P. vulgaris*, starting with BLAST analysis against the reference sequences for the identification of legume anchor (Leg) markers (Fig. S1 and Table S2) (32). For the Leg marker Leg044, we identified an exon of 102 bp within the fragment, and the deduced 34-aa sequence has 85% identity with histidinol dehydrogenase of *Arabidopsis thaliana*. This enzyme catalyses the last two steps in the 1-histidine biosynthesis pathway, which is conserved in bacteria, archaea, fungi, and plants. For Leg100, a coding region of 45 bp was identified at the 3′ extremity of the sequence, and its deduced 15-aa sequence showed 100% identity with biotin synthase of *A. thaliana*, an enzyme that is involved in the biotin biosynthetic process. Three exons were found in Leg133 of 75, 72, and 90 bp. The corresponding 79-aa sequence showed 84% identity with dolichyl-diphosphooligosaccharide-protein glycosyltransferase of *A. thaliana*, an enzyme that is involved in asparagine-linked protein glycosylation. Two exons (88 and 56 bp) were identified in Leg223, with an identity of 71% of the translated protein fragment with the eukaryotic translation initiation factor SU11 family protein of *A. thaliana*. This protein seems to have an important role in accurate initiator codon recognition during translation initiation. The structure of *PvSHP1* was characterized in the work by Nanni et al. (19) (Fig. S1 and Table S2), and the fragment sequenced in this study included three coding regions of 12, 42, and 42 bp. This sequence was identified as homologous to *Arabidopsis SHP1* (**SHATTERPROOF 1**), a gene that is involved in the control of fruit shattering.

We considered the variation in the coding regions of the loci studied in all of the *P. vulgaris* accessions. We found that the coding regions of three loci (Leg044, Leg100, and Leg223) did not show nucleotide substitutions. For Leg133, there was one synonymous substitution in three Mexican accessions and all of

### Table 1. Population genetics statistics for the five gene fragments and the concatenate sequences in the different gene pools of *P. vulgaris*

<table>
<thead>
<tr>
<th>Population</th>
<th>N</th>
<th>V</th>
<th>Pi</th>
<th>S</th>
<th>H</th>
<th>Hd</th>
<th>θw × 10⁻³</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Concatenate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>84</td>
<td>137</td>
<td>123</td>
<td>14</td>
<td>56</td>
<td>0.96</td>
<td>9.9</td>
<td>8.3</td>
</tr>
<tr>
<td>MW</td>
<td>37</td>
<td>119</td>
<td>98</td>
<td>21</td>
<td>34*</td>
<td>0.99*</td>
<td>10.6*</td>
<td>8.7*</td>
</tr>
<tr>
<td>AW</td>
<td>43</td>
<td>32</td>
<td>22</td>
<td>10</td>
<td>18</td>
<td>0.86</td>
<td>1.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Phi</td>
<td>4</td>
<td>18</td>
<td>0</td>
<td>18</td>
<td>4</td>
<td>1.00</td>
<td>2.7</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Leg044</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>96</td>
<td>23</td>
<td>23</td>
<td>0</td>
<td>13</td>
<td>0.83</td>
<td>6.1</td>
<td>5.4</td>
</tr>
<tr>
<td>MW</td>
<td>44</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>9*</td>
<td>0.82*</td>
<td>4.3*</td>
<td>5.6*</td>
</tr>
<tr>
<td>AW</td>
<td>46</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>0.47</td>
<td>0.9</td>
<td>2.5</td>
</tr>
<tr>
<td>Phi</td>
<td>6</td>
<td>14</td>
<td>0</td>
<td>14</td>
<td>2</td>
<td>0.33</td>
<td>5.7</td>
<td>7.4</td>
</tr>
<tr>
<td><strong>Leg100</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>99</td>
<td>45</td>
<td>43</td>
<td>2</td>
<td>13</td>
<td>0.70</td>
<td>22.6</td>
<td>15.4</td>
</tr>
<tr>
<td>MW</td>
<td>47</td>
<td>40</td>
<td>39</td>
<td>1</td>
<td>10*</td>
<td>0.83*</td>
<td>22.9*</td>
<td>16.0*</td>
</tr>
<tr>
<td>AW</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0.04</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Phi</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0.53</td>
<td>1.9</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>Leg133</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>101</td>
<td>15</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>0.63</td>
<td>5.1</td>
<td>5.0</td>
</tr>
<tr>
<td>MW</td>
<td>48</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>6*</td>
<td>0.75*</td>
<td>6.2*</td>
<td>5.1*</td>
</tr>
<tr>
<td>AW</td>
<td>47</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0.20</td>
<td>0.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Phi</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Leg223</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>94</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>8</td>
<td>0.78</td>
<td>3.2</td>
<td>2.9</td>
</tr>
<tr>
<td>MW</td>
<td>43</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>7*</td>
<td>0.79*</td>
<td>2.9*</td>
<td>3.0*</td>
</tr>
<tr>
<td>AW</td>
<td>47</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0.45</td>
<td>1.5</td>
<td>1.9</td>
</tr>
<tr>
<td>Phi</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>PvSHP1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>98</td>
<td>49</td>
<td>44</td>
<td>5</td>
<td>26</td>
<td>0.84</td>
<td>14.0</td>
<td>11.2</td>
</tr>
<tr>
<td>MW</td>
<td>47</td>
<td>42</td>
<td>36</td>
<td>6</td>
<td>20*</td>
<td>0.92*</td>
<td>16.0*</td>
<td>11.2*</td>
</tr>
<tr>
<td>AW</td>
<td>45</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>4</td>
<td>0.32</td>
<td>1.7</td>
<td>4.0</td>
</tr>
<tr>
<td>Phi</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0.60</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

