Origin of year-long bean (*Phaseolus dumosus* Macfady, Fabaceae) from reticulated hybridization events between multiple *Phaseolus* species

Angela M. Mina-Vargas¹, Peter C. McKeown¹*, Nicola S. Flanagan², Daniel G. Debouck³, Andrzej Kilian⁴, Trevor R. Hodkinson⁵ and Charles Spillane¹*

¹Genetics & Biotechnology Laboratory, Plant and AgriBiosciences Research Centre (PABC), Áras de Brún 2006, University Road, National University of Ireland Galway, Galway H91 REW, Ireland, ²Pontificia Universidad Javeriana, Cali, Colombia, ³International Center for Tropical Agriculture (CIAT), Apartado Aéreo 6713, Cali, Colombia, ⁴Diversity Arrays Technology Pty Ltd, Building 3, Level D, University of Canberra, Kirinari Street, Bruce, ACT 2617, Australia, and ⁵Department of Botany, School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland

*For correspondence. E-mail charles.spillane@nuigalway.ie or peter.mckeown@nuigalway.ie

Received: 18 February 2016  Returned for revision: 11 May 2016  Accepted: 24 May 2016  Published electronically: 6 August 2016

**Background and Aims**  Improved understanding of the secondary gene pools of crops is essential for advancing genetic gain in breeding programmes. Common bean, *Phaseolus vulgaris*, is a staple crop with several wild relatives in its secondary gene pool. The year-long bean, *P. dumosus*, an important crop in Guatemala, is considered particularly closely related to *P. vulgaris* and a potential source of novel variation. However, the genetic diversity and relationship to other *Phaseolus* species of *P. dumosus* remain unclear.

**Methods**  We conducted the first comprehensive investigation of *P. dumosus* genetic diversity using both nuclear and chloroplast genome markers. Our nuclear marker set included over 700 markers present within the *Phaseolus* DaRT (Diversity Arrays Technology) array, which we applied to *P. dumosus* and other relatives of *P. vulgaris* (including every secondary gene pool species: *P. acutifolius*, *P. albescens*, *P. coccineus* and *P. costaricensis*).

**Key Results**  *Phaseolus dumosus* arose from hybridization of *P. vulgaris* and *P. coccineus*, followed by at least two later hybridizations with sympatric congener populations. Existing *P. dumosus* collections have low genetic diversity.

**Conclusions**  The under-utilized crop *P. dumosus* has a complex hybrid origin. Further sampling in the region in which it arose may uncover additional germplasm for introgressing favourable traits into crops within the *P. vulgaris* gene pool.

**Key words:**  *Phaseolus dumosus*, year-long bean, common bean, genetic diversity, secondary gene pool, hybridization, crop wild relative (CWR), reticulate evolution, sympathy, inter-species.

**INTRODUCTION**

The conservation and sustainable utilization of crop genetic resources is essential for humanity to address food security in the face of a growing global population and climate change challenges (McCouch et al., 2013). Crop genetic resources consist of the genetic diversity of germplasm that can be used in crop breeding programmes (Glazmann et al., 2010) and can be organized into gene pools in accordance with the gene pool concept of Harlan and de Wet (1971). The primary gene pool of a crop consists of its own cultivars and landraces (and sometimes its wild progenitor), while the secondary gene pool consists of other related species from which agriculturally useful traits can be introgressed (Maxted et al., 2006).

Beans are a globally important component of human diets, food security and livelihoods (Broughton et al., 2003). The common bean, *Phaseolus vulgaris*, is a staple crop throughout South America, Central America and Africa, where it is an important source of dietary protein and a target of ongoing breeding and biofortification programmes (Gepts and Debouck, 1991; Schmutz et al., 2014). To meet current and emerging production challenges, novel genetic variation needs to be identified within the genetic diversity of germplasm collections and *in situ* populations (Ramirez-Villegas et al., 2010; Beebe et al., 2012; Porch et al., 2013).

While *Phaseolus* is a large genus of ∼50 species (Delgado-Salinas et al., 2006), the secondary gene pool of *P. vulgaris* is generally restricted to include only *P. dumosus* Macfady (syn. *Phaseolus polyanthus* Greemnn.), *P. coccineus* L., *P. costaricensis* Freytag & Debouck and *P. albescens* McVaugh ex R. Ramirez. & A. Delgado (Singh, 2001). Other *Phaseolus* species, such as *P. acutifolius*, constitute a tertiary gene pool because they are more distantly related, with sexual hybridization barriers (Delgado-Salinas et al., 2006). Greater understanding of the genetic variation within the secondary gene pool can guide breeding via inter-specific crosses, and can identify populations of conservation priority (Porch et al., 2013). Of the species within the secondary gene pool, *P. dumosus* (year-long bean) is closely related to both *P. vulgaris* and to *P. coccineus*, which is also a widespread crop. However, *P. dumosus* is considered to be less genetically diverse than either of these other two species, suggesting a more recent domestication (Freytag and Debouck, 2002). Importantly, *P. dumosus* displays characteristics of agronomic interest for climate change adaptation, including
tolerance to wet and high-altitude conditions (Singh et al., 1991; Schmit, 1992), and field resistance to ascochyta leaf blight, anthracnosis and white mould (Baudoin et al., 1997; Mahuku et al., 2002).

