

Title	Molecular analysis of evolutionary relationships of Phaseolus dumosus with the gene pool of common bean
Authors	Mina Vargas, Angela Maria
Publication date	2015
Original Citation	Mina Vargas, A.M. 2015. Molecular analysis of evolutionary relationships of Phaseolus dumosus with the gene pool of common bean. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2015, Angela M. Mina Vargas. - http://creativecommons.org/licenses/by-nc-nd/3.0/
Download date	2026-10-09 09:19:35
Item downloaded from	https://hdl.handle.net/10468/2270

**Molecular analysis of evolutionary relationships of *Phaseolus dumosus* with the
gene pool of common bean.**

Angela Maria Mina Vargas

**A thesis submitted to the National University of Ireland, Cork, in fulfilment of
the requirement for the degree of**

Doctor of Philosophy

Supervisor: Professor Charlie Spillane

Co-supervisor Marcel Jansen

Head of Department: Professor David Sheehan

University College Cork, Department of Biochemistry, 2015

TABLE OF CONTENTS

List of figures	vi
List of tables	vii
Abbreviations	viii
Declaration	x
Acknowledgements	xi
Abstract	xiv
1 Introduction	1
1.1 The meaning of genetic diversity for crop genetic resources and gene pool classification	2
1.2 Common bean: Significance for food security and gene pools	5
1.2.1 Available genetic resources: in-situ and ex-situ conservation of wild beans	8
1.2.1.1 In-situ conservation	8
1.2.1.2 Ex-situ conservation	9
1.3 The common bean secondary gene pool, wild and cultivars with extensive genetic variation to explore	11
1.3.1 Description and classification of <i>P. vulgaris</i> and its secondary genepool	12
1.3.1.1 <i>Phaseolus vulgaris</i> L.	12
1.3.1.2 <i>Phaseolus coccineus</i> L.	14
1.3.1.3 <i>Phaseolus costaricensis</i> Freytag & Debouck	15
1.3.1.4 <i>Phaseolus albescens</i> Mc Vaugh ex R. Ramirez & A. Delgado	16
1.3.1.5 <i>Phaseolus persistentus</i> Freytag & Debouck	16
1.3.1.6 <i>Phaseolus dumosus</i> Macfady	17
1.4 <i>Phaseolus dumosus</i> a source of genetic diversity with an unclear origin	18
1.4.1 What is known of <i>P. dumosus</i> origin and phylogenetic relationships?	19
1.5 How to evaluate genetic diversity and detect ancestral hybridisation?	22
1.6 Discerning the origin by hybridisation of <i>P. dumosus</i>	24
2 Molecular diversity of <i>P. dumosus</i> assessed from sequences of ITS and non-coding chloroplast markers	27

2.1 Introduction.....	28
2.2 Materials and Methods	30
2.2.1 Plant material selection	30
2.2.2 Plant growing conditions	30
2.2.3 Molecular analysis: sequencing of the cpDNA and ITS loci.	35
2.2.4 Data Analysis	36
2.2.4.1 Sequence alignments	36
2.2.4.2 Statistical Analysis	37
2.2.4.2.1 Description of genetic diversity estimators used to assess molecular diversity	38
2.2.4.2.2 Pairwise differences: measure of sequence divergences within the secondary gene pool of the common bean	40
2.3 Results	41
2.3.1 Overall information from the ITS and cpDNA sequences	41
2.3.2 What is the molecular diversity of wild <i>P. dumosus</i> based on nuclear and chloroplast DNA polymorphism?	43
2.3.2.1 Diversity from nrDNA sequences	43
2.3.2.2 Diversity from cpDNA sequences	46
2.3.3 What are the similarities and divergences between <i>P. dumosus</i> and the other species in the bean secondary gene pool?	50
2.4 Discussion	60
2.4.1 Wild <i>P. dumosus</i> displayed a low level of genetic diversity.	60
2.4.2 Do variation from cpDNA and nuclear ITS reveal that there has been hybridisation between <i>P. dumosus</i> , <i>P. vulgaris</i> and <i>P. coccineus</i> ?	64
2.4.3 The case of <i>Phaseolus persistens</i>	66
2.5 Next Steps	68
 3 Phylogenetic relationships between the common bean secondary gene pool..	 69
3.1 Introduction.....	70
3.2 Phylogeny reconstruction in wild bean	72
3.2.1 Alignments and a priori analysis.....	73
3.2.3 Methods used to infer the phylogenetic relationship of <i>P. dumosus</i> and its close relatives	75

3.2.3.1 The Maximum Likelihood (ML) method	76
3.2.3.2 Bayesian Markov chain Monte Carlo (MCMC) method.....	77
3.2.3.3 Consensus networks to represent incongruence among phylogenetic trees of the common bean secondary gene pool.	80
3.3 Results	82
3.3.1 What are the wild bean relationships when phylogenetic inference is made from sequences of chloroplast?	82
3.3.1 What are the wild bean relationships when phylogenetic inference is made from sequences of the ITS?	85
3.3.2 Phylogenetic network analysis, can the divergences between cpDNA and ITS topologies be explained by reticulations?	86
3.4 Discussion	88
3.4.1 The phylogenetic relationships of the common bean secondary gene pool do not follow a tree-like shape.	88
3.4.2 Hybridisation may have given origin to <i>P. dumosus</i>	89
 4 Genetic structure and relationships of <i>P. dumosus</i> inferred from whole genome DArT analysis.....	 92
 4.1 Introduction.....	 93
4.1.1 DArT markers: uses, advantages and disadvantages	94
4.2 Materials and Methods	96
4.2.1 Plant material and DNA sets.....	96
4.2.2 Parameters of the DArT array constructed for genotyping of the GP2.....	98
4.2.2.1 DArT Array quality analysis.....	100
4.2.3 Data analysis	101
4.2.3.1 Analysis of genetic diversity of <i>P. dumosus</i>	101
4.2.3.2 Analysis of ancestry and relationship of wild <i>P. dumosus</i> and the common bean secondary gene pool assessed by DArT markers	103
4.4 Results	105
4.4.1 DArT markers generated for the screening of the common bean GP2	105
4.4.2 Genetic diversity of a germplasm collection of wild <i>P. dumosus</i> assessed by nuclear markers.....	106

4.4.3 Analysis of relationships and ancestry of <i>P. dumosus</i> , using whole genome DArT markers.....	108
4.4.3.1 Structure analysis	110
4. 5 Discussion.	116
4.5.1 Whole genome molecular markers reveal a similar level of diversity for <i>P. dumosus</i> compared to diversity inferred by cpDNA and ITS	116
4.5.2 Does the genome-wide analysis give evidence of hybridisation of <i>P. dumosus</i> ?	118
5 Conclusions and future directions.....	121
5.1 Diversity and hybrid origin of <i>P. dumosus</i>	122
5.2 Remarks on the common bean secondary gene pool	124
5.3 Contribution to bean breeding and future directions.....	126
References	128
Appendices	142
Appendix I.....	143
A1. Seeds identification and plant growing conditions	143
A2 Haplotypes identified for each accession of bean.	144
A3 Haplotypes sequences and nucleotide variation.	145
A4 AMOVA derived from sequences of ITS and cpDNA.	149
Appendix II	151
B1. Galled consensus network	151
Appendix III.....	152
C1 Accessions lists	152
List of accessions representatives of the species on the common bean secondary gene pool	152
C2 AMOVA derived from DArT array	154
C3 Estimation of K by the median values of loge of likelihood value L (k)	154
C4 Estimation of K by the Δk value.	155

List of figures

2.1	Summary of pairwise differences of ITS sequences (A) and cpDNA sequences (B).....	48
2.2	Divergences between haplotypes observed at the ITS sequences (A) and haplotypes observed at the cpDNA sequences (B).....	52
2.3	cpDNA junctions of common haplotypes between bean species.....	57
2.4	Median-joining networks representing cpDNA haplotype frequencies.....	58
3.1	Illustration of methods and parameters used in the construction of phylogenetic trees.....	74
3.2	Figure 3.2 cpDNA and ITS phylogenies for wild accessions of the common bean GP2.....	84
3.3	Hybridisation network showing the hypothetical reticulation events among species of the common bean GP2.....	87
4.1	Geographic distribution of species of the GP2.....	97
4.2	Illustration of methods for array construction and characterization with DArT.....	99
4.3	Principal coordinates analysis of 742 DArT markers for <i>P. dumosus</i>	107
4.4	Plot from principal coordinates analysis, showing the genetic distance among species of common bean GP2.....	109
4.5	Patterns of ancestry of wild bean accessions.....	112
4.6	Weighted Neighbor-Joining Dendrogram for wild germplasm accessions of the common bean secondary gene pool.....	113
4.7	Unrooted network for wild germplasm accessions of the common bean secondary gene pool.....	115

List of tables

2.1	Accessions of bean GP2 and close relatives obtained from CIAT germplasm bank.....	32
2.2	General information of the sequences of the loci: ITS, <i>trnL</i> spacer, <i>trnL</i> intron, <i>rps14-psaB</i> spacer and <i>accD-psaI</i> spacer.....	42
2.3	Genetic diversity estimates, for sequences of ITS and four cpDNA of <i>P. dumosus</i> , <i>P. coccineus</i> and <i>P. vulgaris</i>	45
2.4	Genetic Distances (P-distance) calculated for sequences of ITS and cpDNA and ITS loci of the species of the bean GP2 and some close relatives.....	54
4.1	Libraries of the GP2 available in DArT technology.....	100
4.2	Average gene diversity of <i>P. dumosus</i> , <i>P. coccineus</i> and <i>P. vulgaris</i> based on 742 polymorphic DArT markers.....	106

Abbreviations

AICc	Akaike Information Criteria
AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
BS	Bootstrap value
CIAT	International Center for Tropical Agriculture
cpDNA	Chloroplast DNA
cpSSRs	Chloroplast simple sequence repeat
DArT	Diversity Array Technology
DNA	Deoxyribose nucleic acid
FAO	Food and Agriculture Organization of the United Nations
GD	Genetic distance
GP1	Primary gene pool
GP2	Secondary gene pool
GP3	Tertiary gene pool
GTR	General Time Reversible
INDELS	Insertion/deletions
ITS	Internal transcribed spacer
mce	Model of character evolution
MCMC	Markov chain Monte Carlo
ML	Maximum Likelihood
NCBI	National Center for Biotechnology Information
nDNA	Nuclear DNA

NJ	Neighbor Joining
PCoA	Principal Coordinates analysis
PCR	Polymerase chain reaction
PIC	Polymorphism Information Content
QC	Quality Control
SNPs	Single nucleotide polymorphisms
TPM	Three Parameter Model
USDA	United States Department of Agriculture

Declaration

The experiments, analysis and results reported in this thesis are entirely my own work and it has not been submitted in full or in part for another degree, either at University College Cork or elsewhere.

