Biodiversity studies in *Phaseolus* species by DNA barcoding

Silvia Nicolè, David L. Erickson, Daria Ambrosi, Elisa Bellucci, Margherita Lucchin, Roberto Papa, W. John Kress, and Gianni Barcaccia

Abstract: The potential of DNA barcoding was tested as a system for studying genetic diversity and genetic traceability in bean germplasm. This technique was applied to several pure lines of *Phaseolus vulgaris* L. belonging to wild, domesticated, and cultivated common beans, along with some accessions of *Phaseolus coccineus* L., *Phaseolus lunatus* L., and *Vigna unguiculata* (L.) Walp. A multilocus approach was exploited using three chloroplast genic regions (*rbcL*, *trnL*, and *matK*), four intergenic spacers (*rpoB-trnC*, *atpBrbcL*, *trnT-trnL,* and *psbA-trnH*), and nuclear ITS1 and ITS2 rDNA sequences. Our main goals were to identify the markers and SNPs that show the best discriminant power at the variety level in common bean germplasm, to examine two methods (tree based versus character based) for biodiversity analysis and traceability assays, and to evaluate the overall utility of chloroplast DNA barcodes for reconstructing the origins of modern Italian varieties. Our results indicate that the neighbor-joining method is a powerful approach for comparing genetic diversity within plant species, but it is relatively uninformative for the genetic traceability of plant varieties. In contrast, the character-based method was able to identify several distinct haplotypes over all target regions corresponding to Mesoamerican or Andean accessions; Italian accessions originated from both gene pools. On the whole, our findings raise some concerns about the use of DNA barcoding for intraspecific genetic diversity studies in common beans and highlights its limitations for resolving genetic relationships between landraces and varieties.

Key words: Phaseolus spp., plastid DNA, internal transcribed spacers, DNA barcoding, varietal groups, single-nucleotide polymorphisms.

Résumé : Les auteurs ont exploré le potentiel des codes barres génétiques pour étudier la diversité et la traçabilité génétiques au sein du germoplasme du haricot. Cette technique a été employée sur plusieurs lignées pures sauvages, domestiquées et cultivées du *Phaseolus vulgaris*, ainsi qu'à quelques accessions du *P. coccineus*, du *P. lunatus* et du *Vigna unguiculata*. Une approche multilocus a été exploitée au moyen de trois régions géniques chloroplastiques (*rbcL*, *trnL* et *matK*), de quatre espaceurs intergéniques (*rpoB-trnC*, *atpB-rbcL*, *trnT-trnL* et *psbA-trnH*), et les séquences nucléaires ITS1 et ITS2 de l'ADNr. Les buts principaux étaient d'identifier les marqueurs et SNP qui offraient le pouvoir discriminant le plus grand entre les variétés chez le haricot, de comparer deux méthodes (fondée sur les arbres ou les caractères) pour l'analyse de la biodiversité et pour des essais de traçabilité, et d'évaluer l'utilité globale des codes barres d'ADN chloroplastique pour retracer l'origine des variétés italiennes modernes. Les résultats obtenus montrent que la méthode NJ constitue une approche puissante pour comparer la diversité génétique au sein des espèces, mais qu'elle s'avère relativement peu informative pour ce qui est de la traçabilité génétique des cultivars. Au contraire, la méthode basée sur l'examen des caractères a permis d'identifier plusieurs haplotypes distincts pour toutes les régions étudiées au sein des accessions mésoaméricaines ou andéennes, ces deux pools génétiques étant la source des accessions italiennes. Globalement, ces observations soulèvent des interrogations sur l'emploi des codes barres génétiques pour des études de diversité génétique intraspécifique chez le haricot et soulignent les limites de cet outil pour la résolution des relations génétiques entre variétés de pays et cultivars.

Mots-clés : Phaseolus spp., ADN plastidique, espaceurs internes transcrits, codes barres génétiques, groupes variétaux, polymorphisme mononucléotidique.

[Traduit par la Rédaction]

Received 28 December 2010. Accepted 19 February 2011. Published at www.nrcresearchpress.com/gen on 21 July 2011.

Paper handled by Associate Editor Paolo Donini.

S. Nicolè, D. Ambrosi, M. Lucchin, and G. Barcaccia. Department of Environmental Agronomy and Crop Science, Università degli Studi di Padova, Via dell'Università 16 – Campus of Agripolis, 35020 Legnaro, Padova, Italy.
D.L. Erickson and W.J. Kress. Department of Botany and Laboratory of Analytical Biology, National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington, DC 20013-7012, USA.
E. Bellucci. Department of Environmental Sciences and Crop Production, Università Politecnica delle Marche, Ancona, Via Brecce Bianche, 60131 Ancona, Italy.
R. Papa. Department of Environmental Sciences and Crop Production, Università Politecnica delle Marche, Ancona, Via Brecce Bianche, 60131 Ancona, Italy; Cereal Research Centre, Agricultural Research Council, S.S. 16, Km 675, 71122 Foggia, Italy.

Corresponding author: Gianni Barcaccia (e-mail: gianni.barcaccia@unipd.it).

Introduction

The genomic advances of the last decade have provided the technological tools for the development of a universal, DNA-enhanced system of taxonomy suitable for addressing the current "biodiversity crisis" that requires innovative and informative technologies (Tautz et al. 2003). DNA barcoding has been proposed as a cost-effective technology (Hebert et al. 2003) able to contribute to the study of biodiversity, which, until recently, relied primarily on morphology in the Linnaean classification system. DNA-based methods are fast and not limited by taxonomic impediments such as missing morphological features of a particular life stage (e.g., eggs and juvenile forms) (Velzen et al. 2007), missing body parts (Wong and Hanner 2008), or homoplasy of some characters (Vences et al. 2005). Although the application of DNA fingerprinting as an identification tool is not a new idea, DNA barcoding has earned remarkable success attributable to the standardization of the procedure by the use of a universal barcode sequence across a wide range of organisms (Hebert et al. 2004). The proposal of using DNA barcoding as a new identification tool turned on a heated debate between the advocates and the opponents to the potential uses of this technique because of some theoretical and methodological weakness (Will and Rubinoff 2004; Will et al. 2005; Hickerson et al. 2006). The ambitious idea of using the polymorphism information in a short sequence of DNA to distinguish every species in the world has already been translated into a powerful tool in the animal kingdom (Ward et al. 2005), even if other studies demonstrated that some taxa are problematic for the application of DNA barcoding (Brower 2006; Meier et al. 2006; Wiemers and Fiedler 2007). Regarding the utility of the approach for land plants, biologists have been slower in adapting a universal gene region as a barcode because of the difficulty of finding a region analogous to the animal COI gene (also known as *cox1*). Recently, the CBOL Plant Working Group (2009) recommended the combination of the chloroplast genic regions *rbcL* and *matK* as the plant barcode. This core, two-locus DNA-barcoding approach has been proposed as a universal framework for the routine use of DNA sequence data to identify specimens and contribute to the discovery of unknown species of land plants. In the same publication, a minority position of the CBOL Plant Working Group supported the inclusion of the *trnH-psbA* intergenic spacer in the plant barcode following earlier publications that outlined practical difficulties related to the acquisition of *matK* sequences (Kress and Erickson 2007; Fazekas et al. 2008). The combination of the *rbcL* gene with the *trnH-psbA* intergenic spacer, a more rapidly evolving region than *rbcL* and *matK*, seems to be a valid alternative to a simple two-locus model: the former distinguishes distantly related plants, and the latter recognizes closely related sister species or species groups that have only recently diverged (Kress and Erickson 2007). Finally, even if organellar DNA sequences are used as the main source of information for a barcoding system, then one or more nuclear genes may also be required for the supplemental analysis of hybrids. Nuclear genes such as internal transcribed spacers (ITS), which are frequently used for phylogenetic analyses and single-copy nuclear regions, have been considered by some research groups (as, for instance, Cowan et al. 2006), even if with some reserves (see also http://www.kew.org/barcoding/).

Several DNA fingerprinting and genotyping assays based on molecular markers such as RFLPs and SNPs have been developed in the past and are still used in plant genetics and breeding (Mohler and Schwarz 2008). DNA barcoding could provide an additional system to identify not only species but also crop varieties and germplasm resources to assess the distinctiveness of genotypes and relatedness among genotypes (Pallottini et al. 2004). Assessment of the potential of DNA barcoding to distinguish between plant varieties of agri-food interest would be valuable for both breeders and farmers. Whereas the utility of DNA barcoding in species identification has been widely investigated, the intraspecific discrimination of single varietal genotypes, such as clones, pure lines, and hybrids, has been poorly investigated, and few studies have focused on the use of DNA barcoding as a sufficiently informative technique to be exploited for the genetic identification of closely related crop varieties (Tsai et al. 2008).

Our work focuses on the application of DNA barcoding to cultivated bean germplasm as a new tool for discrimination among *Phaseolus* spp. and, most of all, for identification of *Phaseolus vulgaris* L. varieties. *Phaseolus* is a genus in the family Fabaceae, the third largest family of flowering plants (Gepts et al. 2005), and it represents multiple domestications of distinct, but related, species and multiple populations within the same species, e.g., as found in *P. vulgaris* and *Phaseolus lunatus* L. The original natural distribution of this species, before its introduction throughout Europe and Africa in the post-Columbian period, consists of a fragmented area throughout Central and South America. On the basis of the available data, at least two primary centers of origin have been recognized: a relatively heterogeneous one in the Andes (Colombia, Ecuador, Peru, Bolivia, Chile, and Argentina) and a more homogeneous one in Mesoamerica (primarily Mexico, Guatemala, Honduras, El Salvador, Nicaragua, and Costa Rica). These two centers of origin are called the Andean and Mesoamerican gene pools, respectively (Chacón et al. 2005; Papa et al. 2006).

In this paper, we present results on the use of DNA barcoding in several pure lines of wild, domesticated, and cultivated common beans, using both coding and noncoding regions from the chloroplast and nuclear genomes. Our objectives were the following: (*i*) analysis of the performance of different markers as DNA barcodes, primarily below the species level (i.e., Andean and Mesoamerican gene pools); and (*ii*) evaluation of the effectiveness of different methods (i.e., tree based versus character based) of DNA barcoding.

Materials and methods

Germplasm sampling of *Phaseolus*

In total, 33 varieties of *P. vulgaris* were selected as representative of the Mesoamerican and Andean gene pools, based on morphological seed traits, plant descriptors, and molecular markers (Rossi et al. 2009). Eight wild and nine domesticated accessions from Central America (Mexico, Costa Rica, Honduras, and El Salvador) and ten wild and six domesticated accessions from South America (Argentina, Bolivia, Brazil, Colombia and Peru) were used, including two wild accessions from northern Peru and Ecuador characterized by the ancestral phaseolin type I (Debouck et al. 1993; Kami et al.