Unlike P. coccineus and P. vulgaris, ancestral wild types of P. dumosus are known only from the highlands in Guatemala where it is commonly grown. Wild P. dumosus displays an intermediate morphology with characteristics shared with P. coccineus (semituberosus root, large seeds) and P. vulgaris (epigeal cotyledons, introrse anthers). These features led to the suggestion that P. dumosus might derive from the hybridization of P. coccineus and P. vulgaris (Hernandez-Xolocotzi et al., 1959; Miranda Colin, 1967). This hypothesis was contradicted by Smartt (1973), and common ancestry was proposed as a more plausible explanation (Pinerio and Eguiarte, 1988). In recent years, the hypothesis of a hybrid origin has received support from discrepancies observed between plastid DNA (cpDNA) and nuclear DNA markers (Delgado-Salinas et al., 1999; Spataro et al., 2011). Typically, P. dumosus is resolved to be more closely related to P. vulgaris when cpDNA is used, but to P. coccineus with nuclear markers. It has also been proposed that a reticulation event may have occurred during the evolution of P. dumosus, possibly also involving P. costaricensis (llaca et al., 1994; Debouck, 1999). However, most recent analyses in Phaseolus have focused on the relationships of P. vulgaris or P. coccineus, and only small numbers of P. dumosus individuals have been investigated.

In this study, we performed a multilocus population-level study on samples of P. dumosus deposited in the germplasm banks of the International Center for Tropical Agriculture (CIAT) and the USDA. We conducted a comparison with the other members of the secondary gene pool of P. vulgaris to investigate genetic diversity, hybridization and introgression in the group. Specifically, we analysed genetic markers from the nuclear genomes of P. dumosus and the other primary and secondary gene pool species using a DArT array (Diversity Arrays Technology; Jaccoud et al., 2001), augmented with chloroplast genome markers obtained by DNA sequencing. Our study presents the first robust model for the evolution of P. dumosus and its relationship to common bean and other members of its gene pool.

**MATERIALS AND METHODS**

**Plant materials**

Sixty-three accessions of different Phaseolus species were obtained from the CIAT, Cali, Colombia, and the USDA GRIN-NGP Germplasm Collection (Table 1). Seeds of each line were germinated by scarifying and placing them on a layer of water-moistened cotton wool inside a beaker. Germinated seedlings were transferred to a growth chamber (Percival Scientific Inc., Germany) and grown under a 12-h light/12-h dark (24 h) light cycle at 22 °C/21 °C and harvested at 10 d post-germination for DNA extraction.

**DArT array**

Genomic DNA was isolated for inclusion in the Phaseolus DArT array (http://www.diversityarrays.com/index.html). DNA was isolated from single plants following a proteinase protocol (Afanador et al., 1993) and DNA quality tested by digestion with HindIII and PstI (New England Biolabs) according to the manufacturer’s instructions. Digested and non-digested DNA was run on 1 % w/v agarose electrophoresis gels and visualized under UV using a Syngene Gel Genius Bio Imaging System. DNA concentrations were quantified using a NanoDrop ND-1000 spectrophotometer (Labtech, UK). Approximately 1 µg of pure DNA from each sample was pooled into skirted V-bottom 96-well plates and used for library development following digestion with PstI/BstNI to reduce complexity. Digested samples were cloned, amplified by PCR, isolated, arrayed in solid-phase slides and hybridized as described previously (Bríñez et al., 2012). A total of 4208 polymorphic markers were detected among the clones of the P. vulgaris secondary gene pool. Marker quality was assessed using the call rate, discordance and P value; the 742 markers that displayed call rate > 95 %, discordance = 0 and P > 0.75 were used for analysis.