Angela Maria Mina-Vargas

Acknowledgements

The realization of this PhD thesis would not be possible without the collaboration of remarkable people that have stood by my side over the years. First, I would like to express my gratitude to my supervisor, Professor Charlie Spillane, for giving me the opportunity to join his lab, and for his guidance, patience and support which have made it possible for me to conclude this research.

I especially want to thank Dr. Peter McKeown, words cannot express my admiration and gratitude for all your help, advice, encouragement and patience, being an advisor and a good friend, thank you very much.

My gratitude to Dr. Aisling Doyle, for welcoming me into her home on my first day in Ireland, and since then has been an unconditional, supportive, patient and extraordinarily kind friend, thanks so much.

Thanks to Dr Nicola Flanagan for her help on the outline of the research project and who was the bridge for me to come to Ireland. Thanks to the Irish Research Council (IRCSET), for the financial support for my PhD. I also want to express my gratitude to all the people I met in *Charlie's lab*, I used to call it Babel! So many names, So many cultures, so much things I learned, it was a great life experience, thanks to you all.

To those who have helped me in the past, Professor Heiber Cardenas and Dr Joe Tohme, for having giving me opportunities and advice that opened up my perspective on life.

Finally my gratitude to the unconditional supporters that have keep me going:

Thanks to my sisters Carmen, Patricia and my niece Lina for being there for me always and unconditionally. My ladies you are my inspiration! To Tim Healy, thank you for walking beside me on the road to achieving this goal. Your love, understanding and encouragement have been the fuel that keeps me going, thanks for believing in me and for not letting me give up!

In memory of:

Parmenides Mezu Mina

*Here it is for you. I never got to see you again, I never got to say goodbye, but I will
never forget that you believed in me.*

*D&A 'Dedicado al esfuerzo de nuestras mente por alcanzar el sueño que nos
mantiene vivos'.*

Abstract

Valuable genetic variation for bean breeding programs is held within the common bean secondary gene pool which consists of *Phaseolus albescens*, *P. coccineus*, *P. costaricensis*, and *P. dumosus*. However, the use of close relatives for bean improvement is limited due to the lack of knowledge about genetic variation and genetic plasticity of many of these species. Characterisation and analysis of the genetic diversity is necessary among beans' wild relatives; in addition, conflicting phylogenies and relationships need to be understood and a hypothesis of a hybrid origin of *P. dumosus* needs to be tested. This thesis research was orientated to generate information about the patterns of relationships among the common bean secondary gene pool, with particular focus on the species *Phaseolus dumosus*. This species displays a set of characteristics of agronomic interest, not only for the direct improvement of common bean but also as a source of valuable genes for adaptation to climate change.

Here I undertake the first comprehensive study of the genetic diversity of *P. dumosus* as ascertained from both nuclear and chloroplast genome markers. A germplasm collection of the ancestral forms of *P. dumosus* together with wild, landrace and cultivar representatives of all other species of the common bean secondary gene pool, were used to analyse genetic diversity, phylogenetic relationships and structure of *P. dumosus*. Data on molecular variation was generated from sequences of cpDNA loci *accD-psaI* spacer, *trnT-trnL* spacer, *trnL* intron and *rps14-psaB* spacer and from the nrDNA the ITS region. A whole genome DArT array was developed and used for the genotyping of *P. dumosus* and its closes relatives. 4208 polymorphic markers were

generated in the DArT array and from those, 742 markers presented a call rate >95% and zero discordance. DArT markers revealed a moderate genetic polymorphism among *P. dumosus* samples (13% of polymorphic loci), while *P. coccineus* presented the highest level of polymorphism (88% of polymorphic loci). At the cpDNA one ancestral haplotype was detected among all samples of all species in the secondary genepool. The ITS region of *P. dumosus* revealed high homogeneity and polymorphism bias to *P. coccineus* genome. Phylogenetic reconstructions made with Maximum likelihood and Bayesian methods confirmed previously reported discrepancies among the nuclear and chloroplast genomes of *P. dumosus*. The outline of relationships by hybridization networks displayed a considerable number of interactions within and between species. This research provides compelling evidence that *P. dumosus* arose from hybridisation between *P. vulgaris* and *P. coccineus* and confirms that *P. costaricensis* has likely been involved in the genesis or backcrossing events (or both) in the history of *P. dumosus*. The classification of the specie *P. persistentus* was analysed based on cpDNA and ITS sequences, the results found this species to be highly related to *P. vulgaris* but not too similar to *P. leptostachyus* as previously proposed.

This research demonstrates that wild types of the secondary genepool carry a significant genetic variation which makes this a valuable genetic resource for common bean improvement. The DArT array generated in this research is a valuable resource for breeding programs since it has the potential to be used in several approaches including genotyping, discovery of novel traits, mapping and marker-trait associations. Efforts should be made to search for potential populations of *P. persistentus* and to increase the collection of new populations of *P. dumosus*, *P.*

albescens and *P. costaricensis* that may provide valuable traits for introgression into common bean and other *Phaseolus* crops.

1 Introduction

1.1 The meaning of genetic diversity for crop genetic resources and gene pool classification

Genomic variation is a principal element of the evolutionary process in nature and under domestication. Charles Darwin recognised that heritable variation was a primordial element for evolution. He established that variation was the fundamental cause of differences between ancestral and descendent characteristics, and that new **species were originated by “descent with modification” [1,2]. Generated by mutation**, recombination, DNA transposition, or epigenetic processes [3], genomic variation constitutes the raw material for the evolutionary processes of selection and adaptation [4]. The pool of genomic variation that is heritable within and between populations of species represents the species genetic diversity [5]. Genetic diversity is characterised by variation in morphological characteristics, in physiological properties and in DNA sequences (nucleotide diversity).

The level of genetic diversity found in a species informs about the history of selection, recombination, migration and mating systems among and within species populations [6]. Analysis of genetic diversity of cultivar species is an important area of biological research. Research in genetic diversity is highly valuable for human society; the diversity of cultivated plants and their wild relatives is a primary source of knowledge for agriculture, breeding programmes and the generation of food security strategies. There is a demand for research focused on evaluating the genetic diversity of cultivar species, and this is in order to generate strategies that guarantee food security

(<http://www.fao.org/biodiversity/components/plants/en/>). Reduced genetic diversity in crops represents a major threat for world-wide food security because it limits the resources available to respond to current and future diseases, plagues, insects and changes in environmental conditions. Three main factors contribute to the reduction of crop genetic diversity: replacement of landraces by modern cultivars, the loss of growing habitat of wild types and the intense breeding and highly stringent quality requirements imposed by processors and consumers [7,8].

Significant examples of crop losses due to low genetic diversity are; the devastation of around 95% of the Taro crop (*Colocasia esculenta*) in Samoa in 1993 due to infestation by the taro leaf blight (*Phytophthora colocasiae*) [9], the drastic reduction in corn (*Zea mays*) yields in the USA in the 1970s, caused by an outbreak of corn blight (*Cochliobolus heterostrophus*) [10], the rice brown spot disease (*Bipolaris oryzae*) in Bengal contributing to the Indian famine in 1943 [11] and the devastation of the potato crop in Ireland in the 1840s caused by the potato blight (*Phytophthora infestans*). In the last two decades banana production around the world has been adversely affected by the risk of new strains of Panama disease caused by *Fusarium oxysporum* f. sp. with the risk of eradication of some of the wild and cultivated banana varieties, including the most widely planted clone, ‘Cavendish’ [12].

“Genetic diversity is a defence against the problems caused by genetic vulnerability” [13]. With the challenges of a constantly growing human population a major concern for world-wide food security is the preservation and expansion of the diversity of Crop Genetic Resources (CGR). In recent times and due to the constraints generated by climate change, there is a rising interest and need for research that assesses the genetic

diversity of crop wild relatives (CWR) as a fundamental resource for world food security [14]. CWR are a source of genetic material that typically possess a greater set of genetic variation than the ones found in the cultivar types. CWR resources have been used to improve resistance against pests and diseases, tolerance to stressful abiotic conditions, improving nutritional quality and agronomic performance [15]. Overall the genetic resources of a crop species are represented by gene pools. A gene pool of a population is the total set of genotypes, genes, gene combinations and alternate or allelic forms of those genes within a population. Legume CGRs are described by gene pools in accordance with the Harlan and the Wet classification [16]. This classification is based on the ability of related taxa to cross and generate viable and fertile offspring in this order

Primary gene pool (GP1) hybridisation is easy and the hybrid is healthy and fertile.

Secondary gene pool (GP2) hybridisation is possible but difficult, hybrids are weak with low fertility.

Tertiary gene pool (GP3) hybridisation most difficult, low viability in hybrids, very low fertility.

The GP1 is formed by the cultivar and wild types of the species, generally without reproductive barriers. Progeny are viable, fertile and easy to generate. The resources in GP2 and GP3 correspond to taxa closely related to GP1, these resources have a high level of genetic similarity with GP1 but the ability to hybridise reduced due to reproductive barriers. Recently and thanks to the advances of molecular technology which has allowed evaluation of wider resources, that the gene pool classification has being extended further to quaternary gene pool (GP4) which includes all potential

genes and germplasm of species that are not sexually compatible with the crop but can be used for its transformation [17,18].

Overall the gene pool categorisation has provided an ample framework for the analysis of relationships among crops and its wild relatives. The genetic diversity captured among crop gene pools is a key element for crop improvement, sustainability and for food security. To comprehensively manage the diversity of a crop, it is fundamental to understand the structure and evolution of its gene pools.