1995). These accessions were obtained from the germplasm banks held at the International Center for Tropical Agriculture (CIAT) and the United States Department of Agriculture (USDA) (Table 1). In addition, 22 Italian, cultivated, commercially available accessions from unknown progenitor gene pools were obtained from the Agricultural Research Council (CRA), Research Unit for Horticulture of Montanaso Lombardo (Fig. 1). Several *Phaseolus coccineus* L., *P. lunatus*, and *Vigna unguiculata* (L.) Walp accessions were used as reference standards and outgroups. A list of varieties and landraces with information on their origins can be found in Table 1.

Genomic DNA extraction

Genomic DNA was isolated from 0.5–1.0 g of powdered, frozen, young leaf tissue using the Nucleon PhytoPure DNA extraction kit (Amersham Biosciences, Little Chalfont, Buckinghamshire, UK), following the manufacturer's instructions. A purification step with NaOAc was performed to remove excess salts, and the DNA pellets were resuspended in 80–100 µL of 1× TE buffer (100 mmol/L Tris–HCl, 0.1 mmol/L EDTA, pH 8). DNA concentration was estimated by electrophoresis on an 0.8% agarose/TAE gel using the 1 kb Plus DNA ladder (Invitrogen, Carlsbad, California) as a size standard.

DNA barcode markers and PCR assays

To employ a multilocus barcoding technique (Kress and Erickson 2007; Newmaster et al. 2006), a subset of bean samples was tested at several genomic regions to determine the markers that provided the highest polymorphism information content at the intraspecific level. Only 7 of 12 chloroplast gene regions, including both coding (*rbcL* and *matK*) and noncoding regions (the *atpB-rbcL*, *trnH-psbA*, *trnT-trnL*, and *rpoB-trnC* intergenic spacers and the *trnL* intron), proved variable and informative, whereas the other regions (*rpl32-trnL*, *ndhF-rpl32*, *trnD-trnT*, *trnS-trnG*, and *rpoC1*) were found to be monomorphic and were not adopted for further analysis (data not shown). ITS1 and ITS2, the two ITS that separate the 5.8S ribosomal gene from the 18S and 25S loci in rDNA, were used to compare the utility of the nuclear and chloroplast genomes for resolving relationships at the variety level. For three of the selected chloroplast DNA (cpDNA) barcode regions, *rbcL*, *trnL*, and *atpB-rbcL*, primers were designed based on the sequences in the National Center for Biotechnology Information (NCBI) databases for the Fabaceae (legume) family. After removal of redundant and unverified entries, serial local multiple sequence alignments were performed by the Vector NT software. We used the PRIMER3 software to design specific primer pairs, ranging from 18 to 28 base pairs (bp) and located in highly conserved short stretches (300–500 bp) flanking the most variable portions of each region. In the other cases, universal primers were adopted (Table 2).

All PCR experiments were performed in duplicate using the GeneAmp PCR System 9700 (Applied Biosystems, Foster City, California) with an initial denaturation step of 5 min at 95 °C; followed by 35 cycles of 30 s at 95 °C, 1.10 min at 54 °C or 56 °C, and 1.20 min at 72 °C; followed by 7 min at 72 °C; and then held at 4 °C. PCR conditions were modified for the *matK* marker: an initial denaturation step of 5 min at 95 °C; followed by 40 cycles of 30 s at 95 °C, 1 min at 56 °C, and 2 min at 72 °C; followed by 7 min at 72 °C. The 25 µL PCR volume included 1× PCR buffer (100 mmol/L Tris–HCl pH 9.0, 15 mmol/L $MgCl_2$, and 500 mmol/L KCl), 0.2 mmol/L dNTPs, 0.2 µmol/L of each primer, 0.5 U of *Taq* DNA polymerase, 15 ng of genomic DNA as template, and 1× Hi Specific Additive (Bioline, London, UK) to facilitate amplification. The PCR products were resolved on 2% agarose/TAE gels and visualized under UV light using ethidium bromide staining. When faint double bands indicating the presence of nonspecific products were visualized on a gel, a second PCR was performed using more stringent conditions (higher annealing temperatures and fewer cycle numbers). Positive and negative controls were used as references. All amplification products were purified enzymatically by digestion with exonuclease I and shrimp alkaline phosphatase (Amersham Biosciences) and then sequenced using forward and reverse primers according to the original Rhodamine terminator cycle sequencing kit (ABI PRISM; Applied Biosystems). For some regions, an additional forward or reverse primer located outside the amplified region was adopted for sequencing replicates. For sequencing *matK*, dimethyl sulfoxide at 4% of the reaction volume was used to overcome some secondary structural problems.

Tree-based analysis

DNA sequences were visualized and manually edited using Sequencer 4.8 software to minimize sequencing errors and remove gaps in the coding regions that could cause shifts in the open reading frames of *rbcL*.

The BLASTn algorithm (http://www.ncbi.nlm.nih.gov/BLAST) was used to perform sequence similarity searches against the nonredundant nucleotide databases of NCBI. Then, the correspondence between the sequences of the PCR amplicons and the known sequences was tested. We carried out separate data analyses for each individual sequence and for the combined chloroplast and nuclear data sets, individually and together. Multiple sequence alignments were performed by the Se–Al v2.0a11 software, and the inter- and intraspecific genetic divergences were calculated by the MEGA 4.1 beta software (Tamura et al. 2007) according to the Kimura 2-parameter distance model (Kimura 1980). Based on the pairwise nucleotide sequence divergences, the neighbor-joining (NJ) tree was estimated and rooted using the accessions from different species as outgroups. A bootstrap analysis was conducted to measure the stability of the computed branches with 1000 resampling replicates. All nucleotide positions containing gaps and missing bases were eliminated from the data set (the complete deletion option). To assign each accession to the correct gene pool, we used a phenetic approach based on the computation of genetic distance to detect the "barcode gap", a discontinuity between intra- and interspecific variation (Hebert et al. 2003; Barrett and Hebert 2005), and the derived "10× rule" in *Phaseolus* spp. polymorphism analysis was performed on the complete sequence, a combination of the cpDNA regions, and the nuclear ITS regions.

Character-based analysis

The character-based technique was employed to look for unique sets of diagnostic characters related to single varieties

Table 1. List of 63 bean entries with the common name, accession number, origin area, and voucher information.

Sample	Species	Accessions	Classification	Origin	Gene pool	Voucher No.
PvF8wanc	*Phaseolus vulgaris*	G23585	Wild-ancestral	South America (Peru)	Ancestral	i.p.
PvG8wanc	*Phaseolus vulgaris*	G23587	Wild-ancestral	South America (Peru)	Ancestral	i.p.
PvH2mw	*Phaseolus vulgaris*	G23652	Wild	Central America (Mexico)	Mesoamerican	i.p.
PvA3mw	*Phaseolus vulgaris*	G12979	Wild	Central America (Mexico)	Mesoamerican	i.p.
PvC3mw	*Phaseolus vulgaris*	G23463	Wild	Central America (Mexico)	Mesoamerican	i.p.
PvD3mw	*Phaseolus vulgaris*	G22837	Wild	Central America (Mexico)	Mesoamerican	i.p.
PvB7mw	*Phaseolus vulgaris*	G12873	Wild	Central America (Mexico)	Mesoamerican	3901-8
PvG7mw	*Phaseolus vulgaris*	G12922	Wild	Central America (Mexico)	Mesoamerican	i.p.
PvB8mw	*Phaseolus vulgaris*	G11050	Wild	Central America (Mexico)	Mesoamerican	i.p.
PvC8mw	*Phaseolus vulgaris*	G12949	Wild	Central America (Mexico)	n.d.	i.p.
PvD8aw	*Phaseolus vulgaris*	G21113	Wild	South America (Colombia)	Mesoamerican	i.p.
PvE6aw	*Phaseolus vulgaris*	G23445	Wild	South America (Bolivia)	Andean	i.p.
PvF6aw	*Phaseolus vulgaris*	G23444	Wild	South America (Bolivia)	Andean	i.p.
PvG6aw	*Phaseolus vulgaris*	W618821	Wild	South America (Bolivia)	Andean	i.p.
PvH6aw	*Phaseolus vulgaris*	G23455	Wild	South America (Peru)	Andean	i.p.
PvG3aw	*Phaseolus vulgaris*	G23420	Wild	South America (Peru)	Andean	i.p.
PvB6aw	*Phaseolus vulgaris*	G19893	Wild	South America (Argentina)	Andean	i.p.
PvC6aw	*Phaseolus vulgaris*	G19898	Wild	South America (Argentina)	Andean	i.p.
PvD6aw	*Phaseolus vulgaris*	G21198	Wild	South America (Argentina)	Andean	i.p.
PvH5aw	*Phaseolus vulgaris*	W617499	Wild	South America (Argentina)	n.d.	i.p.
PvF7md	*Phaseolus vulgaris*	PI201349	Domesticated	Central America (Mexico)	Mesoamerican	i.p.
PvG1md	*Phaseolus vulgaris*	PI165435	Domesticated	Central America (Mexico)	Mesoamerican	3901-10
PvH1md	*Phaseolus vulgaris*	PI165440	Domesticated	Central America (Mexico)	Mesoamerican	i.p.
PvA2md	*Phaseolus vulgaris*	PI309785	Domesticated	Central America (Mexico)	Mesoamerican	i.p.
PvH4md	*Phaseolus vulgaris*	PI207370	Domesticated	Central America (Mexico)	Andean	i.p.
PvE7md	*Phaseolus vulgaris*	PI309885	Domesticated	Central America (Costa Rica)	Mesoamerican	i.p.
PvD1md	*Phaseolus vulgaris*	PI309831	Domesticated	Central America (Costa Rica)	Mesoamerican	i.p.
PvF1md	*Phaseolus vulgaris*	PI310577	Domesticated	Central America (Honduras)	Mesoamerican	i.p.
PvE1md	*Phaseolus vulgaris*	PI304110	Domesticated	Central America (El Salvador)	n.d.	i.p.
PvC1ad	*Phaseolus vulgaris*	BAT93–1	Domesticated	South America (Colombia)	Mesoamerican	i.p.
PvC2ad	*Phaseolus vulgaris*	BAT93–2	Domesticated	South America (Colombia)	Mesoamerican	i.p.
PvH8ad	*Phaseolus vulgaris*	BAT881	Domesticated	South America (Colombia)	n.d.	3901-11
PvB4ad	*Phaseolus vulgaris*	MIDAS	Domesticated	South America (Argentina)	Andean	i.p.
PvD5ad	*Phaseolus vulgaris*	PI290992	Domesticated	South America (Peru)	Andean	3901-9
PvA7ad	*Phaseolus vulgaris*	JALOEEP558	Domesticated	South America (Brasile)	Andean	3901-7
Pv1itc	*Phaseolus vulgaris*	Cannellino rosso	Cultivated	Italy	—	3901-16
Pv3itc	*Phaseolus vulgaris*	Montalbano	Cultivated	Italy	—	3901-18
Pv6itc	*Phaseolus vulgaris*	Munachedda nera	Cultivated	Italy	—	3901-19
Pv9itc	*Phaseolus vulgaris*	San Michele	Cultivated	Italy	—	i.p.
Pv10itc	*Phaseolus vulgaris*	Nasieddu viola	Cultivated	Italy	—	i.p.
Pv13itc	*Phaseolus vulgaris*	Maruchedda	Cultivated	Italy	—	i.p.
Pv14itc	*Phaseolus vulgaris*	Riso bianco	Cultivated	Italy	—	3901-20
Pv16itc	*Phaseolus vulgaris*	Cannellino	Cultivated	Italy	—	3901-21
Pv19itc	*Phaseolus vulgaris*	Verdolino	Cultivated	Italy	—	3901-22
Pv22itc	*Phaseolus vulgaris*	Blu Lake	Cultivated	Italy	—	3901-23
Pv23itc	*Phaseolus vulgaris*	Goldrush	Cultivated	Italy	—	3901-24
Pv24itc	*Phaseolus vulgaris*	Borlotto Clio	Cultivated	Italy	—	i.p.
Pv27itc	*Phaseolus vulgaris*	Lena	Cultivated	Italy	—	3901-25