**Analysis of genetic diversity of P. dumosus**

A presence/absence matrix of DArT markers was used to investigate the genetic diversity and structure of the collection, and also to perform an ancestry analysis of P. dumosus for evaluation of the possibility of past hybridization events involving P. coccineus and P. vulgaris. Genetic structure was characterized by genetic distance- and model-based clustering. Genetic distance-based clustering was undertaken using principal coordinates analysis (PCoA) with a pairwise Euclidian distance matrix as an input. The matrix and PCoA were calculated using GenAlEx 6.5 (Peakall and Smouse, 2006); all available covariance and distance options were tested. The matrix was also used to calculate an analogue of FST called ΦPT, which allows population genetic differentiation to be estimated from dominant markers. Analysis of molecular variance (AMOVA) among and within all species was also estimated using Arlequin 3.5 (Excoffier and Lischer, 2010). Model-based clustering analysis was performed by applying a Markov chain Monte Carlo (MCMC) algorithm with STRUCTURE 2.3.4 (Pritchard et al., 2000; Falush et al., 2003). The estimated membership or ancestry (Q) was also determined for each sample; allelic frequencies were set as correlated among populations. K was set from 1 to 10 and the simulation was run with 105 iterations for the burn-in period and 106 iterations for the MCMC; the Evanno test (Δk; Evanno et al., 2005) and the probability of the best K (Pritchard et al., 2000) were applied to infer the optimal K value. Both tests were performed with CLUMPP (Jakobsson and Rosenberg, 2007) and Structure Harvester (Earl, 2012).

**Phylogenetic reconstructions**

Genetic distances between genotypes were computed using the Jaccard dissimilarity measure (Perrier et al., 2003). A dendrogram was constructed using the weighted neighbour-joining method (Saitou and Nei, 1987) as implemented in DARwin 5 (Perrier and Jacquemoud-Collet, 2006). The significance of each node was evaluated by bootstrapping with 1000 replications. To study conflicting phylogenetic signals potentially caused by reticulation events, the genetic distance matrix for
<table>
<thead>
<tr>
<th>Species</th>
<th>Accession</th>
<th>County of origin</th>
<th>Department/state</th>
<th>Area (where known)</th>
<th>Altitude (m)</th>
<th>Longitude</th>
<th>Date of collection</th>
<th>Date of deposit</th>
<th>Date of receipt</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. acutifolius</em></td>
<td>G35057</td>
<td>Wild</td>
<td>Mexico</td>
<td>Hidalgo</td>
<td>Tulancingo</td>
<td>1730</td>
<td>14.165</td>
<td>06/05/52</td>
<td>01/02/95</td>
</tr>
<tr>
<td></td>
<td>G36006</td>
<td>Wild</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Guatemala City</td>
<td>1455</td>
<td>19.78</td>
<td>25/01/53</td>
<td>01/02/95</td>
</tr>
<tr>
<td></td>
<td>G36026</td>
<td>Wild</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Guastatengo</td>
<td>1928</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
</tr>
<tr>
<td></td>
<td>G36027</td>
<td>Wild</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Solola</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36035</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36034</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Solola</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36031</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G35057</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36006</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Guastatengo</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36026</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Guastatengo</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36027</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Solola</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36035</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36034</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Solola</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36031</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G35057</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36006</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Guastatengo</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36026</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Guastatengo</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36027</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Solola</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36035</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36034</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Solola</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G36031</td>
<td>Landrace</td>
<td>Guatemala</td>
<td>Guatemala Dept.</td>
<td>Quetzaltenango</td>
<td>14.77</td>
<td>25/01/53</td>
<td>01/02/95</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
the wild genotypes was used to generate a phylogenetic network with SplitsTree 4 (Huson and Bryant, 2006) using NeighborNet uncorrected distances and the EqualAngle splits transformation method. The goodness of fit was assessed by the least squares fit value (LS fit) as described below.

**cpDNA and nuclear internal transcribed spacer (ITS) sequencing**

Markers for 11 regions of the cpDNA were assessed (*trnL-trnF* spacer, *matK-trnK*, *rpoC1-rpoC2* spacer, *petA-psbE* spacer, *rpl16* intron, *atpB-rbcL* spacer, *ndhA* intron, *accD-psal* spacer, *trnF-trnL* spacer, *trnL* intron and *rps14-psaB* spacer). As the last four markers could be amplified from all species, they were chosen for further use. The internal transcribed spacer (ITS) region of nuclear ribosomal DNA was also sequenced for comparison. PCR was performed with a Dyad Disciple TM Peltier Thermal Cycler (MJ Research) using GoTaq® Flexi DNA polymerase (Promega, WI) under the manufacturer’s recommended conditions. PCR fragments were resolved on 1.5% w/v agarose/Tris-Borate-EDTA (TBE) gels and visualized under UV with 1× SYBR® Safe DNA Gel Stain (Invitrogen, UK). For primer sequences and PCR programmes for both cpDNA and ITS loci see Supplementary Data Table S1. Products were purified with QIAquick PCR purification kits (Qiagen, Germany) according to the manufacturer’s instructions and eluted into 30 μL of autoclaved Millipore ultra-pure water. Products obtained from two independent samples of each accession were sequenced in both directions (GATC Biotech, Germany) using single-read 11× ABI 3730xl sequencing technology. It was intended to obtain complete sequences of two samples per accession and those that failed after four reactions were eliminated from the analysis.