1.2 Common bean: Significance for food security and gene pools

Beans (dry, snap and green) are an ethnic food that has been integrated into human diets all around the world. By 2012 the global production of bean (dry and green) was approximately 44.5 million tonnes. India and Brazil are the major producers of dry seeds and China the major producer of green beans (<http://faostat3.fao.org/faostat-gateway/go/to/home/>). Europe is a minor producer with around one million tons of bean (dry and green). Beans in Europe are grown for consumption as well as ornamental plants [19]. In Latin America, Asia and Africa beans besides being a primary food are also an important source of household income for small farmers [20,21]. Nutritionally beans are characterised by their high protein content which is roughly double that of most cereals. Beans are also rich in fibre and essential micronutrients (iron, B vitamin as folic acid and thiamine) and macronutrients such as

lysine [22]. Fortification of these nutrients is part of the objectives of current breeding programmes, along with breeding for yield increase and breeding for better response to environmental constraints [23].

The common bean (*Phaseolus vulgaris* L.) is listed as a key food security crop for food insecure areas of East and South Africa, where beans are a primary source of protein and minerals for more than 200 million people [22]. In Burundi, Rwanda, Uganda, Kenya, Tanzania and Malawi more than 15% of the protein consumption comes from beans [24,25]. Bean production in food insecure areas is mainly threatened by diseases, insect pests and changing climate giving rise to weather conditions resulting in severe drought, excessive temperature as well as excessive rainfall periods [26–28]. Drought alone is a major threat to bean production in Ethiopia, Kenya, and Tanzania [29]. Currently breeding programs for environmental constraints target resistance and tolerance genes and traits already present in the bean gene pool. These genes can potentially be used for improvement of the cultivated varieties.

The main genetic resources related to common bean are classified as: Primary gene pool formed by *P. vulgaris* wild and cultivars, secondary gene pool formed by *P. dumosus*, *P. coccineus*, *P. costaricensis*, and *P. albescens* and tertiary gene pool formed by *P. acutifolius* and *P. parvifolius* [30]. Hernández-López et al. [31] and Porch et al. [32] are two recent reviews that provide a summary of the research done in common bean genetic resources, diversity, genomics and breeding programs up to 2013. It is acknowledged that within the primary gene pool of *P. vulgaris* there are two primary centres of diversity (Andean and Mesoamerican), which correspond with

the identification of two wild gene pools [33,34]. The recently sequenced genome of *P. vulgaris* [35] provided further evidence that beans were domesticated independently in Central America and South America. The dispersion of *P. vulgaris* around the world has generated secondary centres of diversity in Europe, Asia and Africa [36].

The knowledge generated on the primary gene pool of *P. vulgaris* has greatly contributed to the success of breeding programs and bean improvement [37–39]. However, the use of a narrow source of diversity limits the capacity for improvement. Additionally some characteristics, such as adaptation to cold or dry seasons and acid soils, can be poorly present or absent in the GP1 [30].

Examples of conditions that affect bean production around the world are: abiotic stresses such as low soil fertility, lack of minerals, excess of toxic elements, and biotic stresses that reduce production due to susceptibility to bacterial diseases such as, halo blight (*Pseudomonas syringae* pv. *phaseolicola*), common bacterial blight (*Xanthomonas campestris* pv.) and bean anthracnose (*Colletotrichum lindemuthianum*), among others [27,40]. Hence to face today's environmental and production challenges for common bean, novel sources of diversity need to be explored [32,38,41]. Among the species in the secondary and tertiary gene pool of the common bean, there are several wild forms with wide genetic diversity. These wild forms have not been subjected to the genetic bottlenecks generated by selection for domestication; therefore contain greater variation to survive environmental constrictions [42,43]. However, the use of wild relatives and close relatives material for bean improvement is limited due to the lack of knowledge about the genetic

variation and plasticity of these materials [32]. A greater understanding of the nature of within species genomic variability and the genetic discontinuities between species, will facilitate the breakdown of hybrid incompatibilities for crop improvement.

1.2.1 Available genetic resources: in-situ and ex-situ conservation of wild beans

1.2.1.1 In-situ conservation

In-situ conservation of wild bean basically corresponds to populations that are within protected areas and within those the number and conditions of the populations are still unreported. In Guatemala, Mexico and Costa Rica, the extent of distribution of wild bean populations still under exploration. The actual in-situ conservation corresponds to populations that happen to be within areas of National parks, or in the called 'Reserva de la Biosfera' areas; a strategy implemented in Mexico and in other countries in Central America and South America which aims to protect wild relatives of several crops in Latin America. However there is no precise information about what species of bean and how many populations are included within this protected areas. Porch et al [32] presented a list of the possible wild populations for *P. vulgaris* and the other species of the GP2, based on herbarium records and seed bank records. For instance they report that there is possibly eleven populations of wild *P. dumosus* in Guatemala and one population in Mexico, but the gap between hypothetical populations and the germplasm representations is of 100% missing collections. The data also shows that with the exception of *P. vulgaris* and *P. coccineus* in Mexico, the numbers of known

populations are low for all species of the GP2, and the actual existing populations are likely lower than reported or populations may not exist anymore since the distribution areas are highly reduce by urbanization and agriculture.

Recent explorations and collections of *Phaseolus* species across Mexico [44] have found high genetic erosion in some regions to the point that entire populations of wild bean have disappear. Similar situation is occurring in Guatemala where the distribution area for wild *P. dumosus* Guatemala is threatened by urbanization and agriculture (<https://www.hort.purdue.edu/newcrop/1492/beans.html>) and in Costa Rica where populations of wild *P. costaricensis* are highly affect by human perturbation of agroecosystems [45].

1.2.1.2 Ex-situ conservation

The FAO report on germplasm collection (FAO 2010) indicates that around the world the ex-situ germplasm collections for *Phaseolus* beans includes approximately 250.000 accessions. The collections include different species as well as landraces, cultivars, breeding materials and wild types. The largest collection of bean seeds is held at the International Center for Tropical Agriculture (CIAT); Cali, Colombia, where 36,000 accessions representing 44 taxa from 110 countries are safeguarded. This gene bank is part of the Consultative Group for International Agricultural Research (CGIAR). The USDA germplasm bank at Western Regional Plant Introduction Station, Pullman, Washington, USA, is the second largest collection with almost 15,000 accessions, which are also represented in the CIAT collection [36]. The

majority of accessions represented in both of these collections correspond to *P. vulgaris*, this includes cultivar, landraces, breeding materials and wild types. These collections represent the primary centres of origin of *Phaseolus* in the Americas and also includes smaller representations from materials from Europe, Asia, and Africa. Cultivar and landraces types are also better represented than wild types. There are less representations for other bean species, with only 1,049 accessions of the secondary gene pool and 335 accessions for the tertiary gene pool.

Smaller seed banks are held in Mexico, Brazil and Costa Rica. In Europe the main collections are those of Germany (approximately 9,000 accessions) and Russia (6,000 accessions). Smaller collections of European landraces of *P. coccineus* are held in Italy and Spain. Less information is available on collections of bean in Africa, but reports are of the existence of collections in Uganda, (National Crops Resources Research Institute (NCRRI) and possible collections in Burundi, Democratic Republic of Congo, Ethiopia, Kenya, Republic of South Africa, Rwanda, Tanzania, Uganda and Zimbabwe [37].

1.3 The common bean secondary gene pool, wild and cultivars with extensive genetic variation to explore

Aside from the diversity in tolerance to abiotic stresses and adaptability to a range of environments, the species in the GP2 also show resistance to several diseases such as: white mold (*Sclerotinia sclerotiorum*) [46], ascochyta leaf blight (*Phoma exigua*), angular leaf spot (*Phaeoisariopsis griseola*) bean anthracnose (*Colletotrichum lindemuthianum*), fusarium root rust (*Fusarium oxysporum*), common bacterial blight (*Xanthomonas axenopodis*) and bean golden yellow mosaic (BGMV) [30,40,47,48]. Resistant lines have been generated from interspecific crosses between *P. vulgaris*, and species in the GP2, for example for resistance to anthracnose [49] resistance to white mold [46] and common bacterial blight-resistant [50].

Crosses are relatively easy between the species of GP2 and also between them and *P. vulgaris*, particularly when *P. vulgaris* is used as the female parent [28,51]. However stable phenotypes are difficult to obtain since the hybrid progenies are often inviable or partially sterile [52]. The species *P. costaricensis* and *P. albescens* have not been domesticated. Wild *P. costaricensis* is found in Costa Rica and western Panama. This species is cross-pollinated, evaluation and use of its germplasm collections have shown its capacity for wide-crossing and gene transfer with the other species of the GP2 as well as the common bean [53]. The following sections of this introduction refer to several questions around of *P. dumosus* genetic diversity, origin and evolutionary relationships of the species *P. dumosus* the main subject of this research.

1.3.1 Description and classification of *P. vulgaris* and its secondary genepool

1.3.1.1 *Phaseolus vulgaris* L.

Common names:

English; common bean, haricot bean, kidney bean, navy bean, dwarf bean, flageolet bean, French bean, garden bean, green bean.

French; haricot common, haricot, haricot vert.

Portuguese; feijão, feijoeiro.

Spanish; nuñas, chícharo, frijol, habichuela, judía común, nuña, poroto.

Swahili; mharagwe.