All, all genotypes of *P. vulgaris*; AW, Andean wild; D, the D parameter by Tajima (66) for testing neutrality; H, number of haplotypes; Hd, haplotype diversity; MW, Mesoamerican wild; N, sample size; ns, not significant; Phi, phaseolin I type; Pi, parsimony informative sites; S, singleton variable sites; V, variable sites; θ = 10⁻³ and θw × 10⁻³, two measures of nucleotide diversity from Tajima (64) and Watterson (65) (θ estimator), respectively.

*The MW diversity parameters that are higher than those parameters of the AW.*
the accessions characterized by type I phaseolin. There was also one nonsynonymous substitution that involved only one Andean accession from Argentina [a 1-aa replacement: Asp (D) for Ala (A)]. Finally, for PvSHP1, a nonsynonymous substitution was found in two Mesoamerican accessions from Mexico [a 1-aa replacement: Gln (Q) for His (H)].

The nucleotide diversity for each of these five loci and the concatenate sequence was analyzed across all of the P. vulgaris accessions, taking into account the major population subdivisions that corresponded to the three wild gene pools: the Mesoamerican, Andean, and phaseolin type I (northern Peru–Ecuador) groups (Table 1). For the concatenate sequence, which included 84 accessions of P. vulgaris for which all of the five loci sequences were obtained, a total of 137 variable sites (V) were observed. The number of haplotypes was 56, none of which was shared among the three gene pools. The highest number of haplotypes was 34 for the Mesoamerican gene pool, with 18 haplotypes in the Andean gene pool and 4 haplotypes in the northern Peru–Ecuador accessions. Among all of the groups, the highest diversity was seen in the Mesoamerican wild population (\( \pi = 10.6 \times 10^{-3} \)), which was 10-fold higher than the diversity of the Andean accessions (\( \pi = 1.0 \times 10^{-3} \)). Considering each gene fragment separately, the diversity in the Mesoamerican was also much higher than in the Andes (for all of the statistics considered: number of haplotypes, haplotype diversity, \( \pi \), and \( \theta_w \) (Table 1). According to the binomial distribution, this pattern has a probability of occurring by chance of \( P = 0.03 \). Moreover, considering the estimates from the five loci and using the Wilcoxon–Kruskal–Wallis nonparametric test, the Mesoamerican wild bean showed a significant higher diversity compared with the Andean populations (from \( P \leq 0.009 \) for haplotype, haplotype diversity, and \( \pi \) to \( P \leq 0.016 \) for \( \theta_w \)). This finding was also shown by the loss of diversity (\( L\kappa \)) estimates between the Mesoamerican and Andean populations, with a reduction in the diversity for the latter compared with the former that ranged from 0.49 (Leg223) to 1.00 (Leg100) and was 0.90 for the concatenate sequence (Table 2). Finally, even if Tajima’s D was never significant, in the Andean population, it was negative and always much lower than Tajima’s D from Mesoamerica (Table 1).