**Sequence alignment and diversity analysis**

Forward and reverse sequences were assembled into contigs using the ContigExpress tool in Vector NTI® V.11.5 (Life Technology, USA) and the resulting consensus sequence was exported in FASTA file format. Alignments for each locus were constructed with MUSCLE (Edgar, 2004) and manually verified with Jalview 2.7 (Waterhouse et al., 2009) or Gblocks® (Castresana, 2000). The four cpDNA loci and one nuclear marker (nrITS) were analysed by independently calculating genetic divergence parameters in DnaSP v5 (Librado and Rozas, 2009). Allelic diversity was calculated as haplotype diversity, $\pi = n \left(1 - \sum_{i=1}^{n} j_i^2 / (n-1)\right)$, where $j_i$ is the frequency of the $i$th allele and $n$ is the number of samples, modified from Nei (1987), replacing $2n$ with $n$. Nucleotide diversity ($\pi$) was calculated according to Nei and Li (1979). Number of segregating sites ($S$) was calculated according to Nei and Kumar (2000). Apart from $\pi$ (the significance of which was evaluated by Student’s $t$-test), all tests were applied both between species and between and within accessions, and their significance was assessed by AMOVA in Arlequin 3.5 (Excoffier and Lischer, 2010) using fixation indices ($F$ statistics) calculated according to Wright (1949). Significance was tested using 1000 permutations with $\alpha > 0.05$ (Excoffier et al., 1992).
Determination of pairwise differences

The divergences per locus between individuals of the six species of *Phaseolus* were measured by determining pairwise genetic distance (distance matrix), the total average proportion of nucleotide differences between groups (*D*<sub>XY</sub>) and the net nucleotide substitutions per site between populations (*d*<sub>XY</sub>), given by *d*<sub>XY</sub> = *d*<sub>X</sub> + *d*<sub>Y</sub> − (*d*<sub>X</sub> + *d*<sub>Y</sub>) / 2 where *XY* is the average distance between groups X and Y, and *d*<sub>X</sub> and *d*<sub>Y</sub> are the mean within-group distances. Calculations were performed on the nrITS sequences and a concatenated matrix of the four cpDNA loci. Graphical representations of the divergences were made in R (http://www.R-project.org).

**Phylogenetic trees and hybridization networks**

Gene trees were built using the cpDNA and ITS sequences described above. The cpDNA alignments were analysed as a matrix of concatenated loci or as gene region partitions of this. The concatenated matrix was assembled using Mesquite 2.75 (Maddison and Maddison, 2011). Maximum likelihood (ML) trees were calculated using the GTR model (Tavaré, 1986) and drawn for the cpDNA and ITS, using MEGA5 (Tamura et al., 2011) Matrices were subjected to bootstrap testing (Felsenstein, 1985) calculated from 1000 replicates for the nrITS phylogeny and 10 000 replicates for that of the cpDNA. Branches with bootstrap values <60% were collapsed. Bayesian MCMC phylogenetic trees were created with MrBayes version 3.2 (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012). Optimal substitution models were identified for each for each locus using jModelTest and the Akaike information criterion (Akaike, 1973) following Kelchner and Thomas (2007), Posada (2008) and Darriba et al. (2012). For the priors of the model for the chloroplast sequences, three partitions were established: partition 1 corresponds to the *trnT-trnL* spacer and *trnL* intron, analysed under the GTR model; partition 2 corresponds to the *rps14-psaB* spacer, analysed under the F81 model (Felsenstein, 1981); and partition 3 corresponds to the *accD-psaI* spacer, analysed under the GTR model. A discrete *γ* distribution with six discrete categories was used to model evolutionary rate differences among sites. All parameters were unlinked to allow all partitions to vary under different rates. The nrITS region was analysed under the GTR model, with a discrete distribution and four discrete categories. Tree topologies were determined in two parallel runs, each performed with four 'heated' and one 'cold'; searches were made in R (http://www.R-project.org).

RESULTS

**DArT analysis indicates that P. dumosus is intermediate between P. vulgaris and P. coccineus**

To investigate the genetic diversity of *P. dumosus* within the *P. vulgaris* secondary gene pool, we analysed the genomic DNA of individual plants from 45 wild accessions and 28 cultivars using a *Phaseolus* DArT array (Table 1). The wild accessions included plants from all but one of the locations from which *P. dumosus* has been collected (Schmit and Debouck, 1991). Wild and domesticated forms of *P. vulgaris* and the other species of the secondary gene pool were also analysed, together with samples of the tertiary gene pool species *P. acutifolius* as an outgroup (Table 1). For each accession, DNA was taken from at least three plants, generating 152 samples once DNA quality controls were applied. In total, 4208 of the clones that we contributed to the *Phaseolus* DArT array were suitable for use as polymorphic markers in *P. dumosus*: we used these to build a binary (presence versus absence) matrix for each marker in each sample, performed a PCoA and calculated pairwise genetic diversity.