Widely consumed around the world, the common bean *Phaseolus vulgaris* L. is one of the most important crop plants. *P. vulgaris* is a legume, taxonomically included in the Fabaceae family which also includes other major crops such as *Glycine max* (L.) Merr. (soybean), *Cajanus cajan* (L.) Huth. (pigeon pea), *Medicago sativa* L. (alfalfa), *Arachis hypogaea* L. (peanuts) and *Lens culinaris* Medik. (lentils). The exact number of *Phaseolus* species is unknown, currently it includes seventy six species native to Central America and South America [51]. Wild *P. vulgaris* is an indeterminate vine plant, with a stigma lateral introrse, aerial shoot and a fibrous root. Leaves are trifoliate and bracteoles are large (Mesoamerican form bracteoles are larger than calix). Flowers are papilionaceous, zygomorphic with a bi-petalled, keel in open unbranched clusters (racemes), each flower contains 10 stamens, nine of which are fused into a partial tube, with the tenth stamen free. One multi-ovuled, ovary is positioned above the sepals, petals and stamens. The style is upturned and spiralled, with a collar of fine hairs below the stigma. Flowers may be white, pink, yellow, purple or bicolour, the flowers colours are independent of seed colour. *P. vulgaris* is predominantly self-fertilized, seeds of wild types are small oblong rounded, kidney-

shaped or ellipsoid. Seeds vary in colour from black, brown and yellow types to red, white, speckled and flecked forms. Mature pods are dehiscent. Germination is epigeal (cotyledons of the seedlings located above the soil surface) and requires five to seven days [51,54,55]. Cultivated varieties of *P. vulgaris* have larger seeds and growing types can be determinate or indeterminate and grow as bush, twining or climbing forms. Segregated populations from spontaneous crosses between cultivar types and wild types are commonly found in areas of sympatry [51].

Common bean grows at intermediate altitudes 800 to 2000 meters and temperatures between 17 to 26 °C, its distribution is associated with the pattern of rainfall [56]. *P. vulgaris* best grows in well-drained soils during the raining season, until blooming and beginning of pod filling, and maturation takes place during the dry season. Wild *P. vulgaris* is a perennial plant with tendency to annual plant while domesticated forms are predominantly annual [31].

Wild gene pools of *P. vulgaris* are found in Central America (Mesoamerican gene pool) and South America (Andes gene pool) [31,57,58]. Within these two gene pools further diversification has been detected [59]. Overall the Mesoamerican gene pool displays a higher diversity and stronger genetic structure than the Andean gene pool. Wild populations are fragmented by the geological discontinuities of the Lake Nicaragua, the Isthmus of Panama, and the Andean mountain range of the Americas [57].

Domestication of *P. vulgaris* occurred independently in Central and South America and the separation of wild gene pools is extended to the cultivated forms [60–63]. A large-seeded type was developed in Peru around 8,000 years ago, and a smaller-seeded

type, developed in Mexico around 7,000 years ago. There are partial reproductive barriers between the Andean and Mesoamerican gene pools for wild as well as for domesticated populations [64,65]. Domestication syndrome has been documented among populations of *P. vulgaris*. Reduction of diversity as a consequence of domestication is more evident in the Mesoamerican genepool than in the Andean genepool [66].

1.3.1.2 *Phaseolus coccineus* L.

Common names:

English: scarlet runner bean.

French: haricot d'Espagne.

Portuguese: feijão-trepador.

Spanish: ayocote, judía.

Wild *P. coccineus* roots are tuberous, strong and fleshy. Inflorescence are zygomorphic, the flowers are red scarlet or white the length of the corollas ranges from 1.5 to 2.5 cm, with the larger flowers on the cultivars. The stigma can be located introrsely, apically or extrorsely, an extrorse stigma and stamens with short filaments contributes to avoid self-pollination. *P. coccineus* is predominantly an allogamous plant. Wild *P. coccineus* have explosively dehiscent pods; most of domesticated populations generally have indehiscent pods. The hilum is ovate to oblanceolate in *P. coccineus* (orbicular in *P. dumosus*). Cultigens rarely develop more than six seeds per pod, whereas feral types develop 4-6 seeds and wild forms 4-9 seeds. The cultivated forms have longer flowers and impermeable seed coat, seeds are linear oblong to oblong with different colours: white, black, beige and violet. Seedling size is larger than wild plants. *P. coccineus* has a hypogeal germination (cotyledons of the seedlings

located below the soil surface) [51,67,68]. The wild forms are distributed in temperate regions from Mexico to Panama. It is a perennial plant, commonly grown as an annual crop in highlands mountains, between 1,800 and 3,000 meters; as it is more tolerant of cool conditions than other *Phaseolus* species.

Archaeological remains indicates that scarlet runner bean was probably domesticated in Mexico. The states of Chiapas, Puebla and Oaxaca are pointed as centre of origin of this specie [69]. Populations of wild and cultivar *P. coccineus* are highly polymorphic and display strong genetic differentiation between accessions [19,67]. In Central and South America the runner bean is grown as a substitute of the common bean in areas with wet and cold summers. This is the bean generally grown in the United Kingdom, The Netherlands, South of Italy Spain and northern areas of Europe [70]. It is occasionally grown in Africa e.g. Ethiopia, Kenya, Uganda, South Africa and Asia.

1.3.1.3 *Phaseolus costaricensis* Freytag & Debouck

The species *P. costaricensis* have not been domesticated only wild forms are known to exist, the wild plants are large climbing vines with a strong climbing growth habit, reaching 3 to 6 meters in height. This species is characterised by its fibrous roots, large and long bracteoles, and very numerous flowers: dark pink, lilac or fuchsia. Germination is epigeal. Its geographical distribution occurs in the humid montane forests of Costa Rica and western Panama at altitudes between 1,400 and 2,100 meters. *P. costaricensis* is cross-pollinated and the evaluation and use of its germplasm

collections have shown its capacity for wide-crossing and gene transfer with the other species of the GP2 as well as the common bean [53,71].

1.3.1.4 *Phaseolus albescens* Mc Vaugh ex R. Ramirez & A. Delgado

There are only a few collections of *P. albescens*, found in a very restricted area of mountain forests in western Mexico [51] some aspects of its morphology are still unconfirmed. Bracteoles are as long as calix 3-4 mm, flowers are long, colour pale purple fading yellow. Pods contain up to 5 seeds, germination is epigeal. This species is considered endemic to southern Mexico where it occurs in the regions of Colima, Guerrero, Jalisco and Michoacán. Based on occurrence data provided by Crop Wild Relatives and Climate Change (2013), its extent of occurrence (EOO) is less than 20,000 km².

1.3.1.5 *Phaseolus persistentus* Freytag & Debouck

With only one collected specimen, this species was described as small climbing vine, with wide bracteoles and unique short pods. Root seeds and flowers for this species are to date unknown [51]. The species type was collected in Guatemala in the area of Sacatepéquez. The taxonomic classification for this species is yet to be clarified, based on morphology, the species has been classified into the Falcati section (*Leptostachyus* group) while molecular analysis has assigned the species to the *Phaseoli* section (*Vulgaris* group) [72,73].

1.3.1.6 *Phaseolus dumosus* Macfady

Synonyms:

(=) *Phaseolus coccineus* subsp. *darwinianus* Hern.-Xol. & Miranda-Colín

(=) *Phaseolus coccineus* subsp. *polyanthus* (Greenm.) Maréchal et al.

(=) *Phaseolus flavescens* Piper.

(=) *Phaseolus polyanthus* Greenan.

Common names:

English: year-long bean, yearlong bean.

Spanish: Acalete, botil, piloya, cache.

The species *Phaseolus dumosus* Macfady (syn. *Phaseolus polyanthus* Greenm) is known as the year bean or year-long bean and is cultivated mainly in Central America. The vines of *P. dumosus* are very large and have vigorous climbing growth habits, roots are fibrous, fasciculate and thick. The inflorescences have primary bracts and long, narrow bracteoles that give the appearance of a spike. The flower head contains six to sixteen stems, flowers are purple, white or purplish pink in the wild form. The stigma is capitata and introrse. Bracteoles are linear to narrow oblong and heavily covered with strigose hair. Pods are straight and seeds are rounded with an orbicular helium. Seeds are medium to large and are highly variable in colour, the seed coat is thick, as in *P. coccineus*, but is permeable to water. *P. dumosus* has an epigeal germination.

1.4 *Phaseolus dumosus* a source of genetic diversity with an unclear origin

The taxonomic status of *P. dumosus* was reconfirmed by the identification of wild ancestral forms collected in Guatemala [74]. No archaeological records have been reported for this species. Unlike *P. vulgaris* and *P. coccineus*, ancestral wild types of *P. dumosus* have been found only in the highlands in Guatemala. This species did not undergo widespread dispersion and it is mainly cultivated in Central America and in South America. *P. dumosus* is often grown by farmers as substitute for common bean, it is found in an altitude range from 800 to 2600 meters [74,75]. *P. dumosus* has a long flowering period (two to five months) and can have two flowering and fruit-bearing periods per year if the rainy season is heavy. *P. dumosus* grows better in deep, organic, damp and well-drained soils with pH 6.2 to 6.5 and it tolerates a degree of shade.

Compared with the wild forms of other cultivated species, *P. dumosus* shows less differentiation from its wild form, which could be the result of its more recent domestication [51]. The cultivated form is found in Mexico (Puebla, Veracruz, Oaxaca and Chiapas), Guatemala (Huehuetenango, San Marcos, Quezaltenango, Totonicapán, Baja and Alta Verapaz, Sololá, Chimaltenango and Sacatepéquez) and in some areas of Costa Rica, Jamaica, and Dominican Republic. In South America, *P. dumosus* is found in secondary vegetation, including wooded vegetation, in Peru, Venezuela, Colombia, and Ecuador [76,77].

The year -long bean displays a set of characteristics of agronomic interest, not only for the direct improvement of *P. vulgaris* but also as a source of valuable genes for

climate change defence. Compared with other *Phaseolus* species, *P. dumosus* has a higher tolerance to high precipitation, and has tolerance to acid soils and low temperatures. *P. dumosus* can grow in wet and humid tropical conditions having resistance to several fungal and bacterial diseases associated to these conditions [33,75]. For instance *P. dumosus* has shown a high and stable field resistance to ascochyta leaf blight and anthracnose, and some of its accessions have displayed the highest reported level of resistance to white mold [39,49, 50]. There are approximately five hundred accessions of *P. dumosus* kept in the germplasm banks of CIAT and USDA, of those less than three percent are accessions of ancestral wild types collected in Guatemala. The majority (around 450 accessions) of the seed kept in the collections correspond to accessions from cultivars and landraces.