Table 1 (*concluded*).

Sample	Species	Accessions	Classification	Origin	Gene pool	Voucher No.
Pv28itc	*Phaseolus vulgaris*	Giulia	Cultivated	Italy	—	3901-26
Pv29itc	*Phaseolus vulgaris*	Saluggia	Cultivated	Italy	—	3901-27
Pv31itc	*Phaseolus vulgaris*	Borlotto Lamon	Cultivated	Italy	—	3901-28
Pv32itc	*Phaseolus vulgaris*	Saluggia	Cultivated	Italy	—	3901-29
Pv33itc	*Phaseolus vulgaris*	Cannellini	Cultivated	Italy	—	3901-30
Pv34itc	*Phaseolus vulgaris*	Verdoni	Cultivated	Italy	—	3901-34
Pv35itc	*Phaseolus vulgaris*	S. Matteo	Cultivated	Italy	—	3901-31
Pv36itc	*Phaseolus vulgaris*	Zolferini Rovigotti	Cultivated	Italy	—	3901-32
Pv37itc	*Phaseolus vulgaris*	Neri Messicani	Cultivated	Italy	—	3901-33
PcA1mw	*Phaseolus coccineus*	PI417608	Wild	Central America (Mexico)	n.d.	i.p.
Pc30itc	*Phaseolus coccineus*	Venere	Cultivated	Italy	—	i.p.
Pc39itc	*Phaseolus coccineus*	Spagna	Cultivated	Italy	—	i.p.
PlB1md	*Phaseolus lunatus*	PI310620	Domesticated	Central America (Guatemala)	n.d.	i.p.
Pl38itc	*Phaseolus lunatus*	Lima	Cultivated	Italy	—	3901-2
Vu40itc	*Vigna unguiculata*	Fagiolino dall'occhio	Cultivated	Italy	—	3905-2

Note: Voucher No., plants with flowers and pods are conserved in the herbarium of the Botanical Garden of the University of Padua (Italy); i.p., voucher attainment in progress; n.d., not determined.

or variety groups of *P. vulgaris*. Rather than using hierarchies or distance trees, character-based analysis classifies taxonomic groups based on shared specific informative character states, SNPs or insertions or deletions (indels), at either one or multiple nucleotide positions (DeSalle et al. 2005). Analysis of polymorphism distribution was performed using the DnaSP v.4 software (Rozas et al. 2003) to generate a map containing haplotype data without considering sites with alignment gaps. This program detects positions characterized by the presence of specific character states that are limited to a particular subgroup within *P. vulgaris* species and shared by all the members of that cluster. In addition, the haplotype number, H_n, and the haplotype diversity, H_d (Nei 1987), were estimated.

Population structure analysis

The population structure of the *P. vulgaris* germplasm was investigated using the Bayesian model-based clustering algorithm implemented in the STRUCTURE software (Pritchard et al. 2000; Falush et al. 2003), which identifies subgroups according to combination and distribution of molecular markers. This software was also used to assign each DNA sample of varieties and landraces, predefined according to geographical origin and (or) gene pool, to an inferred cluster. All simulations were executed assuming the admixture model, with no a priori population information. Analyses of SNP data were performed with 500 000 iterations and 500 000 burn-ins by assuming the allele frequencies among populations to be correlated (Falush et al. 2003). Ten replicate runs were performed, with each run exploring a range of K spanning from 1 to 16. The most likely value of K was estimated using ΔK, as reported in other studies (Evanno et al. 2005). Individuals with membership coefficients of $q_i \geq 0.7$ were assigned to a specific group, whereas individuals with $q_i < 0.7$ were identified as admixed.

Results

DNA barcoding success and levels of variability

For the selected chloroplast and nuclear markers examined in all 63 accessions of *Phaseolus* spp., our PCR amplifications were successful 100% of the time, although low quality sequences were sometimes produced because of specific gene regions (Table 3). For all dubious amplicons and sequences, the reactions were repeated. The only particularly problematic barcode marker was *matK*, with multiple failed amplifications and low sequence quality. Similar difficulties have been reported by others (Kress and Erickson 2007; Fazekas et al. 2008). Therefore, we removed this region from our analyses.

The primer pairs designed for *trnT-trnL* and *trnH-psbA* proved highly universal with a 100% success rate for both PCR and sequencing, whereas primers for the other markers (i.e., *rbcL*, *atpB-rbcL*, *trnL*, and *rpoB-trnC*) were also highly universal but unreliable in sequence quality. Although double PCR products were usually not detectable in the gel, sequencing problems likely arose from multiple comigrating amplicons of similar size but different sequence. When nonspecific amplicons of unexpected length were visible in the gel (i.e., for *rbcL* and *atpB-rbcL*), a second, more stringent PCR was performed, or new primer pairs were adopted for

Fig. 1. Seeds of the common bean (*Phaeolus vulgaris* L.) varieties analyzed in this study as representatives of the Italian cultivated germplasm (1, Cannellino rosso; 2, Riso giallo; 3, Montalbano; 4, Munachedda nera; 5, San Michele; 6, Nasieddu Viola; 7, Maruchedda; 8, Riso bianco; 9, Cannellino nano; 10, Verdolino; 11, Blu lake; 12, Goldrush; 13, Clio; 14, Zolferino rovigotto; 15, Lena; 16, Giulia; 17, Saluggia nano; 18, Venere; 19, Borlotto Lamon; 20, Saluggia; 21, Cannellino; 22, Verdone; 23, San Matteo; 24, Nero messicano; 25, BAT881 (reference breeding line)). Also analyzed in this study seeds of *Phaseolus lunatus* L. (26, sieva bean from Lima), *Phaseolus coccineus* L. (27, scarlet runner bean or Spanish bean), and *Vigna unguiculata* L. Walp. (28, blackeyed pea).

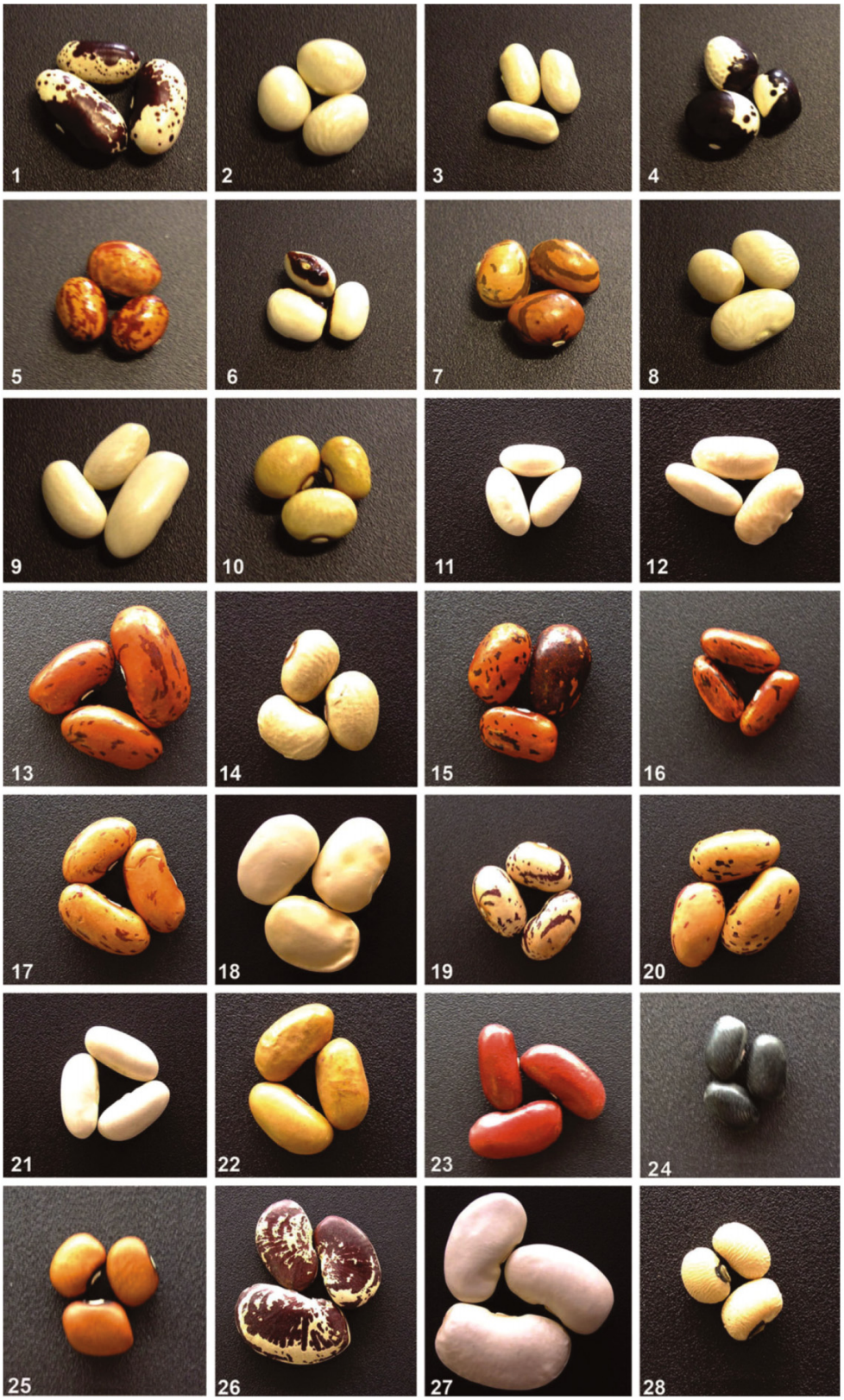


Table 2. List of primers used for each chloroplast and nuclear marker with their nucleotide sequence, amplicon length, and reference source.