Within *P. dumosus*, three principal coordinates separated the populations into three groups. The wild accession G36286 formed a second group with two landraces (P1195389 and P1317563; top right of Fig. 1A), but all other accessions resolved to a broad group occupying an intermediate position (Fig. 1A). Hence, the three principal coordinates did not resolve accessions, landraces and cultivars into separate categories. This was confirmed by AMOVA, which indicated that variation between categories only accounted for ~15% of total divergence (*F*<sub>ST</sub> = 0.15385, *P* = 0.0127), while the majority (~85%) of the variation corresponded to polymorphisms present within the categories (Supplementary Data Table S2).

To determine the genetic structure of the secondary gene pool we also performed PCoA across all *Phaseolus* species tested. We again discovered that much of the variation (72.52%) was explained by three coordinates (47?81, 20?85 and 3?86%; Table S2) which clearly separated accessions by species (Fig. 1B). Certain cultivars of *P. vulgaris* formed an additional group with two cultivated hybrids (*P. vulgaris × coccineus*; *P. vulgaris × dumosus*). Separation between Mesoamerican and South American genotypes could be observed in this case, but only for *P. vulgaris*. Notably, *P. dumosus* genotypes occupied an intermediate position between *P. coccineus* and *P. vulgaris*. G36286 again diverged from most *P. dumosus* populations, clustering with *P. albescens*. We conclude that *P. dumosus* nuclear markers resolve the species as intermediate between *P. vulgaris* and *P. coccineus*, but provide little differentiation between *P. dumosus* populations.

**P. dumosus is closely related to sympatric populations of P. coccineus**

To analyse the population structure within *P. dumosus* accessions in more detail, we performed a Bayesian structure analysis. Using median values of log likelihood value, *L*(k), the K with the highest probability of best representing the data was *K* = 8 (Fig. 2; Supplementary Data Table S3). The probability distribution was, however, somewhat bimodal, with another peak at *K* = 6, which suggests some secondary structure among genotypes of *P. vulgaris* and *P. coccineus*. When *K* = 2, *P. dumosus* and *P. coccineus* were grouped together, but at *K* > 2 all *P. dumosus* accessions formed a single, separate cluster, in agreement with the
Fig. 1. Genetic distance among species of the *P. vulgaris* secondary gene pool. (A) Principal coordinates analysis of 742 DArT markers for *P. dumosus* indicating that 72.10% of the total genetic variation is found among the first three components. Pink diamonds, black outline, wild accessions; pink diamonds, pink outline, landraces; open diamonds, cultivars. (B) Principal coordinates analysis accounting for 72.52% of the molecular variation found among 704 DArT markers, with 85% of variation in *P. dumosus* accounted for by the first three components. *P. dumosus*: wild (pink diamonds, black outline), landrace (pink diamonds, pink outline), cultivar (open diamonds). *P. coccineus*: wild (solid dark blue circles), landrace (solid light blue circles), cultivar (open circles). *P. vulgaris*: wild (solid green squares), cultivar (open squares). *P. costaricensis* (yellow +), *P. albescens* (black +), *P. acutifolius* (grey +; outgroup). Hybrids: *P. vulgaris* × *P. dumosus* (solid pink squares), *P. vulgaris* × *P. coccineus* (solid blue squares). Controls (red triangles).

PCoA. Among *P. dumosus* samples, admixture was only detected for accession G36286, which was admixed with both *P. dumosus* (*Q* = 0.856) and *P. coccineus* (*Q* = 0.144). This accession was collected in the Guatemalan region of Quetzaltenango, in an area of Santa Maria de Jesus to which *P. coccineus* is also endemic. *Phaseolus coccineus* itself displayed much more widespread admixture; 14% of samples showed some evidence of admixture. Three percent of *P. vulgaris* samples were also admixed. Differentiation between Guatemalan and Mexican genotypes was again detected for both *P. coccineus* and *P. vulgaris* but not *P. dumosus*. For *P. vulgaris*, Peruvian genotypes also resolved separately.

To further investigate the evolutionary relationships within the secondary gene pool, a weighted neighbour-joining dendrogram was computed using Jaccard dissimilarity (Saitou and Nei, 1987; Perrier et al., 2003), using all 4208 markers to account for any variation not captured by the most polymorphic markers. The dendrogram resolved two main clusters, which we termed A and B, both of which were strongly supported (bootstrap support [BS] = 100%; Fig. 3). Cluster A comprised all accessions of *P. vulgaris*, while cluster B included representatives of *P. dumosus*, *P. coccineus* and *P. albescens*. The tree clearly resolved *P. dumosus* as monophyletic with a close affinity with *P. coccineus*, and more closely related to *P. coccineus* accessions from Guatemala than with those from Mexico (Fig. 3).