1.4.1 What is known of *P. dumosus* origin and phylogenetic relationships?

As with all cultivated beans, *P. dumosus* is a diploid $2n=2x=22$. Wild *P. dumosus* displays an intermediate morphology: with characteristics shared with *P. coccineus* (semi tuberous root and a larger seed) and with *P. vulgaris* (epigeal cotyledons and introrse stigma) [78]. Early observations of the intermediate morphology were the basis for the hypothesis that this species derives from a hybridisation between *P. vulgaris* and *P. coccineus* [55,79]. This hypothesis was initially contradicted by Smartt [80], indicating that other aspects of the morphology and the intermediate characters were indications of a common ancestry rather than a hybridisation. Schmit and

Debouck [74], considered the hypothesis of common ancestry a plausible event in which the three species (*P. vulgaris*, *P. coccineus* and *P. dumosus*) have evolved from a common ancestor in Mesoamerica. They did not discard the possibility of *P. dumosus* being either a hybrid or a variant of *P. coccineus* without hybridisation involved. This was supported by the analysis of restriction enzymes carried out by Piñero and Eguiarte [81] which found little variation among the evaluated accessions of *P. dumosus*.

In recent years the hypothesis of *P. dumosus* being originated from a hybridisation event has been reinforced by results from molecular analysis. A number of phylogenetic analyses have provided an insight into the relationships around the GP2, with some studies including a few of the accessions of *P. dumosus*. Evaluations have been made using chloroplast DNA (cpDNA) markers [56,57,82] and microsatellites (cpSSRs) [83,84]. Nuclear DNA (nDNA) sequences [85] and microsatellites [19] have also been evaluated. One study [73] combined data from nuclear ITS and chloroplast *trnK* loci. A common aspect of these studies is the lack of inclusion of a wide range of accessions of *P. dumosus*, particularly of the ancestral materials.

Nevertheless the overall indication from those studies of *P. dumosus* relationships is that, according to DNA analysis, *P. dumosus* seems to be closer to *P. vulgaris* according to cpDNA, but more related to *P. coccineus* by the nDNA. In addition, experimental crosses using *P. dumosus* as cytoplasm donor have shown a higher (but not total) fertility for the cross *P. dumosus*/*P. coccineus* than for the cross *P. dumosus*/*P. vulgaris* [51,80]. Overall, interspecific crosses using the GP2 are possible and less difficult when *P. vulgaris* is used as a female donor. However, there are some

difficulties for the recovery of F1, showing little introgression, and hybrid progeny usually turns back into a parental cytoplasm donor [36,80]. It has been proposed [82,86] that a reticulation (net like evolution) may have occurred during the genesis of *P. dumosus*, involving *P. costaricensis*, and also several backcrossing events which may explain the unbalance in fertility from the interspecific crosses and the phylogenetic discordance.

Summarizing, there is no strongly supported conclusion regarding the origin and relationship of *P. dumosus*. Since analyses of the *Phaseoli* group have been mostly directed towards outlining the relationships of *P. vulgaris* or *P. coccineus*, they have usually included only a small number of individuals of *P. dumosus*. Consequently they have not addressed the possibility of population-level variability within *P. dumosus*, since interspecific crosses between different accessions of the same species may produce widely disparate results in terms of hybrid fertility, and gene introgression success. To better understand the possibilities of interspecific introgression for crop improvement using *P. dumosus*, a multi-locus population-level study is required, since the particular questions of the origin and relationship of *P. dumosus* remain unresolved. Likewise the actual levels of diversity of the germplasm representatives that are held for this species is another aspect that requires evaluation. Desiderio et al. [84] recently stressed the need for further analysis that includes wild relatives to analyse relationships and speciation processes that have occurred around the wild relatives of the common bean.

It is necessary to confirm the discrepancies among the nuclear and the chloroplast genomes by a wide analysis of wild representatives of the three species. Such analysis

should include the other close relative species such as *P. costaricensis*, *P. albescens* and *P. persistentus* of which little is known. This thesis research addresses the lack of information on *P. dumosus* genetic diversity and provides further insights into the origin of this species.

1.5 How to evaluate genetic diversity and detect ancestral hybridisation?

Phylogenetic discordance alone is not enough to validate the hybrid origin of a species, since discordance between topologies can also be explained by other evolutionary processes. In closely related species, common ancestry and the presence of ancestral polymorphism across species boundaries (incomplete lineage sorting) can also generate discordance among topologies [87–89]. Additionally, ancestral hybridisation is difficult to detect since over time processes such as genetic drift, mutation and selection generates genetic changes that are accumulated over generations. Hence the genetic signal of hybridisation is blurred and difficult to detect [90]. Likewise after a hybridisation event, processes of reticulated evolution may take place [91], therefore evidence of introgression between hybrids and putative parents, will strongly suggest a hybrid origin. In the same way the presence of some biological characteristics such as an intermediate morphology of the hybrid taxa, partial fertility, sympatry of parental species, would reinforce a hybridisation hypothesis [92,93].

Based upon these parameters, discordance among phylogenetic trees, has been used as a criterion to identify events of hybridisation. Several cases of homoploid hybrids

in plant and animal species have been documented by the analysis of nuclear and cytoplasmic DNA [88, 94–96]. The advances from using chloroplast regions for evolutionary and genetic diversity studies are well known. It is understood that in around 80% of angiosperms, chloroplast are uniparentally transmitted [97], for which cpDNA is often used to outline the direction of hybridisation. In close related species chloroplast can easily move across species boundaries by introgressive hybridisation (repeated backcrossing of the hybrids to parental species) [97,98]. In *Phaseolus* species chloroplast are of maternal heritage [99,100], thus a chloroplast-based phylogeny outlines the maternal relationships between the *Phaseolus* species.

The nuclear genome alone can reveal events of hybridisation when the genetic signal from the two parental species still persists in the hybrid. In this case, the hybrid would likely show inheritance from both, the paternal and the maternal genome, except when there has been homogenization of alleles by concerted evolution [101,102] or introgressive hybridisation [103]. To find evidence of hybridisation, the evaluation of several loci is essential, since a multi-locus analysis provides information on the species rather than a locus. It is also essential in cases where a locus or a marker present homogenization, as the hybrid signal could still be detected by incongruencies of topologies from different loci. Moreover, a multi-locus analysis has the power to discern amongst some of the evolutionary processes that have acted during speciation [88,104]. For instance, a comparison of polymorphism between closely related species denotes the variation accumulated by each species after divergence. Shared polymorphism between species gives insights of recent divergence, common ancestry and gene flow, while the loss of shared variants and the accumulation of fixed differences between species are consequences of mutational events, genetic drift and

selection [105]. The reconstruction of phylogenies with data from different loci, are useful for revealing divergent phylogenies that suggest hybridisation. However, as phylogenetic trees do not represent evolutionary processes that deviate from the assumption of homology [106,107], consensus networks and network diagrams are used to represent contradictory evidence from species phylogenies and events of hybridisation and reticulation [108,109].

1.6 Discerning the origin by hybridisation of *P. dumosus*

In this thesis research I conducted an analysis of genetic diversity, phylogenetic and genetic structure of a germplasm collection of *P. dumosus*. This analysis was performed with the aim of assessing the molecular diversity of this species as well as validating its hybrid origin. Using a germplasm representation of wild, landrace and cultivar types of *P. dumosus* and the other species in the GP2, the following research approaches were applied:

First, sequencing of multiple cpDNA and the ITS (nrDNA) loci was performed. Chapter 2 describes how sequenced traces were used for analysis of within species nucleotide variation to establish the diversity of *P. dumosus*. The nucleotide variation between species of the GP2 was used to find haplotypes, shared polymorphism and common ancestry.

Second, Chapter 3 of this thesis presents the phylogenetic reconstruction inferred from the sequences of cpDNA and ITS. Reconstructions were made by applying the Maximum likelihood and Bayesian methods. These traditional phylogenetic methods were applied to confirm the discrepancy between relationships observed by the nuclear and chloroplast genome of *P. dumosus*. Trees from individual loci were tested against concatenated matrices and different models of evolution were considered. Additionally, a phylogenetic hybridisation network was generated to identify putative events of reticulation and provide further insight into the origin of *P. dumosus*.

Third, a genome-wide screen was performed by using DArT (Diversity Array Technology) molecular markers. This is a sequence-independent and high-throughput marker technology, currently one of the most useful tool to outlines phylogenetic relationships between crops and wild relatives [110]. This technique is able to identify polymorphism along the genome and provide information from hundreds of polymorphic loci in a single array [111,112]. Addressing the lack of information on *P. dumosus* diversity, Chapter 4 of this thesis, presents the use of DArT markers to characterize a germplasm collection of *P. dumosus* and representatives of the other species in the GP2 including *P. vulgaris*. The scoring from DArT markers was used to generate an analysis of genetic structure, a fine-scale analysis by Principal Coordinates Analysis (PCoA) and an ancestry analysis (STRUCTURE) which provides a deep understanding of the relationships among the species of the GP2, The ancestry analysis allows identification of major but also minor structure among and within species. Biological interpretation of major findings, main conclusions and potential further research from this work are presented in Chapter 5.

Overall this research provides a deeper understanding of the diversity of *P. dumosus* and its relationships with the species GP2 of the common bean, as well as providing valuable data about *P. persistentus*, *P. albescens* and *P. costaricensis*.

2 Molecular diversity of *P. dumosus* assessed from sequences of ITS and non-coding chloroplast markers

2.1 Introduction

The common bean secondary gene pool (GP2) is formed by close relatives of the common bean: wild and cultivated forms of *P. dumosus*, *P. coccineus*, and *P. costaricensis* [86]. This thesis chapter is focused on evaluating the molecular diversity of wild germplasm of *P. dumosus*. Among species of the (GP2), *P. dumosus* is a key genetic resource, as described in Chapter 1. Its cultivars are grown in central and South America where it is known as the year-long bean, due to the traditional cultivation as a yearly crop by Indians in Guatemala [74]. Little is known about the diversity of wild ancestral forms of *P. dumosus*. Most of the genetic diversity of this species remains undiscovered; analysis of genetic diversity and characterisation of the wild types is necessary.

To date, many questions about the origin and relationship of *P. dumosus* remain unresolved. Notwithstanding the close link among species of the common bean GP2 is well known [61,85,86,113], the origin and direction of the relationship of *P. dumosus* with its closest relatives are still under debate [83,84]. A necessary first step towards clarifying the origin of *P. dumosus* is the evaluation of the molecular diversity of the wild forms. Molecular diversity provides information on aspects as level of polymorphism, shared haplotypes and levels of diversity [114].