	Amplicon length (bp)							
Marker	*Phaseolus vulgaris*	*Phaseolus coccineus*	*Phaseolus lunatus*	*Vigna uguiculata*	Primer name	Primer sequence (5′-3′)	T_a (°C)	References
rbcL gene	543	543	543	543	rbcL_F	GCAGCATTYCGAGTAASTCCYCA	56	Nicolé et al. unpublished
					rbcL_R	GAAACGYTCTCTCCAWCGCATAAA		Nicolé et al. unpublished
					rbcL 724R*	TCACATGTACCTGCAGTAGC		Lledó et al. 1998
matK gene	695	695	695	695	matK4La	CCTTCGATACTGGGTGAAAGAT	56	Wojciechowski et al. 2004
					matK1932Ra	CCAGACCGGCTTACTAATGGG		Wojciechowski et al. 2004
trnL intron	350	350	296	357	trnL_F	GGATAGGTGCAGAGACTCRATGGAAG	56	Nicolé et al. unpublished
					trnL_R	TGACATGTAGAATGGGACTCTATCTTTAT		Nicolé et al. unpublished
					5′trnLUAAF*	CGAAATCGGTAGACGCTACG		Taberlet et al. 1991
					3′trnLUAAR*	GGGGATAGAGGGACTTGAAC		Taberlet et al. 1991
atpB-rbcL IGS	329	325	326	331	atpB_F	GGTACTATTCAATCAATCCTCTTTAATTGT	56	Nicolé et al. unpublished
					atpB_R	ATGTAAATCCTAGATGTRAAAATAKGCAG		Nicolé et al. unpublished
					atpB_R2*	CGCAACCCAATCTTTGTTTC		Nicolé et al. unpublished
trnH-psbA IGS	365	365	365	369	psbA3′f	GTTATGCATGAACGTAATGCTC	56	Sang et al. 1997
					trnHf	CGCATGGTGGATTCACAATCC		Tate and Simpson 2003
rpoB-trnC IGS	1117	1117	1124	1136	rpoB_F	CKACAAAAYCCYTCRAATTG	54	Shaw and Small 2005
					trnCGCAR	CACCCRGATTYGAACTGGGG		Shaw and Small 2005
					rpoB_R3*	TTCTTTACAATCCCGAATGG		Nicolé et al. unpublished
trnT-trnL IGS	813	837	823	871	trnTUGU2F	CAAATGCGATGCTCTAACCT	56	Cronn et al. 2002
					5′trnLUAAR	TCTACCGATTTCGCCATATC		Taberlet et al. 1991
Total length	3556	3576	3509	3627				
ITS1	373	382	355–364	314	ITS5	GGAAGTAAAAGTCGTAACAAGG	54	White et al. 1990
					ITS2	GCTGCGTTCTTCATCGATGC		White et al. 1990
ITS2	419	418	413	401	ITS3	GCATCGATGAAGAACGCAGC	54	White et al. 1990
					ITS4	TCCTCCGCTTATTGATATGC		White et al. 1990

*Primers used only for sequencing.

Table 3. Basic information on the cpDNA and internal transcribed spacers (ITS) barcode regions, including sequence length of amplicons, inter- and intraspecific number and frequency of SNPs, and insertions or deletions (indels).

	rbcL	*matK*	*trnL*	*atpB-rbcL*	*trnH-psbA*	*trnT-trnL*	*rpoB-trnC*	ITS1	ITS2
Total No. of *Phaseolus* entries	63	63	63	63	63	63	63	63	63
Average amplicon length (bp)	543	695	338	328	366	836	1124	358	413
No. of SNPs in *Phaseolus* spp.	8	n.d.	21	14	14	53	48	65	58
Interspecific frequency (SNPs/100 bp)	1.5	n.d.	6.0	4.3	3.8	6.5	4.2	17.4	13.8
No. of SNPs in *P. vulgaris*	0	n.d.	4	0	8	3	2	6	4
Intraspecific frequency (SNPs/100 bp)	0	n.d.	1.1	0	2.2	0.4	0.2	1.6	1.0
No. of indels in *Phaseolus* spp.	0	n.d.	1	4	0	5	5	10	5
Average indel size (bp)	0	n.d.	58	2	0	7	2	4	5
No. of heterozygous sites	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3	7
Amplification success (%)	100	100	100	100	100	100	100	100	100
Sequencing success (%)	100	62	100	100	100	100	90	97	100

Note: n.d., not determined; n.a., not applicable. The percentage of sequence-tagged site PCR and sequencing success is also reported.

sequencing (see Table 2). Similar problems were experienced and solved for the ITS1 and ITS2 markers (Table 3).

The sequences of accessions corresponding to different varieties differed only at SNPs and were, therefore, easily aligned, but the sequences corresponding to different species or genera contained indels in some portions of the noncoding cpDNA, requiring manual editing of the alignments. For the ITS regions, heterozygosity was detected at only a few nucleotide positions (see Table 3), and the sites of nucleotide substitutions were recorded using the conventional code for degenerate bases of the International Union of Biochemistry.

The single sequences analyzed for cpDNA markers ranged from 328 to 1124 bp, covering a total length of 4229 bp, whereas amplicons for ITS1 and ITS2 markers averaged 358 and 413 bp, respectively. The occurrence of polymorphisms among *P. vulgaris* accessions was limited to single nucleotides; 17 SNPs were documented across the six chloroplast markers, and 10 SNPs were found for the two nuclear markers (Table 3).

The tree-based genetic identification method

The distance matrices based on the K2P substitution model for both chloroplast and nuclear regions were generated, and the average values were calculated between *Phaseolus* spp. and between subpopulations of *P. vulgaris*. Combined DNA barcode sequences showed high interspecific and low intraspecific variation rates (Table 4). The genetic distances between *P. vulgaris* and *V. unguiculata*, calculated over all barcode regions, were 0.0618 and 0.1651 on the basis of cpDNA and ITS polymorphisms, respectively. Moreover, *P. vulgaris* proved to be more closely related to *P. coccineus* than to *P. lunatus*, according to both chloroplast and nuclear markers. The average genetic distance of the former was 0.0104 and 0.0173, whereas with the latter it was 0.0231 and 0.0432 on the basis of cpDNA and ITS sequences, respectively (see Supplementary data,[1] Table S1). In *P. vulgaris*, the genetic distance estimated within varietal groups, classified on the basis of the known gene pool membership, was 0.0011 and 0 for the Andean gene pool according to cpDNA and ITS markers, respectively; for the Mesoamerican gene pool it was 0.0021 for cpDNA and 0.0020 for ITS regions (Fig. 2).

Because our focus was on the detection of polymorphisms useful for discriminating among *P. vulgaris* landraces and varieties within Mesoamerican, Andean, and Italian plant materials, further analysis was based on the DNA markers scored as polymorphic at the intraspecific level. The degree of nucleotide differentiation between congeneric species was at least 5-fold higher than were values estimated within species, whereas no significant sequence divergence rate was scored between the two different gene pools of *P. vulgaris*. Furthermore, out of 1600 intraspecific comparisons of the chloroplast and nuclear markers, 180 (11.25%) showed no significant differences between varieties.

We used the NJ tree method to analyze genetic distinctiveness using cpDNA markers. The NJ tree allows the conversion of sequence polymorphisms into genetic distances using nucleotide substitution models (Wiemers and Fiedler 2007). Based on the coalescence of conspecific populations with incomplete sampling, the NJ tree assembles all the accessions derived from one species into a single group. Separate analyses for each marker yielded NJ trees that correctly distinguished sister species and different genera, forming separate clusters for *V. unguiculata*, *P. lunatus*, *P. coccineus*, and *P. vulgaris* (data not shown). In contrast, the NJ tree built for each barcode sequence of *P. vulgaris* species was not unique because of tie trees retrieved due to low divergence values among common bean accessions. Moreover, the NJ tree constructed from the whole set of cpDNA polymorphisms produced low discrimination among accessions within the species *P. vulgaris*, owing to the complete lack or paucity of informative characters in the investigated chloroplast regions.

In the NJ tree constructed with a combination of sequence polymorphisms of the four variable chloroplast markers, members of the species *P. vulgaris*, *P. coccineus*, and *P. lunatus* were split into defined clusters, with bootstrap values as high as 99%–100%, whereas the branching nodes of *P. vulgaris* subgroups were weakly supported, with bootstrap values ≤ 60% in most cases (see Supplementary data, Figure S1). The accessions of *P. vulgaris* derived from either Mesoamerican or Andean gene pools grouped together and formed a few subclusters slightly separated from each other, with several exceptions. In four cases the gene pool

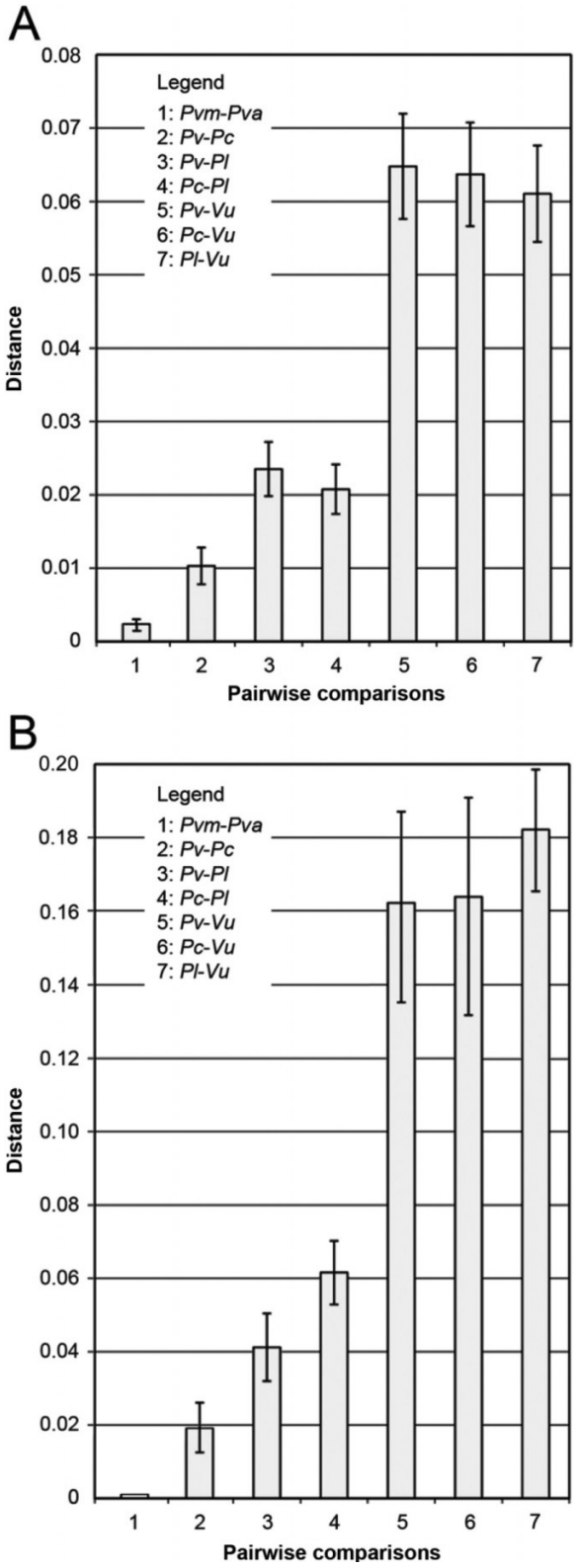
[1]Supplementary data are available with the article at www.nrcresearchpress.com/gen.