**Population Structure.** The population structure analysis determined six subpopulations that best define the population (Fig. 1, B1 to B6). All of the Andean accessions were clearly assigned to cluster B6, with high percentages of membership (\( q_{B6} \geq 0.76 \)). The same was seen for the type I phaseolin of the northern Peru–Ecuador accessions, which were assigned to cluster B5 (\( q_{B5} \geq 0.91 \)). The Mesoamerican accessions showed a different scenario; indeed, they were subdivided into four different clusters (B1–B4), and they showed higher levels of admixture. We considered a threshold of \( q \geq 0.70 \) to assign the individuals to these four clusters. Cluster B1 included 17 Mesoamerican accessions (\( q_{B1} \geq 0.85 \)), 6 of which were from Mexico, 6 from Guatemala, 4 from Colombia, and 1 from El Salvador. The other three clusters were composed of only Mexican accessions: cluster B2 included seven accessions (\( q_{B2} \geq 0.93 \)), cluster B3 included eight accessions (\( q_{B3} \geq 0.76 \)), and cluster B4 included four accessions (\( q_{B4} \geq 0.99 \)).

One Mexican accession (PI325677) was not assigned to any specific Mesoamerican cluster because of its high level of admixture (\( q_{B6} = 0.28 \), \( q_{B3} = 0.45 \), \( q_{B6} = 0.22 \), and \( q_{B4} = 0.05 \)).

To better understand the geographical distributions of the Mesoamerican accessions that belong to genetic groups B1–B4, we carried out spatial interpolation of the membership coefficients (Fig. 2). The accessions of cluster B1 were distributed all along the Mesoamerican gene pool from the north of Mexico down to Colombia (Fig. 2A), with a major presence toward the Pacific Ocean. The accessions of cluster B2 were spread from central (particularly on the Caribbean side) down to southern Mexico (Fig. 2B). However, there was a lack of representation of these two clusters in a wide area of central Mexico. Clusters B3 and B4 were represented essentially in Mexico and in particular, above the Transverse Volcanic Axis, with cluster B3 widely spread from northern to central Mexico (Fig. 2C) and cluster B4 more restricted to a small area in central Mexico (Fig. 2D).

The relationships among the clusters identified were revealed by a neighbor-joining (NJ) tree (Fig. 3). There was no clear

**Table 2.** Loss of nucleotide diversity in the Andean wild (AW) population vs. Mesoamerican wild (MW) population calculated as \( L\kappa = 1 – (\pi_{MW}/\pi_{AW}) \), where \( \pi_{MW} \) and \( \pi_{AW} \) are the nucleotide diversities in MW and AW populations, respectively (39)

<table>
<thead>
<tr>
<th>Locus</th>
<th>( L\kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leg044</td>
<td>0.78</td>
</tr>
<tr>
<td>Leg100</td>
<td>1.00</td>
</tr>
<tr>
<td>Leg133</td>
<td>0.93</td>
</tr>
<tr>
<td>Leg223</td>
<td>0.49</td>
</tr>
<tr>
<td>PvSHP1</td>
<td>0.89</td>
</tr>
<tr>
<td>Concatenate</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Fig. 1. Percentages of membership (\( q \)) for each of the clusters identified (B1–B6; color-coded as indicated). Each accession is represented by a vertical line divided into colored segments, the lengths of which indicate the proportions of the genome that are attributed to the specific clusters. The accessions are ordered according to latitude from northern Mexico to northern Argentina. The country of origin is indicated by the horizontal line. AW, Andean wild; ar, Argentina; bl, Bolivia; C_mx, central Mexico; col, Colombia; ec, Ecuador; es, El Salvador; gt, Guatemala; MW, Mesoamerican wild; N_mx, north Mexico; N_pr, northern Peru; Phi, type I phaseolin (northern Peru–Ecuador); S_mx, south Mexico; S_pr, southern Peru.