**Genetic diversity in wild *P. dumosus* is lower than in other *Phaseolus* species**

To determine whether phylogenies based on the chloroplast genome are congruent with those generated from the DArT array, we sequenced DNA from different cpDNA loci and also from the nuclear ITS region (Supplementary Data Fig. S1; Table S1). When applied to seven wild accessions of *P. dumosus*, we noted that the allelic diversity at these markers appeared to be lower (Supplementary Data Table S4). Greater diversity was observed between species, which typically only shared a few alleles (Supplementary Data Fig. S2). To investigate this further, we determined the extent of nucleotide diversity for each species in the secondary gene pool and found that diversity in *P. dumosus* was indeed the lowest of any species (*π* = 0.00035; Supplementary Data Table S5). *Phaseolus coccineus* was the most polymorphic (*π* = 0.00903), with *P. vulgaris* intermediate. For the cpDNA, most molecular variation at the *trnL* intron was due to differences between species (Supplementary Data Table 6A). An AMOVA on the concatenated matrix assessed variation at the other three cpDNA loci and showed a similar result (Supplementary Data Table S6B). The pairwise divergence from *P. vulgaris* was also low, while the *P. dumosus*–*P. coccineus* divergence was somewhat greater (0.06 versus 0.03%; Supplementary Data Fig. S3, Supplementary Data Table S7). The majority of the variation (76%) was again concluded to be due to differences between species rather than within them (Fixation index among groups *F* _CT_ = 0.76396, *P* = 0.01075; Table S7).
Conflict between cpDNA and nuclear ribosomal ITS phylogenetic reconstructions

The cpDNA and nrITS sequences were used to generate phylogenetic trees using both ML and Bayesian methods (Fig. 4). The topologies of the two trees for cpDNA data are summarized in an ML bootstrap support tree with added posterior probability values from the Bayesian analyses (Fig. 4A). Two major clades were evident within the secondary gene pool (Fig. 4A): this division was strongly supported by the Bayesian (posterior probability [PP] > 93 %) and the ML tree topology (BS > 82 %). Clade A comprised certain accessions of wild P. dumosus and a subclade of P. vulgaris (molecular diversity was too low to resolve the P. dumosus subclade further). One accession of P. coccineus (G36206) and one of P. costaricensis (G40604) also grouped in clade A as part of a polytomy with P. dumosus and P. vulgaris. Clade B comprised a P. coccineus subclade and a sister subclade containing the second accession of P. costaricensis and the single accession of P. albescens. In the P. coccineus subclade, a well-supported branch (BS = 83 %, PP = 100 %) separated accession G35650-Mexico. Topologies based on nrITS sequences were largely congruent between the Bayesian and ML analyses. Two major clades (C and D) were moderately supported by both methods (BS > 72 %, PP > 83 %); an optimal tree is displayed (Fig. 4B).

Notably, the results of the cpDNA-based phylogeny (Fig. 4A) resolved P. vulgaris as the sister species of P. dumosus. This was in contrast to the results from the nuclear genomic markers evaluated in the DArT array, in which P. coccineus had been resolved as sister to P. dumosus (Fig. 3). The nrITS trees instead agreed with the DArT analysis, with the Guatemalan subclade of P. coccineus placed as sister to P. dumosus (Fig. 5B).

Furthermore P. costaricensis was resolved as polyphyletic: one accession clustered with P. albescens (G40805), another in an unresolved position close to P. dumosus in both trees (G40604).

Probable hybridization events within the secondary gene pool

To test directly whether the conflict between plastid and nuclear DNA markers can be resolved, we developed a
hybridization network to account for possible hybridization in the *P. vulgaris* secondary gene pool (Fig. 5). The network showed three splits, each corresponding to a species group agreeing with the tree topologies identified above. Importantly, this network identified 21 putative hybridization events, comprising 16 reticulation events within species and five putative hybridizations across species. A reticulation node was generated as an explanation of the divergence between the chloroplast and the nuclear genome of *P. dumosus*, strengthening the suggestion that *P. dumosus* arose via such a hybridization event. To test this hypothesis more thoroughly, a phylogenetic analysis was performed on the DArT array data. To identify any possible reticulations, the data were fitted into a NeighborNet (Fig. 6), which robustly supported two main clusters (LS fit = 99.83 %), one comprising *P. vulgaris* accessions only and a second containing *P. dumosus*, *P. coccineus*, *P. costaricensis* and *P. albescens*. The latter cluster could be further broken down into two levels of sub-cluster, each corresponding to a species and each group of accessions. Conflicting signals or reticulations (represented by ‘boxes’ on the base of the nodes in Fig. 6) were detected both between and within the clusters, confirming that numerous between-population and between-species hybridization events are likely to have occurred.