Table 4. Consensus sequence related to the 17 individual SNPs detected in the target cpDNA regions with information on the haplotypes found across all common bean (*Phaseolus vulgaris* L.) entries.

			Halotype (no. of entries)															
			Ancestral		Mesoamerican								Andean					
Marker	SNP position	Consensus sequence	Hap16 (2)	Hap09 (1)	Hap01 (1)	Hap03 (10)	Hap08 (1)	Hap12 (1)	Hap13 (3)	Hap06 (7)	Hap14 (1)	Hap15 (3)	Hap02 (15)	Hap04 (3)	Hap10 (1)	Hap11 (1)	Hap07 (1)	Hap05 (6)
trnL	14	G	A						A	A	A	A						
	183	A			C						C							
	264	T				G		G	G							G		
	332	T	A						A	A	A	A						
trnH-psbA	156	A		C						C			C					
	219	T	C	C														
	223	A	T	T														
	224	A	T	T														
	225	A	T	T														
	229	G	A	A														
	272	T				G	G	G	G									
	283	C															A	
trnT-trnL	85	A												C	C	C		
	512	A													G			
	673	T				G	G		G									
rpoB-trnC	478	G												T	T	T		
	642	A		n.d.		C	C	C	C								n.d.	

Note: Haplotypes are arranged in three main subgroups for ancestrals, Mesoamerican, and Andean gene pools. n.d., not determined. Hap01: PvA2md; Hap02: PvA7ad, PvG6aw, PvG3aw, PvB4ad, Pv1itc, Pv6itc, Pv9itc, Pv10itc, Pv13itc, Pv14itc, Pv16itc, Pv19itc, Pv24itc, Pv27itc, Pv32itc; Hap03: PvC3mw, PvG1md, PvC1ad, PvH1md, PvC2ad, PvE7md, PvH8ad, PvF1md, Pv22itc, Pv23itc; Hap04: PvH5aw, PvD6aw, Pv3itc; Hap05: PvH2mw, PvA3mw, PvB7mw, PvE6aw, PvF6aw, PvD1md; Hap06: PvH4md, Pv28itc, Pv29itc, Pv31itc, Pv33itc, Pv34itc, Pv36itc; Hap07: PvH6aw; Hap08: PvD3mw; Hap09: PvD5ad; Hap10: PvB6aw; Hap11: PvC6aw; Hap12: PvE1md; Hap13: PvF7md, Pv35itc, Pv37itc; Hap14: PvG7mw; Hap15: PvB8mw, PvC8mw, PvD8aw; Hap16: PvF8wanc, PvG8wanc.

Fig. 2. Histograms representing the inter- and intraspecific divergences calculated using chloroplast (A) and nuclear (B) markers. In addition to the mean value, the standard deviation is reported for each comparison within and between species.



was in disagreement with the geographic origin. In two of these four cases, i.e., PvH4md (from Mexico but belonging to the Andean gene pool, based on Rossi et al. (2009)) and PvD8aw (from Colombia but belonging to the Mesoamerican gene pool after Rossi et al. (2009)), the positions of the two accessions in the NJ tree were not in conflict with those of the other genotypes. In fact, PvH4md grouped with Italian cultivars and PvD8aw clustered with two Mesoamerican accessions. In four different cases, there was no indication of a gene pool, but it was possible to recover this information using NJ analysis. Two of these cases were wild accessions (PvC8mw and PvH5aw), and for these genotypes, the gene pool matched the geographic origin, as expected; the other two were domesticated accessions (PvE1md and PvH8ad), and their position in the tree suggests that they may have been transferred between regions, possibly by human intervention (see Supplementary data, Fig. S1). If all common bean accessions are classified according to their position in the NJ tree, then it is evident that 26 accessions belong to the Andean gene pool and that the remaining 29 belong to the Mesoamerican gene pool (see Table 1). It is worth noting that the ancestral bean accessions were recognized as a separate subcluster with a high confidence value and that they were grouped with another accession from Peru (see Supplementary materials, Fig. S1), the putative primary center of the ancestral wild gene pool (Debouck et al. 1993).

The NJ tree constructed using SNPs from the nuclear ITS regions, based on a lower number of polymorphisms among varieties compared with cpDNA regions, revealed an unstructured distribution of the SNPs with no subgroups for *P. vulgaris* accessions (data not shown).

The character-based genetic characterization method

Owing to the paucity of results from the above genetic distance method, a second, character-based approach was employed to identify diagnostic attributes shared between the members of a given taxonomic group but absent from a different clade that descends from the same node (Rach et al. 2008). This method does not consider indels (which were not found at the intraspecific level anyway); hence, the informative characters employed in the character-based approach were limited to SNPs.

Within *P. vulgaris*, the occurrence of SNPs depended on the marker used: for *rbcL* and *atpB-rbcL* sequences, no SNPs were detected, whereas for the other regions the number varied from two to eight (the latter for *trnH-psbA*). Among the cpDNA markers, *trnH-psbA* and *trnL* showed the highest number of SNPs, proving to be the most suitable regions for discrimination of genotypes within a species, along with the nuclear ITS1 and ITS2 markers. Of the other four chloroplast regions, only *trnT-trnL* and *rpoB-trnC* exhibited SNP markers among accessions, although at a lower frequency (see Table 3). SNP analysis of the entire chloroplast data set revealed 16 haplotypes out of the 57 accessions of *P. vulgaris* (Table 4). It is worth noting that four of these were the most common haplotypes, each being shared by 6–15 accessions. Unique haplotypes were found for 8 of the 57 common bean accessions (Table 4); the number of haplotypes (H_n) was nine for Central American, nine for South American, and five for Italian varieties. The haplotype diver-

sity (H_d) was 0.875, 0.908, and 0.688, respectively, for the three regions (Table 5), with a mean H_d of 0.877 for *P. vulgaris*.

The haplotypes based on chloroplast polymorphisms and corresponding to varietal subgroups within *P. vulgaris* species were used for the construction of a NJ tree (Fig. 3). The majority of haplotypes nested together in tightly clustered subgroups supported by low bootstrap values, with the exception of several haplotypes shared by the northern Peru and Ecuador accessions characterized by the phaseolin type I (e.g., haplotype number 16) and wild accessions. The latter finding is particularly evident for some correlated haplotypes such as Nos. 4, 10, and 11 that are linked to the Andean gene pool, as well as 6, 14, and 15 that are associated with the Mesoamerican gene pool (see Fig. 3 and Table 5). Accessions belonging to *P. coccineus*, *P. lunatus*, and *V. unguiculata* revealed unique haplotypes that were grouped separately for each species.

The number of segregating sites for chloroplast regions was 9 out of 29 Mesoamerican accessions and 13 out of 26 Andean accessions. There were eight haplotypes (H_n) for Mesoamerican accessions and nine for Andean accessions, and the estimate of haplotype diversity (H_d) proved slightly higher for the Mesoamerican (0.823) than the Andean gene pool (0.665). Even without taking the 22 modern Italian varieties into account, the haplotype diversity remained comparable between true Mesoamerican and Andean common bean accessions, with H_d values of 0.875 and 0.908, respectively (Table 5).

The ITS data set for *P. vulgaris* was not informative; all accessions, except the phaseolin type I entries that formed two separate haplotypes, were grouped together in three haplotypes, with one including 52 out of the 57 accessions (data not shown). The Italian accessions did not show any polymorphic sites, whereas the South American accessions were the most variable and scored a haplotype diversity much higher than the Central American ones. The haplotype diversity of the Mesoamerican gene pool was 0.204, but no haplotype diversity was found for the Andean gene pool (see Table 5).

Investigation into the population structure of the *P. vulgaris* germplasm by estimation of ΔK (Evanno et al. 2005) suggested that our core collection of accessions is most likely made up of three genetically distinguishable subgroups ($K = 3$), as shown in Fig. 4. In particular, 23 of the 26 Andean accessions grouped separately from most of the Mesoamerican accessions, showing a high genetic homogeneity within this gene pool and a high estimated membership for each individual. Of the 29 Mesoamerican accessions, 24 were divided into two clearly distinguishable subgroups of 14 and 10 individuals each, whereas the remaining 5 were clustered into a subgroup closely resembling that of the Andean accessions (Fig. 4). On the whole, this analysis showed that genetic diversity is low among accessions of the Andean gene pool and that accessions of the Mesoamerican gene pool are grouped into three genetically differentiated clusters. Accessions with an admixed ancestry were not detected as expected in absence of recombination. It is notable that the two ancestral accessions proved to be closely related to one of the Mesoamerican clusters.

Table 5. Summary of genetic diversity computed separately for chloroplast (A) and nuclear (B) DNA markers for subgroups of geographically distinct accessions and over all accessions of *Phaseolus vulgaris* L. and *Phaseolus* spp. (A, B) and for two different gene pools.