Bitocchi et al.
distinction between the three gene pools (Mesoamerican, Andean, and northern Peru–Ecuador). Interestingly, the Andean cluster (B6) was more closely related to the Mesoamerican group B3, and the northern Peru–Ecuador cluster (B5) was more closely related to the Mesoamerican group B4; however, the Mesoamerican groups B1 and B2 were in an external position.

A similar population structure was revealed by the NJ tree obtained considering the single accessions (Fig. 4) according to the Bayesian model-based approach [Bayesian Analysis of Population Structure (BAPS)]. Also in this case, there was no clear distinction between the three gene pools. Indeed, the Mesoamerican accessions were found to be distributed in all of the tree branches: the Andean accessions (Fig. 4, AW) clustered with the Mesoamerican group B3 (bootstrap value = 97%) and the northern Peru–Ecuador accessions (Fig. 4, Ph1) were more related to the other Mesoamerican groups (bootstrap value = 98%) and particularly, the B4 accessions. The Mesoamerican groups B1 and B2 were included in a clade that was statistically well-supported (bootstrap value = 79%).

Haplotype Networks. The haplotype networks for each of the five loci are shown in Fig. 5. The number of haplotypes found for each gene ranged from 8 (Leg223) to 26 (PvSHP1). The Leg044, Leg133, and Leg223 haplotypes were interrelated through a few mutational steps, whereas a higher number of evolutionary steps was seen for the Leg100 and PvSHP1 haplotypes. However, a consistent observation arose from this analysis: the Mesoamerican gene pool had the greatest number of haplotypes for all of the loci, from 6 (Leg133) to 20 (PvSHP1), and the Andean gene pool showed a star-like structure with 1 major haplotype with a high frequency (higher than 0.72) and a few (from 1 to 3) additional minor haplotypes. Furthermore, there was no clear distinction between the Mesoamerican and Andean group of haplotypes, where the former seemed to be distributed all along the tree and the latter showed haplotypes shared always with the Mesoamerican accessions; the only exception here was PvSHP1, where there were no common haplotypes between these gene pools. However, a clear relationship was evident between a group of Mesoamerican haplotypes (from central Mexico) and the three

![Fig. 2. Spatial interpolation of membership coefficients (q) for the four Mesoamerican clusters identified by the Bayesian clustering analysis. (A) B1_blue. (B) B2_light-blue. (C) B3_green. (D) B4_orange. Latitude and longitude are expressed in the Universal Transverse Mercator system.](image)

![Fig. 3. Unrooted NJ tree showing the phylogenetic relationships of genetic clusters identified by cluster analysis.](image)
might not re
and as our results suggest, the current distribution of phaseolin
several reasons (33), this strict relationship has not always held,
phylogeny is identical to the phylogeny of the gene. However, for
seolin (type I), and it was based on the assumption that the species
phaseolin (the major seed storage protein) as the ancestral pha-

major Andean haplotypes, which were separated by six muta-
tional steps (Fig. 5, Pr-SHP1). A confirmation of the clear rela-
tionship between the Andean and Mesoamerican accessions of
group B3 that was highlighted by the NJ trees (Figs. 3 and 4) was
provided by these haplotype networks, even if, at this level,
a similar situation can be seen for B1. Finally, for all of the genes,
the type I phaseolin accessions showed haplotypes that were
closer to the Mesoamerican accessions and often separated from
the majority of the Andean accessions.