The first aim of this thesis chapter was to assess the molecular diversity of the available wild germplasm accessions of *P. dumosus*, using sequences of ITS and cpDNA. As previously stated in Chapter 1 the combined information from sequences of these two

markers are particularly useful to detect process of hybridisation. At the specie level, multi-copy markers such as the ITS regions are highly useful to detect patterns of reticulation, concerted evolution, and lineage sorting [115,116]. For this research eleven markers (listed in section 2.2.3) for non-coding regions of the chloroplast DNA were evaluated among the GP2. These markers were selected because they have previously demonstrated to be effective for phylogenetic studies of *Phaseolus* [56], and informative among and within the *P. vulgaris* group and its closes relatives [57,73,85].

This is the first analysis that uses markers of both nuclear and chloroplast genomes in all available accessions of wild *P. dumosus*, whereby this research provides explanations to the pattern of diversity found among the ancestral wild *P. dumosus*.

In the course of this chapter I will be comparing the molecular diversity of *P. dumosus* with the diversity of *P. coccineus* and *P. vulgaris*. Whether the data from molecular diversity supports the hypothesis of hybridisation or not is discussed. Finally this chapter also reports data about *P. persistentus*, a species for which classification was unclear. An explanation of the classification of this species is also provided.

2.2 Materials and Methods

2.2.1 Plant material selection

A selection of 45 accessions representative of the GP2, were sampled from the germplasm bank held at the International Center for Tropical Agriculture (CIAT), Cali-Colombia. The selection included samples from the centres of origin in Central America, Mexico, Guatemala and Costa Rica. The identification of each seed accession was kept as written in the germplasm vouchers. The germplasm bank provided the seeds in envelopes that are labelled with the accession number, this number was used to identify each planted seed and each plant, photographic records were also taking as shown in Appendix A1.

A sample of the only reported accession of *P. persistentus* [51] was obtained from the Royal Botanical Gardens, Kew, United Kingdom. The plant accessions used in this chapter are listed in Table 2.1. The locations of all the samples included in this thesis are illustrated in Figure 4.1, Chapter 4.

2.2.2 Plant growing conditions

Five seeds of each accession were scarified before being placed on a layer of water-moistened cotton wool inside a beaker. Each beaker was numbered and labelled, with accession number, seed number and placed in a growing cabinet (Percival Scientific

2. Molecular diversity of *P. dumosus* assessed from sequences of ITS and non-coding chloroplast markers

Inc., Germany). The seeds were grown under the following conditions: 12-hour light cycle at 22°C and a 12-hour dark cycle at 21°C. The cotton was moistened with water every day. The first seedlings appeared at between 5 and 7 days, (Appendix I, A1 presents a photographic illustration of seeds labelled and the growing conditions). Plants were harvested at ten days post-germination for subsequent molecular analysis. As part of the germplasm exchanged agreement, plants materials were destroyed after use, no plants were preserved.

Table 2.1 Accessions of bean GP2 and close relatives obtained from CIAT germplasm bank. Accession identification, biological status and geographical references, Decimal degrees (Dd), (-) Missing information.

Species	Accession	Biological Status	Country	County	Area	Altitude (Meters)	Latitude (Dd)	Longitude (Dd)
<i>P. acutifolius</i>	G40178	Wild	Mexico	Chihuahua	Belleza	1700	26.900	-106.300
<i>P. acutifolius</i>	G40286	Wild	United State	Cochise	—	—	31.717	-109.333
<i>P. albescens</i>	PL3592	Wild	Mexico	—	—	—	—	—
<i>P. coccineus</i>	G35837A	Wild	Guatemala	Huehuetenango	Chiantla	2450	15.383	-90.833
<i>P. coccineus</i>	G36285	Wild	Guatemala	Quezaltenango	Zuñil	1800	14.750	-91.517
<i>P. coccineus</i>	G36353	Wild	Guatemala	Jalapa	Jalapa	2010	14.583	-90.083
<i>P. coccineus</i>	G36006	Wild	Guatemala	Quetzaltenango	Santa Maria De Jesus	1570	14.717	-91.517
<i>P. coccineus</i>	G36206	Wild	Guatemala	Quetzaltenango	Zuñil	2040	14.750	-91.483
<i>P. coccineus</i>	G36256	Cultivar	Ecuador	Imbabura	Cotacachi	1700	0.350	-78.467
<i>P. coccineus</i>	G35650	Wild	Mexico	Morelos	—	—	—	—
<i>P. coccineus</i>	G35757	Wild	Guatemala	Sacatepéquez	Santa Maria de Jesus	2100	14.467	-90.733
<i>P. coccineus</i>	G36002	Wild	Guatemala	Sacatepéquez	San Miguel Dueñas	2080	14.550	-90.850
<i>P. coccineus</i>	G36334	Wild	Mexico	Chihuahua	Gomez Farias	-	29.300	-107.717
<i>P. coccineus</i>	G35859	Wild	Mexico	Durango	Nombre de Dios	1830	23.950	-104.283

2. Molecular diversity of *P. dumosus* assessed from sequences of ITS and non-coding chloroplast markers

<i>P. costaricensis</i>	G40805	Wild	Costa Rica	San Jose	Alajuelita	1620	9.867	-84.117
<i>P. costaricensis</i>	G40604	Wild	Costa Rica	Cartago	La Union	1750	9.950	-83.950
<i>P. dumosus</i>	G35946	Wild	Guatemala	Sacatepéquez	San Miguel Dueñas	1730	14.550	-90.850
<i>P. dumosus</i>	G36127	Cultivar	Peru	Cajamarca	Chota	2430	-6.450	-78.633
<i>P. dumosus</i>	G35641	Cultivar	Mexico	Puebla	Tetela de Ocampo	1730	14.550	-90.850
<i>P. dumosus</i>	G35758	Wild	Guatemala	Sacatepéquez	Santa Maria De Jesus	1300	19.300	-100.417
<i>P. dumosus</i>	G35877	Wild	Guatemala	Sololá	Panajachel	1680	14.767	-91.167
<i>P. dumosus</i>	G35908	Wild	Guatemala	Sacatepéquez	San Miguel Dueñas	1940	14.550	-90.833
<i>P. dumosus</i>	G36045	Wild	Guatemala	Guatemala	Amatitlan	1800	14.433	-90.583
<i>P. dumosus</i>	G36286	Wild	Guatemala	Quetzaltenango	Santa Maria De Jesus	1820	14.550	-90.833
<i>P. dumosus</i>	G35758	Wild	Guatemala	Sacatepéquez	Santa Maria De Jesus	1300	19.300	-100.417
<i>P. leptostachyus</i>	G40553	Wild	Mexico	Oaxaca	Oaxaca	1750	17.333	-96.883
<i>P. lunatus</i>	G25224	Wild	Costa Rica	Guanacaste	Cañas	100	10.767	-85.483
<i>P. lunatus</i>	G25226	Wild	Costa Rica	Guanacaste	Cañas	100	10.767	-85.483
<i>P. persistentus</i>	DQ4457	Wild	Guatemala	Sacatepéquez	San Miguel Dueñas	—	14.330	-90.500
<i>P. vulgaris</i>	G50387	Wild	Guatemala	Quetzaltenango	Zuñil	1730	14.750	-91.517
<i>P. vulgaris</i>	G7479	Wild	Mexico	Guerrero	Teloloapan	1150	18.317	-99.817
<i>P. vulgaris</i>	G50388B	Wild	Guatemala	Quetzaltenango	Zuñil	1760	14.750	-91.517
<i>P. vulgaris</i>	G23463	Wild	Mexico	Chihuahua	Yepachic	1530	28.333	-108.500

2. Molecular diversity of *P. dumosus* assessed from sequences of ITS and non-coding chloroplast markers

<i>P. vulgaris</i>	G12853	Wild	Guatemala	Jutiapa	Quesada	873	14.283	-89.900
<i>P. vulgaris</i>	G19908	Wild	Guatemala	Sacatepéquez	San Miguel Dueñas	1820	14.550	-90.833
<i>P. vulgaris</i>	G19909	Wild	Guatemala	Sacatepéquez	San Miguel Dueñas	1820	14.550	-90.833
<i>P. vulgaris</i>	G21244	Wild	Peru	Cajamarca	San Pablo	2020	-7.183	-78.833
<i>P. vulgaris</i>	G24358	Wild	Mexico	Michoacán	Benito Juarez	1300	19.300	-100.417

2.2.3 Molecular analysis: sequencing of the cpDNA and ITS loci.

Genomic DNA was isolated from five plants of each accession of the samples. Extractions were made from single plants following the protocol of Afanador et al. [117]. Initially, a polymerase chain reaction (PCR)-test with a subset of fourteen samples (three samples for each species) was carried out. This test was to select the cpDNA markers to use for sequencing. Primers for eleven regions of the cpDNA were tested, those regions were: *trnL-trnF* spacer, *matK-trnK*, *rpoC1 - rpoC2* spacer, *petA-psbE* spacer, *rpL16* intron, *atp β -rbcL* spacer, *ndhA* intron, *accD-psaI* spacer, *trnT-trnL* spacer, *trnL* intron and *rps14-psaB* spacer. From these eleven markers, the latter four were positive for amplification among all species and therefore were sequenced in the full sample set. Details of the primers set used for amplification of each region are given in Table 2.2.

Punctual PCR amplifications were made targeting the full ITS region and for each of the four non-coding cpDNA loci. All PCR reactions were carried out using GoTaq® Flexi DNA polymerase (Promega, WI) enzyme, following the manufacturer's conditions for the master mix. Reactions were performed on a Dyad Disciple™ Peltier Thermal Cycler (MJ Research). Cycling conditions consisted of an initial denaturation of 2 minutes at 94°C, followed by 35 cycles of: denaturation at 94°C for 20 seconds, hybridisation of 1 minute at 57°C ($\pm 5^\circ\text{C}$), extension of 1 minute per kb at 72°C, followed by final extension phase of 10 minutes at 72°C. The PCR-derived fragments were resolved on 1.5% agarose/TBE gels and visualized, under UV light

using 1x SYBR® Safe DNA gel stain (Invitrogen, UK). A positive control DNA and negative water blank were used as references. PCR products were purified using the QIAquick PCR purification kit (QIAGEN, Germany) according to the manufacturer's instructions and eluted into 30 µl autoclaved Millipore water.