A

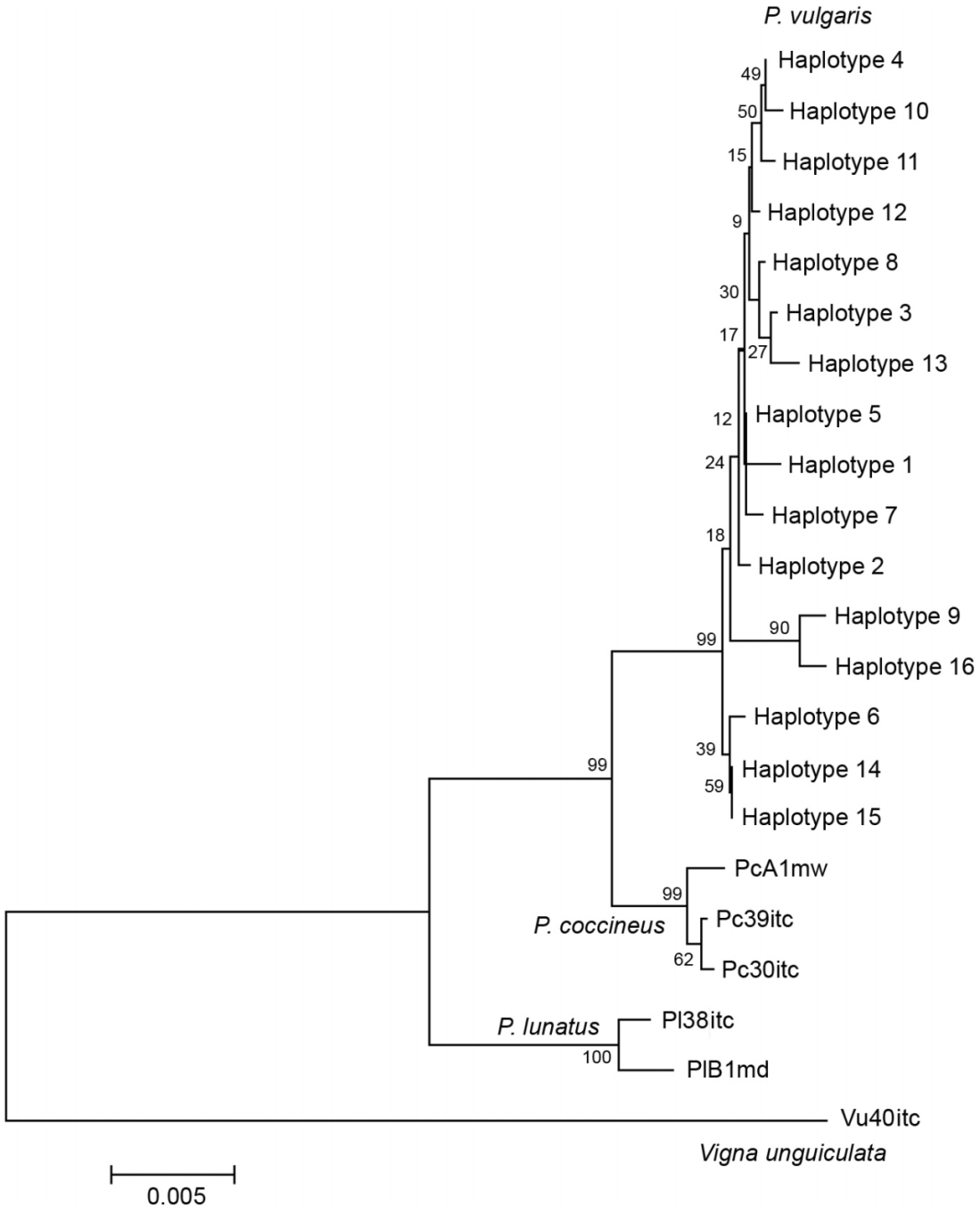
	Germplasm source		Geographical origin			Gene pool	
Genetic diversity statistics	*Phaseolus* spp.	*Phaseolus vulgaris*	Central America	South America	Italy	Mesoamerican[a]	Andean[b]
No. of segregating sites (S)	122	17	9	14	7	9	13
Haplotype number (H_n)	21	16	9	9	5	8	9
Haplotype diversity (H_d)	0.898	0.877	0.875	0.908	0.688	0.823	0.665

B

	Germplasm source		Geographical origin			Gene pool	
Genetic diversity statistics	*Phaseolus* spp.	*Phaseolus vulgaris*	Central America	South America	Italy	Mesoamerican[c]	Andean[b]
No. of segregating sites (S)	69	9	5	7	0	6	0
Haplotype number (H_n)	9	5	2	4	1	3	1
Haplotype diversity (H_d)	0.323	0.171	0.122	0.371	0	0.204	0

[a]29 accessions.
[b]26 accessions.
[c]28 accessions.

Fig. 3. Neighbor-joining tree based on the 16 haplotypes identified from the 57 bean accessions of *Phaseolus vulgaris* L. (for details on haplotypes, see also Table 5).

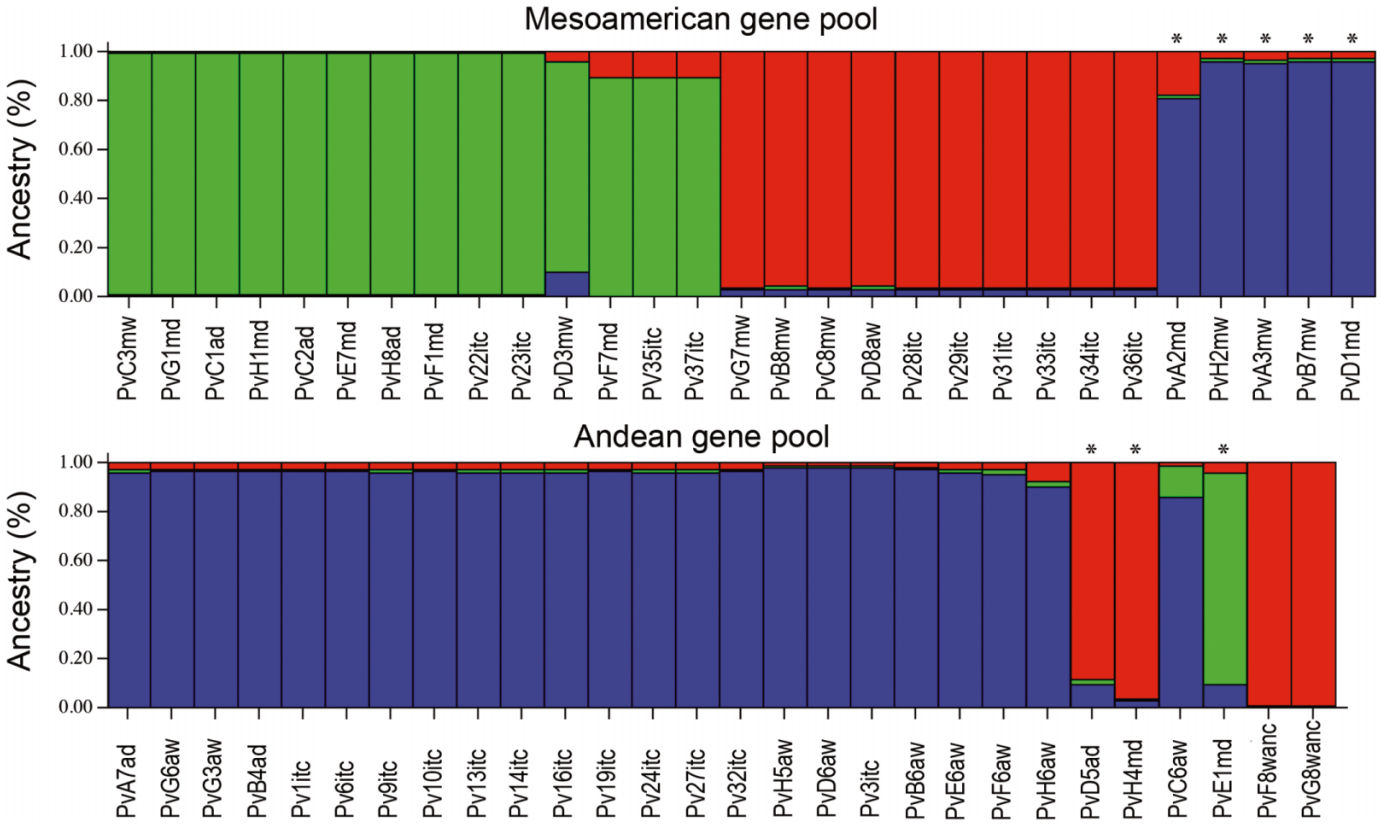


Discussion

Our results in *Phaseolus* spp. further support DNA barcoding as a powerful technique for taxonomic identification and phylogenetic analyses aimed at reconstructing evolutionary patterns and genetic distances between tightly related species. In addition to SNPs, several indels were discovered among *Phaseolus* spp. Most of the interspecific phylogenetic relationships previously identified by Delgado-Salinas et al. (1999) were confirmed by our data, with *P. vulgaris* more closely related to *P. coccineus* than to *P. lunatus*.

Because the main goal of this study was to identify those markers with the greatest polymorphism information and the best performance in intraspecific barcoding, we focused on the relevance of the nucleotide variation among accessions of *P. vulgaris*. Considering the recent criticisms formulated by the CBOL Plant Working Group of the effectiveness of single barcodes and assuming that shallow nucleotide polymorphisms would have previously been detected within species, a multilocus approach was adopted. To investigate the genetic distinctiveness of pure lines, varietal groups, and gene pools for the common bean, we used the following criteria to select the DNA regions suitable for barcoding: (*i*) a

Fig. 4. Population structure of *Phaseolus vulgaris* L. germplasm core collection as estimated with STRUCTURE software. Each accession is represented by a vertical histogram portioned into $K = 3$ colored segments that represent the estimated membership of each individual. Accessions were ordered by gene pool (i.e., Mesoamerican and Andean); improperly clustered accessions are indicated with an asterisk.



high number of sequences available in public gene banks to facilitate both primer design and the identification of species by querying nucleotide databases; and (*ii*) an appropriate substitution rate for intraspecific studies on the basis of information available in the literature.

To evaluate whether DNA barcoding is an efficient tool for the analysis of intraspecific variation and for the identification of landraces and cultivars within a species, two strategies were tested: (*i*) a phenetic tree-building approach using genetic distance data and the derived NJ tree to establish relationships among accessions of *P. vulgaris* and *Phaseolus* spp. and to determine the gene pool of origin for a set of Italian landraces; and (*ii*) a character-based system capable of reconstructing haplotypes on the basis of diagnostic characters, both fixed and variable among accessions and gene pools, for the genetic identification of varietal groups without reference to trees.