Discussion
In this study, we analyzed the nucleotide diversity for five gene
fragments in a large sample of wild P. vulgaris to address the
origins of this species. Indeed, the main aim was to throw light
onto the unique scenario among crop plants that characterizes
the common bean: the existence of two major geographically
distinct evolutionary lineages that predate domestication (Meso-
american and Andean). Our results indicate a clear pattern as-
associated with a Mesoamerican origin of this species from which
different migration events extended the distribution of P. vulgaris
into South America.

To date, the most credited hypothesis relating to the origins of
the common bean has indicated that, from a core area on the
western slopes of the Andes in northern Peru and Ecuador, the
wild beans were dispersed north (to Colombia, Central America,
and Mexico) and south (to southern Peru, Bolivia, and Argentina),
which resulted in the Mesoamerican and Andean gene pools, re-
respectively. This hypothesis has relied on the identification of a
phaseolin (the major seed storage protein) as the ancestral pha-
sein (type I), and it was based on the assumption that the species
phylogeny is identical to the phylogeny of the gene. However, for
several reasons (33), this strict relationship has not always held,
and as our results suggest, the current distribution of phaseolin
might not reflect its ancient distribution. Alternatively, type I
phaseolin might be extinct in Mesoamerica, or it might still be
present but just not included in the samples analyzed.

Predomestication Bottleneck in the Andes. Recently, on the basis
of a comparison of the levels observed for AFLP (13) and SSR (14)
diversity in the wild populations of P. vulgaris from the Andes
and Mesoamerica, the work by Rossi et al. (13) suggested that
a bottleneck had occurred in the Andes before domestication.
Indeed, for molecular markers characterized by a higher mutation
rate (SSRs compared with AFLPs), there were no (or very very small)
differences in the diversities observed in the two main geographical areas, whereas there was much lower di-
versity in the Andes for AFLPs compared with the Mesoamer-
ican wild P. vulgaris. Indirect estimates of the AFLP mutation
rate have shown values that vary from $10^{-10}$ to $10^{-5}$ (34–36),
whereas the SSR mutation rate is higher; it ranges from $10^{-3}$ to
$10^{-2}$ using both indirect (34, 37, 38) and direct (39, 40) esti-
mates. Thus, following the model proposed in the work by Nei
et al. (41) that described the effects of a bottleneck on the ge-
etic diversity of a population at a neutral locus, the work by
Rossi et al. (13) suggested the occurrence of a bottleneck in the
Andes before domestication. This bottleneck was then re-
covered by markers with a high mutation rate compared with
markers showing lower rates of mutation. According to this
hypothesis for nucleotide diversity, an even lower diversity
should be expected in the Andes compared with Mesoamerica.
Indeed, as indicated in the work by Lynch and Conery (42) for
Fabaceae, the nucleotide mutation rate is $\sim 6.1 \times 10^{-9}$ much
lower than both the AFLP and SSR rates. Our sequence data
show a very strong difference in the genetic diversity between
the wild Mesoamerican and Andean accessions ($Lx = 90\%$).
These reductions are about 2- and 13-fold higher than the
reductions computed in a comparable sample of P. vulgaris
genotypes using AFLP (45\%) (13) and SSR data (7\%) (14),
respectively. Thus, our data strongly support the bottleneck
hypothesis of Rossi et al. (13), and they are based on the clear
relationship between the mutation rate and the time of diversity
recovery from the occurrence of a bottleneck: the higher the
mutation rate, the faster the recovery of diversity. To the best of
our knowledge, this example is one of the few (40) of the critical

Fig. 4. Unrooted NJ bootstrap tree inferred from the concatenate sequence data. Each set of accessions (as indicated) is represented by a colored circle, and
each color indicates the membership to the BAPS groups. Small gray and violet circles represent the nodes for which bootstrap values are higher that 50% and
80%, respectively (the 80% threshold highlights the relationships with very strong support). AW, Andean wild; MW, Mesoamerican wild; Phl, type I phaseolin
(northern Peru–Ecuador).
role of marker mutations in describing the diversity of plant populations, thus underlining the need for the careful consideration of mutation rates in diversity studies.