**DISCUSSION**

Wild *P. dumosus* has low levels of genetic diversity

This study investigated the genetic diversity within germplasm of wild *P. dumosus* and close relatives, to better understand the diversity within the germplasm and to generate data that can inform conservation and breeding activities. Our results, which combine markers from nuclear and plastid genomes, indicate that nucleotide diversity of *P. dumosus* is the lowest of any species in the *P. vulgaris* secondary gene pool. Three further analyses based on the DArT data (PCoA, AMOVA and STRUCTURE; Fig. 2, Table S2) all agreed with this conclusion, and confirmed it across cultivars, landraces and wild accessions. As the wild forms of *P. coccineus* and *P. vulgaris* are considerably more diverse (Table S5), the lack of diversity in *P. dumosus* is not due to the use of inherently conserved loci. Low nucleotide diversity in wild populations can be associated with mating system, geographical distribution or gene flow (Ellstrand and Elam, 1993), although as *P. dumosus* is predominantly allogamous (Schmit et al., 1994), mating system is unlikely to be responsible. Instead *P. dumosus* populations may be subject to a constricted distribution or selection under environmental restrictions. Indeed, wild *P. dumosus* currently has a small geographical range restricted to central and south
Discrepancies between the nuclear and plastid genomes of *P. dumosus*

Our results further highlight the existence of discrepancies between nuclear and plastid genomes: DArT array analysis of *P. dumosus* strongly supports an affinity at nuclear loci with *P. coccineus*, as suggested previously (Piñero and Euguiarte, 1988; Delgado-Salinas et al., 1999, 2006). On the other hand, cpDNA sequences of *P. dumosus* resemble those of *P. vulgaris*. It should be noted that there is a discrepancy between the numbers of markers analysed in each genome, with fewer derived from the cpDNA. However, our results are in agreement with those proposed following analysis of cpSSRs (Angioi et al., 2009; Desiderio et al., 2013), which suggests that our study assessed sufficient plastid markers to draw valid inferences. Interestingly, results based on the nuclear genome reflect sympatric associations: wild *P. dumosus* more closely resembles the Guatemalan accessions of *P. vulgaris* from Guatemala (Schmit and Debouck, 1991; Freytag and Debouck, 2002) and also grows under challenging environmental conditions such as high altitude and cool wet climatic conditions. If *P. dumosus* diverged from its sister species relatively recently, this could also contribute to the low genetic diversity. Variation was only observed in G35877-Solola, a Guatemalan accession that was previously identified as unusually divergent by isozyme analysis (Schmit and Debouck, 1991). This accession is included in the core collection in the CIAT germplasm bank (Tohme et al., 1995), which suggests that the core collection captures much of the genetic diversity found among wild *P. dumosus*. Given that the CIAT bean breeding programme has recently succeeded in transferring valuable resistance genes from *P. dumosus* to *P. vulgaris* (J. Tohme, pers. comm.), it is important for ongoing crop improvement that maximal genetic diversity from *P. dumosus* is captured in both the core and base collections.
*P. coccineus* than those of any other locality. This discrepancy could have arisen if *P. dumosus* originated as a hybrid. We note that there was little evidence for introgressions between the nuclear genomes of the three species. Of the two shared cpDNA alleles, one is shared by all species of the secondary gene pool and is likely an ancestral allele retained in the conserved *rps14-psaB* locus. This allele may be associated with the split between the secondary gene pool and *P. vulgaris* itself, and may be derived from their common ancestor. A published analysis of cpSSR (Desiderio *et al.*, 2013) also reported evidence for ancestral clusters among accessions of *P. vulgaris* and *P. coccineus* associated with the separation of these lineages. A further allele (at the *accD-psaI* spacer locus) was also shared by some accessions, but is again unlikely to indicate hybridization, as it is present only in one accession of *P. coccineus* that occurs in sympatry with *P. dumosus* in Guatemala and in one of the two accessions of *P. costaricensis* from Costa Rica.

**Fig. 5.** Hybridization network showing the hypothetical reticulation events among *Phaseolus* species. The network was generated from a 60 % BS consensus tree of ML analyses for sequences of cpDNA and ITS; reticulations derived from the chloroplast genome are indicated by green lines and reticulations from the nuclear genome by orange lines; strong lines indicate reticulations between species and fine lines indicate reticulations within species (accession names have been omitted for clarity).

**Fig. 6.** Unrooted network for wild germplasm accessions of the *P. vulgaris* secondary gene pool. The network was produced from genetic dissimilarity as measured by the DArT array, applying the NeighborNet algorithm.
Phaseolus dumosus originated from ancient hybridizations