The standard tree-building approach proposed by Hebert et al. (2003) to discriminate among closely related species entails the use of sequence divergence values and the criterion of reciprocal monophyly based on the NJ tree. The employment of the distance threshold derived from the barcode gap as a tool for species delimitation is fundamental to DNA barcoding. This concept is controversial because a 10-fold screening threshold of sequence difference is present in some animals, such as birds and insects (Hebert et al. 2004; Hajibabaei et al. 2006), but is absent in others, such as cowries (Meyer and Paulay 2005). The latter observation supports the hypothesis that the barcoding gap may be an artifact of incorrect sampling (Meyer and Paulay 2005; Wiemers and Fiedler 2007). An additional tool is the NJ tree profile that allows the assignment of sequences to the correct species based on the positions of the branches relative to the cluster of the species (Wiemers and Fiedler 2007). In our study, this type of system proved to be a powerful technique to correctly cluster same-species accessions by the use of a standardized genic or intergenic region as a molecular tag. All of the sequences, whether analyzed separately or together, supported the distinctiveness of different species. In fact, even if we investigated a small number of genotypes of *Phaseolus* spp., the high nucleotide variability for these accessions, based on the occurrence of both SNPs and indels, clearly indicated the genetic distinctiveness of *P. coccineus* and *P. lunatus* from *P. vulgaris*. In contrast, the NJ tree proved poorly informative for the genetic traceability of cultivars within *P. vulgaris* species. With the exceptions of the intergenic *trnH-psbA* region and the *trnL* genic intron, the chloroplast sequences contributed little or nothing toward resolving the genetic identities of landraces and varieties. Although some concerns have arisen about the difficulties associated to the use of the *trnH-psbA* spacer (Whitlock et al. 2010), in the present study we have never experienced problems with this marker and, on the contrary, it proved to be the most informative one, followed by the *trnL*. The NJ tree derived from the chloroplast combined data set appeared to exhibit a geographically related branching pattern, with the vast majority of the Andean and Mesoamerican common bean samples clustering separately. In this work, DNA barcoding

failed to provide a clear separation between the Andean and Mesoamerican gene pools, whereas several recent studies successfully distinguished between the two groups by using both chloroplast and nuclear SSR markers or genomic AFLP markers alone (Kwak and Gepts 2009; Angioi et al. 2009; Rossi et al. 2009; Burle et al. 2010). Moreover, 12 of the 22 Italian varieties clustered with the Andean gene pool, whereas 10 accessions were classified as Mesoamerican. This result confirms previous observations about the origin and structure of European (Papa et al. 2006; Logozzo et al. 2007; Angioi et al. 2010) and Italian germplasm of *P. vulgaris* (Sicard et al. 2005; Angioi et al. 2009).

Unlike the NJ tree based on cpDNA, the distance tree generated by combining the sequences of the nuclear markers did not provide greater resolution. However, it confirmed previous studies that discourage the use of ITS for intraspecific phylogeny because of extensive intragenomic sequence variation (Álvarez and Wendel 2003). The SNPs found in ITS regions scored an average intraspecific frequency higher than that of cpDNA regions (1.3 versus 0.65 SNPs/100 bp, respectively). Nevertheless, the random distribution of ITS-related SNPs negatively affected the genetic discrimination between accessions and supports the likelihood of hybridization among accessions, which may favor the occurrence of intragenomic variation. In our study, intragenomic variation is the strongest hypothesis because the inbreeding system of *P. vulgaris* excludes a high frequency of heterozygous genotypes.

The standard tree-building approach to discriminate between gene pools and the DNA barcoding method to identify *P. vulgaris* varieties were not informative because of a slow substitution rate. For this reason, a character-based system was tested. For the DNA barcoding of multiple individuals within a species, where the genetic distances are low, it has been proposed that the character-based barcode is a more appropriate approach than the phenetic system (Rach et al. 2008). The barcode method uses DNA sequence information to generate discrete diagnostics for species identification.

To further explore intraspecific variability, the DnaSP software was used to discover combinations of character states both exclusive to a single variety and polymorphic among varieties. For the 57 *P. vulgaris* accessions (landraces and varieties), this approach allowed the detection of as few as 16 haplotypes over all cpDNA regions. These haplotypes corresponded to an equal number of subgroups, each made up of Mesoamerican or Andean accessions along with Italian accessions that clustered with either gene pool. The only exception was haplotype number 5, which was shared by mostly wild accessions from both the Mesoamerican and Andean groups. This finding raises concerns about the utility of DNA barcoding for intraspecific genetic diversity analysis, even when this technique is based on multiple loci. Although it is true that a number of SNPs and haplotypes were recovered for phaseolin type I, Mesoamerican, and Andean accession groups, it is also true that neither haplotypes nor characters specific for single accessions were found (see Table 4 for details).

In contrast to cpDNA regions, the nuclear ITS data set of *P. vulgaris* proved, as expected, poorly informative; almost all accessions clustered into a single group, except for the ancestral entries, which clustered apart. The corresponding NJ tree revealed an unstructured distribution of SNPs with neither subgroups for *P. vulgaris* accessions (data not shown) nor any segregating site among the Italian accessions. Consistent discordances among molecular data sets (i.e., chloroplast versus nuclear markers) have been observed in other taxa as well, e.g., in the Triticeae of the grasses (Mason-Gamer and Kellogg 1996) and in the Anacardiaceae (Tingshuang et al. 2004).

The estimate of haplotype diversity deserves particular attention because data based on cpDNA markers did not conflict with those based on nuclear ITS markers. When cpDNA barcodes were used, accessions belonging to the Mesoamerican gene pool exhibited a haplotype diversity higher than that estimated for the Andean gene pool (H_d = 0.823 and 0.665, respectively). Conversely, when ITS markers were used, no haplotype diversity was found for the Andean gene pool, but for the Mesoamerican gene pool, H_d = 0.204. Other works have demonstrated that the genetic diversity within the two gene pools is, in general, higher for the Mesoamerican gene pool compared with the Andean one (see, e.g., Chacón et al. 2005; Kwak and Gepts 2009; Rossi et al. 2009). This finding was further supported by independent cluster analyses with the STRUCTURE software: genetic diversity was low among accessions of the Andean gene pool that were grouped in tightly related subclusters, whereas the accessions of the Mesoamerican gene pool were grouped into three genetically differentiated subclusters. In all cases, estimated membership values were high, and admixed individuals were not present.

The 33 wild and domesticated common bean accessions can be considered a core collection of Mesoamerican and Andean gene pools, and the 22 commercial varieties are representative of Italian cultivated germplasm. Both wild and domesticated accessions within Mesoamerican and Andean gene pools proved to be formed by pure lines that are poorly distinguishable genetically from each other on the basis of the cpDNA haplotypes and ITS polymorphisms.

To characterize the genetic diversity among common beans, different approaches have been employed, from the analysis of morphology and the seed protein phaseolin to the examination of several types of molecular markers (for a review see Papa et al. 2006). These methodologies have revealed the existence of at least two major gene pools, the Mesoamerican and the Andean, and several racial groups for *P. vulgaris* (reviewed by Chacón et al. 2005; see also Rossi et al. 2009). In our study, a new molecular tool, DNA barcoding combined with NJ tree-building, was tested to determine the genetic divergence of the modern common bean cultivars and to relate them to wild and domesticated materials from the original bean domestication centers. This technique was shown to be highly reliable for identification purposes at the species level but much less informative at the variety level. Although DNA barcoding, using SNPs and indels of genic or intergenic tagged regions, provided an accurate method for the genetic identification of *Phaseolus* spp., it should not be adopted for the genetic identification of varieties within *P. vulgaris*.

The incorporation of multiple nuclear regions may be necessary to reliably identify single common bean varieties, primarily in groups that exhibit extensive hybridization and repetitive introgression patterns. In addition to ITS, other tar-

get loci for genetic identification of cultivars within *P. vulgaris* could be single- or low-copy nuclear housekeeping genes. However, the existence of high intragenomic variation can limit the utility of ITS rDNA for phylogenetic reconstructions, especially between closely related taxa (Vollmer and Palumbi 2004).

Molecular markers are applied in plant science to overcome the absence of a standard characterization system and appropriate legal protection of modern varieties and germplasm resources, as previously demonstrated in the common bean (Pallottini et al. 2004) and other major crop species such as maize (Barcaccia et al. 2003). In this context, DNA barcoding in plants could be profitably exploited for studying biodiversity at the genus level, but it does not appear useful for assessing the genetic identities of crop varieties and foodstuffs within a species.

Acknowledgements

Thanks are due to the A. Gini Foundation (University of Padova, Italy) to support S.N. during her internship at the Smithsonian Institution (Washington DC). We also thank B. Campion, Agricultural Research Council, Research Unit for Vegetable Crops (CRA-ORL; Montanaso Lombardo, Italy), for supplying the Italian bean varieties. Funding for this project was provided by the Smithsonian Institution, the Ministry of University, Research, Science, and Technology (Italy), and the University of Padova (project CPDA087818/08).

References

Álvarez, I., and Wendel, J.F. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. **29**(3): 417–434. doi:10.1016/S1055-7903(03)00208-2. PMID:14615184.

Angioi, S.A., Desiderio, F., Rau, D., Bitocchi, E., Attene, G., and Papa, R. 2009. Development and use of chloroplast microsatellites in *Phaseolus* spp., and other legumes. Plant Biol. **11**(4): 598–612. doi:10.1111/j.1438-8677.2008.00143.x. PMID:19538398.

Angioi, S.A., Rau, D., Attene, G., Nanni, L., Bellucci, E., Logozzo, G., et al. 2010. Beans in Europe: origin and structure of the European landraces of *Phaseolus vulgaris* L. Theor. Appl. Genet. **121**(5): 829–843. doi:10.1007/s00122-010-1353-2. PMID:20490446.

Barcaccia, G., Lucchin, M., and Parrini, P. 2003. Characterization of a flint maize (*Zea mays* var. *indurata*) Italian landrace. II. Genetic diversity and relatedness assessed by SSR and Inter-SSR molecular markers. Genet. Resour. Crop Evol. **50**(3): 253–271. doi:10.1023/A:1023539901316.

Barrett, R.D.H., and Hebert, P.D.N. 2005. Identifying spiders through DNA barcodes. Can. J. Zool. **83**(3): 481–491. doi:10.1139/z05-024.

Brower, A.V.Z. 2006. Problems with DNA barcodes for speciesdelimitation: 'ten species' of *Astraptes fulgerator* reassessed (Lepidoptera:Hesperiidae). Syst. Biodivers. **4**(02): 127–132. doi:10.1017/S147720000500191X.

Burle, M.L., Fonseca, J.R., Kami, J.A., and Gepts, P. 2010. Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. Theor. Appl. Genet. **121**(5): 801–813. doi:10.1007/s00122-010-1350-5. PMID:20502861.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proc. Natl. Acad. Sci. U.S.A. **106**(31): 12 794–12 797. doi:10.1073/pnas.0905845106. PMID:19666622.