**Population Structure in the Mesoamerican Gene Pool.** The occurrence of three gene pools of the wild common bean (Mesoamerica, Andes, and northern Peru–Ecuador) has been shown extensively (9, 12–14, 23, 24, 31, 43). In particular, the subdivisions of the two major ecogeographical gene pools (Mesoamerica and Andes) have been shown by several studies that have used different types of markers. Even if high population structure has been seen in the Mesoamerican wild gene pool (12), it has usually been considered as a single gene pool. However, in the present study, while the northern Peru–Ecuador and Andean gene pools are, indeed, characterized by homogeneous assignments into specific genetic groups (B5 and B6, respectively), the Mesoamerican accessions are clearly split into four distinct genetic clusters, B1–B4, that are clearly separated also with the NJ tree based on the single individual genotypes (Fig. 4). Moreover, a very important result is seen in the lack of a clear distinction between the Mesoamerican and Andean wild gene pools (Figs. 3–5), whereas the Mesoamerican clusters show different degrees of relatedness with the other gene pools. In particular, as showed by the NJ trees (Figs. 3 and 4), the Andean wild accessions (B6) were more related to the Mesoamerican B3 as were the northern Peru–Ecuador accessions (B5) to the Mesoamerican B4; indeed, the major Mesoamerican clusters (B1 and B2) were less related to the Andean and northern Peru–Ecuador clusters. It is, thus, important to consider the geographical distributions of these groups in Mesoamerica. The B1 group was present essentially across all of the geographical area from the north of Mexico down to Colombia, whereas the B2 group was spread from central to southern Mexico. According to our data, the B1 and B2 clusters were almost absent in a wide area of central Mexico, where they were substituted by the two Mesoamerican groups, as the B3 and B4 clusters, which were more related to the South American populations B3 and B6. This finding is clearly not compatible with the hypothesis of a South American origin, where the phaseolin I genotypes would be expected to be intermediate between the Mesoamerican and Andean clusters.

The Mesoamerican origin of the wild common bean is supported by the large diversity observed, and it is also confirmed by the single locus haplotype networks, where the northern Peru–Ecuador haplotypes were closer to the Mesoamerican groups and often separated from the majority of the Andean accessions that were usually represented as a single, highly frequent haplotype.

In summary, this analysis of the population structure supports the Mesoamerican origin hypothesis. At the same time, it reveals the very complex geographical structure of the genetic diversity in Mesoamerica, with central Mexico and the Transverse Volcanic Axis, which originated ~5 Ma in the Late Miocene, as the cradle of diversity of *P. vulgaris*. As the magnitude of this structure has not been clearly identified using multilocus markers, on the basis of its old origins, this finding would suggest that its signature has been partially hidden because of the combined effects of different mutation rates and recombination.

**Origin of the Common Bean.** Our study presents clear evidence of a Mesoamerican origin of *P. vulgaris*, which was most likely located in Mexico, both from the analysis of the population structure and phylogeny and the confirmation of the occurrence of a bottleneck before domestication in the Andes, which was proposed in the work by Rossi et al. (13). The Mesoamerican origin is consistent with the known distribution of most of the close relatives of *P. vulgaris*, the much higher diversity of Mesoamerican wild compared with the diversity from South America, the occurrence of a severe bottleneck in the Andes before domestication, and finally, the occurrence in Mesoamerica of wild beans that are closely related to those beans found in South America, both from the Andean and northern Peru–Ecuador
gene pool. Thus, we suggest that *P. vulgaris* from northern Peru–Ecuador is a relict population that only represents a fraction of the genetic diversity of the ancestral population and that this population migrated from Central Mexico in ancient times. The results that suggest type I phaseolin as ancestral seem quite robust; thus, the absence of this type of seed protein in the Mesoamerican gene pool could be explained by two alternative hypotheses: the type I phaseolin became extinct in Mesoamerica, or it might still be present but just not included in the samples analyzed in the literature.

Our data also present a scenario for the evolutionary history of the wild common bean, with the magnitude of the population subdivisions in Mesoamerica not having been clearly recognized before this study.