Hybridization has been suggested as a possible explanation for the discrepancy between the chloroplast and nuclear genome results for *P. dumosus*. Confirming that discrepancies between nuclear and cpDNA trees have arisen due to hybridization events is not easy with tree-building methods, which are designed to resolve any conflicting phylogenetic signal (Vriesendorp and Bakker, 2005; Morrison, 2011), so we instead developed a hybridization network to address this question (Figs 1–3; Table 1). This analysis indicates that *P. dumosus* is a hybrid between an ancient lineage of *P. coccineus* (within the Guatemalan clade) and an early diverging lineage of *P. vulgaris*. The location of the split indicates that the hypothetical hybridization event occurred shortly after the split between *P. coccineus* and *P. vulgaris*, which would also contribute to the lack of diversification in the subsequent hybrid lineage. The results of the DArT PCoA analysis also show the intermediate position of *P. dumosus* relative to the two putative parents. A further suggestive finding of the DArT analysis was the presence of admixture and genetic similarity between the representatives of *P. dumosus* and *P. coccineus* from Guatemala, including from the area of Quetzaltenango associated with the Santa Maria de Jesus volcano (Figs 1–3; Table 1). Between closely related species, a higher genetic similarity and higher admixture in regions of sympathy is considered good evidence of hybridization events occurring during speciation (Seelhaus, 2004; Grant et al., 2005; McKinnon, 2005). The evidence for such a hybridization event can be summarized as follows: (1) intermediate morphology between *P. coccineus* and *P. vulgaris*; (2) ancient forms of *P. dumosus* are found only in Guatemala and are in sympathy with ancient forms of its putative parent, *P. coccineus* (Schmit and Debouck, 1991); (3) artificial crosses can produce fertile offspring (Freitag and Debouck, 2002); and (4) molecular evidence indicates an incongruence between relationships observed from nuclear and chloroplast genome analyses, now reconfirmed in this study across all available accessions of wild *P. dumosus*. We conclude that the lineages of *P. coccineus* and *P. vulgaris* that hybridized to give rise to the initial line of *P. dumosus*, or closely related populations, have now been identified.

Consequences of the hybrid origin of *P. dumosus*

Hybrid populations are expected to have high genetic diversity because of the admixture of the parental gene pools. However, this disagrees with the low level of polymorphism identified in the wild *P. dumosus* in this study. The reticulated nature of our hybridization networks suggests this is due to backcrossing events during the evolution of *P. dumosus* (indeed *P. dumosus* groups more closely with *P. coccineus* in the DArT analyses than with *P. vulgaris*). Successive backcrossing to the pollen parent is commonly found after a hybridization event and can dilute the contribution of the maternal parental at the nuclear genome while the cytoplasmic DNA maintains its signal (McKinnon, 2005). In the case of *P. dumosus*, backcrossing may therefore have occurred to *P. coccineus* (pollen donor). In addition, the phylogenetic analysis carried out with the genome-wide DArT markers displays a conflicting signal involving *P. costaricensis*, as its exclusion reduced conflicts in the tree and increased the bootstrap values of the *P. coccineus* node. When hybrid taxa are included in phylogenetic analysis they often generate conflict and reduce bootstrap values on the nodes of the parental taxa (Rieseberg and Solis, 1991). This strongly suggests that *P. costaricensis* was involved in the evolution of *P. dumosus*, as proposed by Llaca et al. (1994). We consider that *P. dumosus* has a reticulate origin involving a hybridization event between *P. coccineus* and *P. vulgaris* and several backcrossing events to *P. coccineus*, and in which *P. costaricensis* was also involved. Timing of speciation events in this group are not known but the diversification of the species in the *Phaseolus* gene pool occurred within the last 2 million years and has been argued to be related to the Central American orogenies (Delgado-Salinas et al., 2006), a hypothesis that remains to be tested.

Conclusions

Molecular marker analyses support the hybrid origin theory of *P. dumosus* from *P. coccineus* and *P. vulgaris*. *Phaseolus dumosus* has relatively low diversity compared with its parental species, but not alarmingly low levels, given its restricted geographical distribution. Our results suggest that exploration for new germplasm is needed. Indeed, small stands of potentially wild populations have been identified in the eastern Chiapas and Alta Verapaz in Guatemala (Ramirez-Villegas et al., 2010). In addition, a geographically wider germplasm collecting mission should be undertaken in the localities of Sololá and Zacapepéquez, where our research detected some nucleotide-level variations from these regions.

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of the following. Figure S1: divergences between haplotypes observed in the sequences of the (A) nrITS and (B) cpDNA. Figure S2: cpDNA alleles shared between species. Figure S3: summary of pairwise differences in nrITS sequences (A) and cpDNA sequences (B). Table S1: markers and primers used in this study. Table S2: AMOVA derived from DArT-seq. Table S3: tabulated output of STRUCTURE analysis. Table S4: allelic diversity within accessions tested. Table S5: genetic diversity estimated from sequences of the ITS and four cpDNA markers. Table S6: analysis of molecular variance (AMOVA). Table S7: genetic distances (P distances) for the species of the *Phaseolus* secondary gene pool and close congegers as calculated from nrITS and cpDNA markers.

ACKNOWLEDGEMENTS

This work was supported by an IRCSET Government of Ireland Postgraduate PhD Fellowship to A.M.M.V., under the PhD supervision of C.S.

LITERATURE CITED