The PCR products of two independent samples of each accession were sequenced in both directions, forward and reverse; by GATC Biotech (Germany) using single read 11x ABI 3730xl sequencing technology. It was intended to obtain complete sequences of two samples per accession. Where sequencing reactions failed, they were repeated and when necessary a new sample (from the five harvested) was used. In recurrent failures new DNA extraction or new PCR product was used. Sequences that failed after four attempts were eliminated from the analysis.

2.2.4 Data Analysis

2.2.4.1 Sequence alignments

Forward and reverse sequences were assembled into contigs using the ContigExpress tool in Vector NTI I® V.11.5 (Life Technology, US). Sequence errors due to poor quality data were identified based on the quality scores from the ABI. As the software allowed the option of keeping associations between the score from the ABI chromatogram and the sequences in the contig, poor quality sequences were removed. The resulting consensus sequence was exported in Fasta file format. The homology of the sequences was checked using GenBank BLASTn algorithm

(<http://www.ncbi.nlm.nih.gov/BLAST>) against the nucleotide databases of NCBI. Alignments for each locus were constructed with the program Muscle [118] and manually verified and corrected with the editor Jalview 2.7 [119].

2.2.4.2 Statistical Analysis

Statistics for genetic diversity and genetic similarity were determined for each of the four loci of the cpDNA and for the ITS sequence independently. Analysis was made by the calculation of estimators of molecular diversity and genetic divergences described in the following section. For analysis, the sequences of *P. dumosus* were divided in two groups: wild and cultivated. Most estimates were made between and within wild accessions of the species, *P. dumosus*, *P. vulgaris* and *P. coccineus* as these are the focus of this study. Where possible the estimates included the full set of species and accessions mentioned earlier.

The statistical analysis was conducted to:

1. Evaluate the molecular polymorphisms found in wild *P. dumosus* and draw information about the diversity of this species.
2. Compare the molecular variation observed in wild *P. dumosus* with the variation observed in wild types of *P. coccineus* and *P. vulgaris*.

2.2.4.2.1 Description of genetic diversity estimators used to assess molecular diversity

Genetic variation in non-coding regions of DNA can be studied only by examining the DNA sequences. This provides information about polymorphisms caused by nucleotide substitutions, insertion/deletions (INDELs) and other events. The DNA polymorphism of the cpDNA and ITS were estimated by measures of genetic variability, including: Haplotype diversity, Nei's Nucleotide diversity [120], number of segregating sites (S), haplotypes junction and haplotypes divergences. All these measures of genetic variability were used to provide information about the level of polymorphism between bean species and within accessions of each species. Analyses were made using the program DnaSP v5 [121], this is a widely used program for analysis of DNA polymorphism from nucleotide sequence data. The **total number of segregating sites (S)** is established by summarizing all nucleotide sites that show two or more different nucleotides states within the population [122]. This measure indicates the number of nucleotide sites that differ among the aligned sequences of beans.

The **Nucleotide diversity (π)** summarises nucleotide polymorphism by averaging the number of nucleotide differences per site between two sequences. This measure provides information about the number of nucleotide substitutions between the haplotypes and is a measure of how variable the bean species are at the evaluated locus. π does not depend on sample size and is defined as:

$$\pi = \sum_{i=j}^m \frac{d_{ij}}{c}$$

Where m is the number of DNA sequences, i and j refer to the i^{th} and j^{th} sequences and c is the total number of sequence comparisons [122]. The significance of π between *P. dumosus*, *P. vulgaris* and *P. coccineus* was tested by t-test.

The **Haplotype (gene) diversity (h)** describes the number and frequency of shared haplotypes and whether these haplotypes can be grouped by species and/or location. For example, it will indicate how many haplotypes are shared within accessions of *P. dumosus*. h is a measure of the heterozygosity at a single locus. It was calculated by:

$$h = n(1 - \sum f_i^2) / (n-1)$$

Where f is the frequency of the i^{th} haplotype, and n is the number of samples. The equation was modified from the original, replacing $2n$ with n [122]. The estimate does not include gaps or missing data.

The significance of the molecular structure between species (variation between species i.e. *P. dumosus*, *P. coccineus*, *P. vulgaris*) and within species (variation between the accessions of a species) was tested using analysis of molecular variance (AMOVA). The AMOVA was calculated using Arlequin 3.5 [123]. The test in this program is based on fixation indices (F_{ST} -statistic) calculated according to Wright [124]. The significance was tested using non-parametric permutational testing procedures [125]. The test was made for 1000 permutations, and a level of significance of 0.05. The

AMOVA computation considers the relative frequency of haplotypes and the number of mutations between pairs of haplotypes.

2.2.4.2.2 Pairwise differences: measure of sequence divergences within the secondary gene pool of the common bean

Any two sequences derived from a common ancestor that have evolved independently of each other eventually diverge [126]. The genetic distance is a measurement of this divergence, based on the number of allelic substitutions per locus, which have occurred during the separate evolution of two populations. It is also a measure of the relatedness of taxa, (whereas genetic diversity represents diversity within a population). The divergences per locus between individuals of the six species of beans were measured by: pairwise genetic distance (as distance matrix), the average absolute divergence (D_{XY}) and the relative average divergence (D_A).

The D_A is given by $d_A = d_{XY} - (d_X + d_Y)/2$ where d_{XY} is the average distance between groups X and Y, and d_X and d_Y are the mean within-group distances. Calculations were made for the ITS sequences and (only for this analysis) a concatenated matrix of the four cpDNA loci. Graphical representation of the divergences were made in R (<http://www.R-project.org>). Further analysis of relationships under different inferences and evolutionary models are presented in Chapter 3.

2.3 Results

2.3.1 Overall information from the ITS and cpDNA sequences

Sequences of nrDNA-ITS: The quality of the majority of sequences was optimal and traces were clean, trimming was required only at the ends of the sequences. After trimming the ITS alignment averaged 610bp, approximately 250bp correspond to the ITS1, the 5.8S rRNA was around 168bp and the ITS2 was around 190bp. Two INDELs were observed within the sequences of *P. coccineus*, and no INDELs were observed within the sequences of *P. dumosus* or *P. vulgaris* (Table 2.2). The ITS sequences from *P. persistentus* and *P. leptostachyus* were removed from final alignments due to poor quality of the sequence traces. These sequences were replaced with sequences available in GenBank: *P. persistentus* accession DQ445768.1 [127] and *P. leptostachyus* accessions EU909001.1 and EU909005.1 [128].

Sequences of cpDNA: Alignments of the cpDNA loci varied in size, the largest retrieved (750bp) was the *trnT-trnL* spacer. The *trnL* intron was the shortest sequence with 480bp. Among the four loci, six INDELs were found within *P. coccineus* accessions and one INDELs within for *P. vulgaris* accessions (Table 2.2)

2. Molecular diversity of *P. dumosus* (year-long bean) assessed from sequences of ITS and non-coding chloroplast markers.

Table 2.2 General information of the sequences of the loci: ITS, *trnL* spacer, *trnL* intron, *rps14-psaB* spacer and *accD-psaI* spacer. Sequence sizes are after edition and are reported in base pairs (bp). Primers set used to amplify each region and primers reference are reported. INDELS present on the sequence of each locus are reported indicating the species were it was found.

Region-(size bp)	Primers name-sequence (5'-3')	Reference	INDELS (number of nucleotides)		
			<i>P.dumosus</i>	<i>P.coccineus</i>	<i>P.vulgaris</i>
<i>ITS-1</i> (250) <i>ITS-2</i> (190)	a1_GTCGTAACAAGGTTTCCGTAGG b2_GAGGTCTCGTACATGGCACA	This study	–	2	–
<i>trnT-trnL</i> spacer (750)	c3_CAT TAC AAA TGC GAT GCT CT d4_TCT ACC GAT TTC GCC ATA TC	Taberlet et al. (1991)	–	4	1
<i>trnL</i> intron (480)	e5_CGA AAT CGG TAG ACG CTA CG f6_GGG GAT AGA GGG ACT TGA AC	Taberlet et al. (1991)	–	–	–
<i>rps14-psaB</i> spacer (550)	g7_CAT TTC ACG AAG TAT GTG TCC G h8_TGG CGT GGA TAT TGG CAG GA	Fofana et al. (1997)	–	1	–
<i>accD-psaI</i> spacer (600)	i9_GGA AGT TTG AGC TTT ATG CAA ATG G j10_AGA AGC CAT TGC AAT TGC CGG AAA	Small et al. (1998)	–	1	–

2.3.2 What is the molecular diversity of wild *P. dumosus* based on nuclear and chloroplast DNA polymorphism?

2.3.2.1 Diversity from nrDNA sequences

Low polymorphism was found in the sequences of wild *P. dumosus*. There was only one segregating site ($S=1$), hence only two haplotypes and a low nucleotide diversity ($\pi=0.00035$), Table 2.3. Six of the seven wild accessions presented the same haplotype (**i9**). The second haplotype (**i10**) was present only in the accession G35877 from Sololá-Guatemala. This data was re-confirmed by manually checking the alignments against the sequence trace. It was confirmed that among the *P. dumosus* sequences, there were no individual variation or sequence conflict in which a nucleotides resembled *P. vulgaris*.

There were six haplotypes among the sequences of wild *P. coccineus*, and four haplotypes among the sequences of *P. vulgaris*. *P. coccineus* was the most polymorphic between species of the GP2, $S=14$ and $\pi=0.00903$ (Table 2.3). The t-test comparisons for the average of nucleotide diversity were highly significant ($P < 0.0001$), meaning that there is a significant difference between the rates of polymorphism presented in the ITS regions of each species.