Conclusions
Evolutionary studies of crop species are crucial for several applications; indeed, the knowledge relating to the level and structure of genetic diversity of crop plants and their wild relatives is the starting point of any breeding program (44–52). Our study indicates that, to explore new genetic diversity that is not incorporated into the current domesticated germplasm and consider this high genetic diversity in the Mesoamerican accessions that is not present in the Andean gene pool, it is the wild Mesoamerican germplasm that should be used in breeding programs, because it has potential for the release of new cultivars. Moreover, it is crucial to consider the high population structure that characterizes the Mesoamerican wild germplasm to sample the largest amount of diversity for introgression into commercial varieties. This finding is very important if we consider that the majority of the improved varieties of the common bean are of Andean origin at present. Furthermore, exploration of new genetic diversity is also fundamental for the meeting of future demands for cultivars that can adapt to climate change, while also maintaining, or improving their yields.

Materials and Methods

Plant Materials
A panel of 102 wild accessions of the common bean was selected to represent the geographical distribution of wild *P. vulgaris* from northern Mexico to northwestern Argentina. The accessions are representative of the different gene pools of the species: 49 Mesoamerican accessions (Mexico, Central America, and Colombia), 47 Andean accessions (South America), and 6 wild accessions from northern Peru and Ecuador that are characterized by the ancestral type I phaseolin (23, 24). The accessions characterized by the type I phaseolin are from few small populations found in restricted geographic areas on the western slope of the Andes. We used six of these accessions to represent the diversity of these populations, including only well-described and characterized accessions. A complete list of the accessions studied is available in Table S1. The seeds were provided by the United States Department of Agriculture Western Regional Plant Introduction Station and the International Centre of Tropical Agriculture in Colombia.

PCR and Sequencing
Genomic DNA was extracted from each accession from young leaves of a single, greenhouse-grown plant using the miniprep extraction method (53). A total of five–500- to 900-bp gene regions across the common bean genome were sequenced (Table S2). Four of these fragment genes (Leg044, Leg100, Leg133, and Leg223) were chosen from a set of these Leg markers developed in the work by Hougaard et al. (32). Single-copy orthologous genes between legume species were identified, and primers were designed in conserved exon regions for the amplification of exons and intron sequences (32, 54–56). The fifth gene fragment, PvSHP1, is homologous to the SHATTERPROOF (SHP1) gene involved in the control of fruit shattering in *A. thaliana*. This finding was developed in the work by Nanni et al. (19) that analyzed PvSHP1 nucleotide variation in a limited sample of wild *P. vulgaris* (29 and 16 Mesoamerican and Andean wild genotypes, respectively). The available PvSHP1 sequence data of 40 wild genotypes from the study by Nanni et al. (19) were included in the present analyses (Table S1).

It is important to note that mapping experiments show that gene duplication and highly repeated gene sequences in *P. vulgaris* are generally low, with most loci occurring as single copies (57–59).

DNA fragments were amplified using 25 μg DNA and the following reagent concentrations: 0.25 μM each forward and reverse primer, 200 μM each dNTP, 2.5 mM MgCl2, 1× Taq polymerase buffer, 1 unit AmpliTaq DNA polymerase, and sterile double-distilled H2O to a final volume of 100 μL. Amplifications were carried out using a 9700 Thermal Cycler (Perkin-Elmer Applied Biosystems) with an initial denaturation of 1 min at 95 °C that was followed by 30 cycles of 1 min at 95 °C, 1 min at X °C, and 2 min at 72 °C plus 10 min of final extension at 72 °C. The X °C refers to the annealing temperature, which is specified for each primer pair in Table S2. The PCR products were purified using GFX PCR DNA and Gel Band Purification Kits (Invitrogen) according to the manufacturer’s instructions. For the Leg100, Leg133, and Leg223 loci, the samples were sequenced on both strands using forward and reverse primers on the cycle sequencing reaction with BigDye Terminator Cycle Sequencing Ready Reaction Kits (Applied Biosystems). The products were resolved on an ABI Prism 3100-Avant Automated Sequencer (Applied Biosystems). The sequence data were analyzed using Pregap4 and Gap4 of the Staden Software Package (http://staden.sourceforge.net). The Pregap4 modules were used to prepare the sequence data for assembly (quality analysis). Gap4 was used for the final sequence assembly of the Pregap4 output files (normal shotgun assembly). The single-strand sequencing reaction for Leg044 (reverse primer) and PvSHP1 (forward primer) was performed by Macrogen. The sequences were resequenced, and there was no ambiguity as to which allele was present. The sequences are accessible at GenBank (GenBank accession nos. JN796475–JN796922).