Chacón, M.I., Pickersgill, B., and Debouck, D.G. 2005. Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. Theor. Appl. Genet. **110**(3): 432–444. doi:10.1007/s00122-004-1842-2. PMID:15655667.

Cowan, R.S., Chase, M.W., Kress, W.J., and Savolainen, V. 2006. 300 000 species to identify: problems, progress and prospects in DNA barcoding of land plants. Taxon, **55**(3): 611–616. doi:10.2307/25065638.

Cronn, R.C., Small, R.L., Haselkorn, T., and Wendel, J.F. 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am. J. Bot. **89**(4): 707–725. doi:10.3732/ajb.89.4.707.

DeSalle, R., Egan, M.G., and Siddall, M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360**(1462): 1905–1916. doi:10.1098/rstb.2005.1722.

Debouck, D.G., Toro, O., Paredes, O.M., Johnson, W.C., and Gepts, P. 1993. Genetic diversity and ecological distribution of *Phaseolus vulgaris* (Fabaceae) in northwestern South America. Econ. Bot. **47**(4): 408–423. doi:10.1007/BF02907356.

Delgado-Salinas, A., Turley, T., Richman, A., and Lavin, M. 1999. Phylogenetic analysis of the cultivated and wild species of *Phaseolus* (Fabaceae). Syst. Bot. **24**(3): 438–460. doi:10.2307/2419699.

Evanno, G., Regnaut, S., and Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. **14**(8): 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x. PMID:15969739.

Falush, D., Stephens, M., and Pritchard, J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, **164**(4): 1567–1587. PMID:12930761.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., et al. 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE, **3**(7): e2802. doi:10.1371/journal.pone.0002802. PMID:18665273.

Gepts, P., Beavis, W.D., Brummer, E.C., Shoemaker, R.C., Stalker, H.T., Weeden, N.F., and Young, N.D. 2005. Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. Plant Physiol. **137**(4): 1228–1235. doi:10.1104/pp.105.060871. PMID:15824285.

Hajibabaei, M., Singer, G.A.C., and Hickey, D.A. 2006. Benchmarking DNA barcodes: an assessment using available primate sequences. Genome, **49**(7): 851–854. doi:10.1139/G06-025. PMID:16936793.

Hebert, P.D.N., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. **270**(1512): 313–321. doi:10.1098/rspb.2002.2218. PMID:12614582.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., and Francis, C.M. 2004. Identification of birds through DNA barcodes. PLoS Biol. **2**(10): e312. doi:10.1371/journal.pbio.0020312. PMID:15455034.

Hickerson, M.J., Meyer, C.P., and Moritz, C. 2006. DNA barcoding will often fail to discover new animal species over broad parameter space. Syst. Biol. **55**(5): 729–739. doi:10.1080/10635150600969898. PMID:17060195.

Kami, J., Velàsquez, V.B., Debouck, D.G., and Gepts, P. 1995. Identification of presumed ancestral DNA sequences of phaseolin in *Phaseolus vulgaris*. Proc. Natl. Acad. Sci. U.S.A. **92**(4): 1101–1104. doi:10.1073/pnas.92.4.1101. PMID:7862642.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**(2): 111–120. doi:10.1007/BF01731581. PMID:7463489.

Kress, W.J., and Erickson, D.L. 2007. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH- psbA* spacer region. PLoS ONE, **2**(6): e508. doi:10.1371/journal.pone.0000508 . PMID:17551588.

Kwak, M., and Gepts, P. 2009. Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor. Appl. Genet. **118**(5): 979–992. doi:10.1007/s00122-008-0955-4. PMID:19130029.

Lledó, M.D., Crespo, M.B., Cameron, K.M., Fay, M.F., and Chase, M.W. 1998. Systematics of Plumbaginaceae based upon cladistic analysis of *rbcL* sequence data. Syst. Bot. **23**(1): 21–29.

Logozzo, G., Donnoli, R., Macaluso, L., Papa, R., Knupffer, H., and Zeuli, P.S. 2007. Analysis of the contribution of Mesoamerican and Andean gene pools to European common bean (*Phaseolus vulgaris* L.) germplasm and strategies to establish a core collection. Genet. Resour. Crop Evol. **54**(8): 1763–1779. doi:10.1007/s10722-006-9185-2.

Mason-Gamer, R.J., and Kellogg, E.A. 1996. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). Syst. Biol. **45**(4): 524–545. doi:10.1093/sysbio/45.4.524.

Meier, R., Shiyang, K., Vaidya, G., and Ng, P.K.L. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. **55**(5): 715–728. doi:10.1080/10635150600969864. PMID:17060194.

Meyer, C.P., and Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. **3**(12): e422. doi:10.1371/journal.pbio.0030422. PMID:16336051.

Mohler, V., and Schwarz, G. 2008. Genotyping tools in plant breeding: from restriction fragment length polymorphisms to single nucleotide polymorphisms. *In* Molecular marker systems in plant breeding and crop improvement. Vol. 55. *Edited by* H. Lorz and G. Wenzel. Springer, Berlin. pp. 23–38.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Newmaster, S.G., Fazekas, A.J., and Ragupathy, S. 2006. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. Can. J. Bot. **84**(3): 335–341. doi:10.1139/B06-047.

Pallottini, L., Garcia, E., Kami, J., Barcaccia, G., and Gepts, P. 2004. The genetic anatomy of a patented yellow bean. Crop Sci. **44**(3): 968–977. doi:10.2135/cropsci2004.0968.

Papa, R., Nanni, L., Sicard, D., Rau, D., and Attene, G. 2006. The evolution of genetic diversity in *Phaseolus vulgaris* L. *In* Darwin's Harvest: new approaches to the origins, evolution and conservation of crops. *Edited by* T.J. Motley, N. Zerega, and H. Cross. Columbia University Press, New York.

Pritchard, J.K., Stephens, P., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics, **155**(2): 945–959. PMID:10835412.

Rach, J., DeSalle, R., Sarkar, I.N., Schierwater, B., and Hadrys, H. 2008. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proc. Biol. Sci. **275**(1632): 237–247. doi:10.1098/rspb.2007.1290. PMID:17999953.

Rossi, M., Bitocchi, E., Bellucci, E., Nanni, L., Rau, D., Attene, G., and Papa, R. 2009. Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. Evol Appl. **2**(4): 504–522. doi:10.1111/j.1752-4571.2009.00082.x.

Rozas, J., Sánchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics, **19**(18): 2496–2497. doi:10.1093/bioinformatics/btg359. PMID:14668244.

Sang, T., Crawford, D.J., and Stuessy, T.F. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am. J. Bot. **84**(8): 1120–1136. doi:10.1111/j.1439-0523.2005.01137.x.

Shaw, J., and Small, R.L. 2005. Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). Am. J. Bot. **92**(12): 2011–2030. doi:10.3732/ajb.92.12.2011.

Sicard, D., Nanni, L., Porfiri, O., Bulfon, D., and Papa, R. 2005. Genetic diversity of *Phaseolus vulgaris* L., and *P. coccineus* L. landraces in central Italy. Plant Breed. **124**(5): 464–472. doi:10.1111/j.1439-0523.2005.01137.x.

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol. Biol. **17**(5): 1105–1109. doi:10.1007/BF00037152 .

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**(8): 1596–1599. doi:10.1093/molbev/msm092. PMID:17488738.

Tate, J.A., and Simpson, B.B. 2003. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploidy species. Syst. Bot. **28**(4): 723–737. doi:10.1016/S0169-5347(02)00041-1.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., and Vogler, A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. **18**(2): 70–74. doi:10.1016/S0169-5347(02)00041-1.

Tingshuang, Y., Miller, A.J., and Wen, J. 2004. Phylogenetic and biogeographic diversification of *Rhus* (Anacardiaceae) in the Northern Hemisphere. Mol. Phylogenet. Evol. **33**(3): 861–879. doi:10.1016/j.ympev.2004.07.006. PMID:15522809.

Tsai, L.-C., Wang, J.-C., Hsieh, H.-M., Liu, K.-L., Linacre, A., and Lee, J.C. 2008. Bidens identification using the noncoding regions of chloroplast genome and nuclear ribosomal DNA. Forensic Sci. Int. Genet. **2**(1): 35–40. doi:10.1016/j.fsigen.2007.07.005. PMID:19083787.

Velzen, R., Bakker, F.T., and Loon, J.J.A. 2007. DNA barcoding reveals hidden species diversity in *Cymothoe* (Nymphalidae). Proc. Neth. Entomol. Soc. Meet. **18**: 95–103.

Vences, M., Thomas, M., Bonett, R.M., and Vieites, D.R. 2005. Deciphering amphibian diversity through DNA barcoding: chances and challenges. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360**(1462): 1859–1868. doi:10.1098/rstb.2005.1717. PMID:16221604.

Vollmer, S.V., and Palumbi, S.R. 2004. Testing the utility of internally transcribed spacer sequences in coral phylogenetics. Mol. Ecol. **13**(9): 2763–2772. doi:10.1111/j.1365-294X.2004.02265.x. PMID:15315687.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R., and Hebert, P.D.N. 2005. DNA barcoding Australia's fish species. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360**(1462): 1847–1857. doi:10.1098/rstb.2005.1716. PMID:16214743.

White, T.J., Bruns, T., Lee, S., and Taylor, J.W. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* PCR Protocols: a guide to methods and applications. *Edited by* M.A. Innis, D.H. Gelfand, J.J. Sninsky, and T.J. White. Academic Press, Inc., New York. pp. 315-322.

Whitlock, B.A., Hale, A.M., and Groff, P.A. 2010. Intraspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. PLos ONE, **5**(7): e11533. doi:10.1371/journal.pone.0011533.

Wiemers, M., and Fiedler, K. 2007. Does the DNA barcoding gap exist? — a case study in blue butterflies (Lepidoptera: Lycanidae). Front. Zool. **4**: 8. doi:10.1186/1742-9994-4-8. PMID:17343734.

Will, K.W., and Rubinoff, D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics, **20**(1): 47–55. doi:10.1111/j.1096-0031.2003.00008.x.

Will, K.W., Mishler, B.D., and Wheeler, Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. Syst. Biol. **54**(5): 844–851. doi:10.1080/10635150500354878. PMID: 16243769.

Wojciechowski, M.F., Lavin, M., and Sanderson, M.J. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. Am. J. Bot. **91**(11): 1846–1862. doi:10.3732/ajb.91.11.1846.

Wong, E.H., and Hanner, R.H. 2008. DNA barcoding detects market substitution in North American seafood. Food Res. Int. **41**(8): 828–837. doi:10.1016/j.foodres.2008.07.005.