Diversity Analyses
For the Leg markers, the identification of exons and introns was carried out by BLAST analysis (60) against the reference sequence of *A. thaliana* (32) (Table S2). For PvSHP1, its structure was previously identified in the work by Nanni et al. (19). Sequence alignment and editing were carried out using MUSCLE v3.7 (61) and BioEdit v.7.0.9.0 (62). Insertion/deletion (indels) were not included in the analyses. Molecular population genetic analyses were conducted using DnaSP 5.10.01 (63). As given in Table 1, estimations for the five gene fragments and the concatenate sequences in the different gene pools of *P. vulgaris* were made for the number of variable sites, singleton variable sites, parsimony informative sites, haplotypes, nucleotide diversity, number of haplotypes, and nucleotide diversity is in reference to Tajima’s D (66). To measure the loss of nucleotide diversity in the Andean wild vs. Mesoamerican wild populations as proposed in the work by Vigo-roux et al. (39), we used the statistic *L~c~* = 1 – (π~w~)/π~M~, where π~w~ is the nucleotide diversity in Mesoamerican and Andean wild populations, respectively, and *L~c~* parameter ranges from zero to one, which indicate a total loss of diversity and no loss of diversity, respectively. Haplotype trees for each gene fragment were constructed using the median-joining network algorithm implemented in the program NETWORK 4.5.1.6 (67). Considering the concatenate sequence, an unrooted phylogenetic tree was constructed based on Kimura two-parameter distances, and the relative support for each node was tested by the bootstrap method using 1,000 replicates in MEGA 4 (68). The sites with gaps were also excluded from these analyses.

Population Structure Analyses
A Bayesian model-based approach was used to infer the hidden genetic population structure of our sample and thus, assign the genotypes into genetically structured groups/populations. This approach was implemented in the software BAPS 5.3 (69–72). This version of the software incorporates the possibility to account for the dependence caused by linkage between the sites within aligned sequences, which is different from the most widely used STRUCTURE software (73) that uses a model that is not designed to deal with background linkage disequilibrium between very tightly linked markers (74). A total of 84 accessions, those accessions showing high-quality sequences for all of the five loci, was used in this analysis. We carried out a genetic mixture analysis to determine the most probable number of populations (*K*) given the data (71, 72). Under its default settings, BAPS includes *K* as a parameter to be estimated, and the best partition of the data into *K* clusters is identified as the one with the highest marginal log likelihood. The clustering with linked loci analysis was chosen to account for the linkage present between sites within aligned sequences; at the same time, the five loci were assumed as independent. Ten iterations of *K* (from 1 to 20) were conducted to determine the optimal number of genetically homogeneous groups. The admixture analysis was then applied to estimate individual admixture proportions with regards to the most likely number of clusters identified (70, 72). Admixture inference was based on 100 realizations from the posterior of the allele frequencies. We repeated the admixture five times to confirm the consistency of the results. Spatial interpolation of mem-
bership coefficients was performed according to the kriging method, which was implemented in the R packages spatial (http://www.r-project.org). To explore the relationship between the identified clusters, an unrooted phylogenetic tree was constructed using the NJ algorithm in MEGA 4 (68).


ACKNOWLEDGMENTS. We thank G. Bertorelle for critical reading of this manuscript and his valuable advice. This study was supported by Government (MIUR) Grant n. 20083PFSSA_001, Project Progetti di Ricerca di Interesse Nazionale (PRIN) 2008, and the Università Politecnica delle Marche (2006–2010).

Bitocchi et al. PNAS | Published online March 5, 2012 | 1795