Complementary to the nucleotide diversity, the AMOVA analysis for the molecular structure did not find significant variation between the wild accessions of *P. dumosus*. Similar results occurred between the accessions of wild *P. vulgaris*. Instead molecular

variation between wild accessions of *P. coccineus* was significant ($P=0.0127$), 56% of the total variation was due to differences between accessions, and the remaining 43% of the variation corresponded to differences between the country of origin (within groups). This indicates that the variation of wild *P. coccineus* is higher within the countries (Guatemala and Mexico) than between them. In AMOVA testing an F_{ST} value ≥ 0.25 indicates a large genetic differentiation; with an $F_{ST}=0.43519$, $P=0.0127$ (Appendix I, A4), the molecular structure of *P. coccineus* at the accessions level is considerable strong.

AMOVA, made to compare the three species (Appendix I, A4), shows that 91% of the variation is attributed to differences between species ($F_{CT}=0.90653$) and the differences between accessions within species were small ($<5\%$) (This is not evidence of structural separations between accessions from different countries but rather individual variation). The variation within accessions from the same country were around 5% when *P. coccineus* was included and reduced to $<1\%$ when *P. coccineus* was excluded.

These results indicate that, based on the ITS sequences, the differences between the nucleotide diversity among the three species, is mainly due to the diversity displayed by wild *P. coccineus*. Also, it is statistically higher than the diversity displayed by the accessions of wild *P. vulgaris* and wild *P. dumosus*, the latter being the least diverse of the three species.

Table 2.3 Genetic diversity estimates, for sequences of ITS and four cpDNA of *P. dumosus*, *P. coccineus* and *P. vulgaris*. Reporting the number of segregating sites (S), average nucleotide diversity (π) and its standard deviation (sd). And haplotypes found for each locus. In parenthesis is the number of sequences include for each species.

<i>Locus/ species</i>	<i>S*</i>	<i>Diversity (π) sd</i>	<i>Haplotypes</i>
ITS			
<i>P. dumosus</i> (10)	1	0.00035 $\pm$ 0.00027	i9, i10
<i>P. coccineus</i> (15)	14	0.00903 $\pm$ 0.00085	i11, i12, i13, i14, i15, i16
<i>P. vulgaris</i> (14)	4	0.00224 $\pm$ 0.00049	i4, i5, i6, i7
<i>trnT-trnL</i> spacer			
<i>P. dumosus</i> (8)	0	0	ts8
<i>P. coccineus</i> (15)	14	0.00635 $\pm$ 0.00166	ts2, ts4, ts5, ts6, ts14
<i>P. vulgaris</i> (8)	3	0.00134 $\pm$ 0.00036	ts9, ts10, ts11, ts12
<i>trnL</i> intron			
<i>P. dumosus wild</i> (10)	4	0.00135 $\pm$ 0.00111	ti4, ti6
<i>P. coccineus</i> (15)	4	0.00230 $\pm$ 0.00058	ti3, ti5, ti7, ti8
<i>P. vulgaris</i> (15)	3	0.00190 $\pm$ 0.00037	ti10
<i>rps14-psaB</i> spacer			
<i>P. dumosus wild</i> (11)	0	0	rs7
<i>P. coccineus</i> (19)	2	0.0008 $\pm$ 0.00045	rs6, rs7
<i>P. vulgaris</i> (15)	2	0.00116 $\pm$ 0.00036	rs9, rs10, rs7
<i>accD-psaI</i> spacer			
<i>P. dumosus wild</i> (10)	0	0	as3
<i>P. coccineus</i> (16)	9	0.00581 $\pm$ 0.00392	as3, as9, as10, as11, as12
<i>P. vulgaris</i> (11)	2	0.00115 $\pm$ 0.00034	as6, as7

*S represent the number of nucleotide sites that differ among the aligned sequences, the number of segregating sites are not necessarily equivalent to the number of haplotypes since an haplotype can be identified by the present of one or more than one segregating site.

2.3.2.2 Diversity from cpDNA sequences

Practically, there was no polymorphism among accessions of wild *P. dumosus*; segregating sites were presented only in the locus *trnL* intron (Table 2.3). The variation in this locus generates two haplotypes; one that includes nine of the ten accessions and the second one present only in the accession G35758 from Sacatepéquez-Guatemala.

Sequences of the other three cpDNA loci (*trnT-trnL* spacer, *rps14-psaB* spacer and *accD-psaI* spacer) did not present polymorphism among wild *P. dumosus*. The lack of polymorphism in the other loci did not allow further statistical comparisons of wild *P. dumosus*. However to clearly demonstrate the differences between the species molecular variation, an AMOVA was carried out for the *trnL* intron (Appendix I, A4). As expected most of the molecular variation at this locus was due to differences between species (78%). The variation between accessions of the same country was 22%, this variation mainly correspond to the variation of the accessions of *P. vulgaris* and *P. coccineus*.

In an extended analysis, a concatenated matrix of the four cpDNA loci was used for comparisons among the species (Appendix I, A3). It was confirmed that across the four cpDNA loci the molecular variation at intraspecific levels was low. Most (76%) of the molecular variation was due to differences between species and the differences were statistically supported ($F_{CT}=0.76396$, $P=0.01075$). The variation due to differences within accessions of the same country was 22%, it was statistically supported ($F_{ST}=0.78149$, $P=0.0000$). In this case the within accessions variation,

accounts for the differences between accessions of *P. vulgaris* and *P. coccineus* since as stated above, practically there was no variation among accessions of *P. dumosus*.

Using the concatenated matrix the differences between accessions of *P. coccineus* were tested and, despite being statistically significant, the differences were very low (5.5%) and therefore meaningless in terms of structural differentiation. It is statistically incorrect to consider this percentage of variation as explanatory of genetic structure at the accession level. However, the statistical significance of this low percentage could be an indication of recent splits that are still important because they generate a statistical difference but have not had enough time to accumulate considerable variation. Nevertheless due to the number of accessions and samples included here not being representative of the extended variation of wild *P. vulgaris* and wild *P. coccineus* (Chapter 1) and as the main focus of this research is wild *P. dumosus*, we are cautious about making further interpretations.

A summary of the genetic diversity analysis between all species and accessions is presented in Figure 2.1. In this the nucleotide diversity π (orange in diagonal) shows the levels of diversity found through the accessions of each species. The white area across wild *P. dumosus* (dumG, in the figure) clearly illustrates the lack of variation of the accessions. For the ITS (Figure 2.1A), only *P. coccineus* shows visible variation across accessions. On the other hand for the cpDNA (Figure 2.1B) accession variation is seen in *P. vulgaris* as well as in *P. coccineus*.

Figure 2.1A ITS

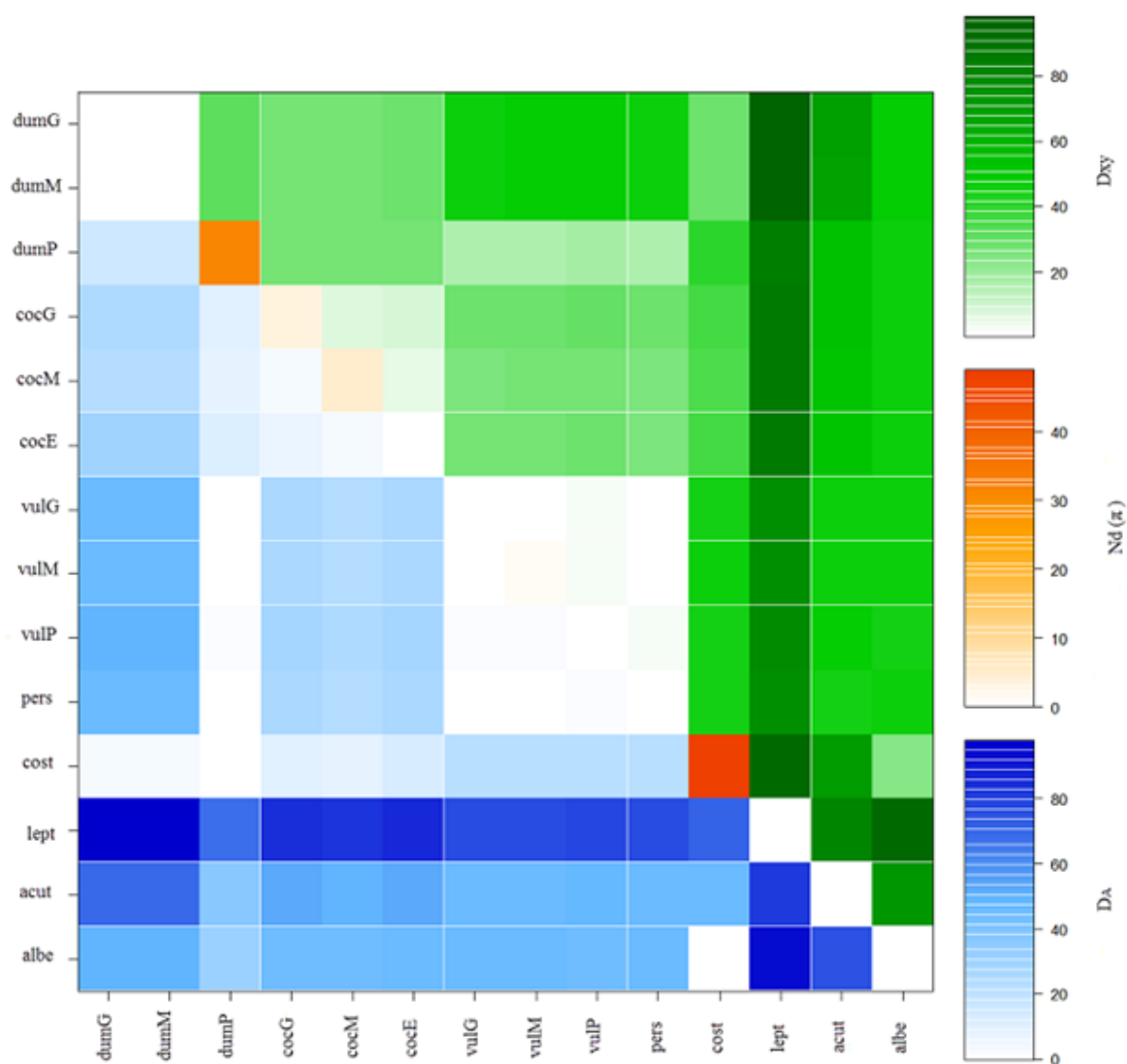


Figure 2.1B cpDNA

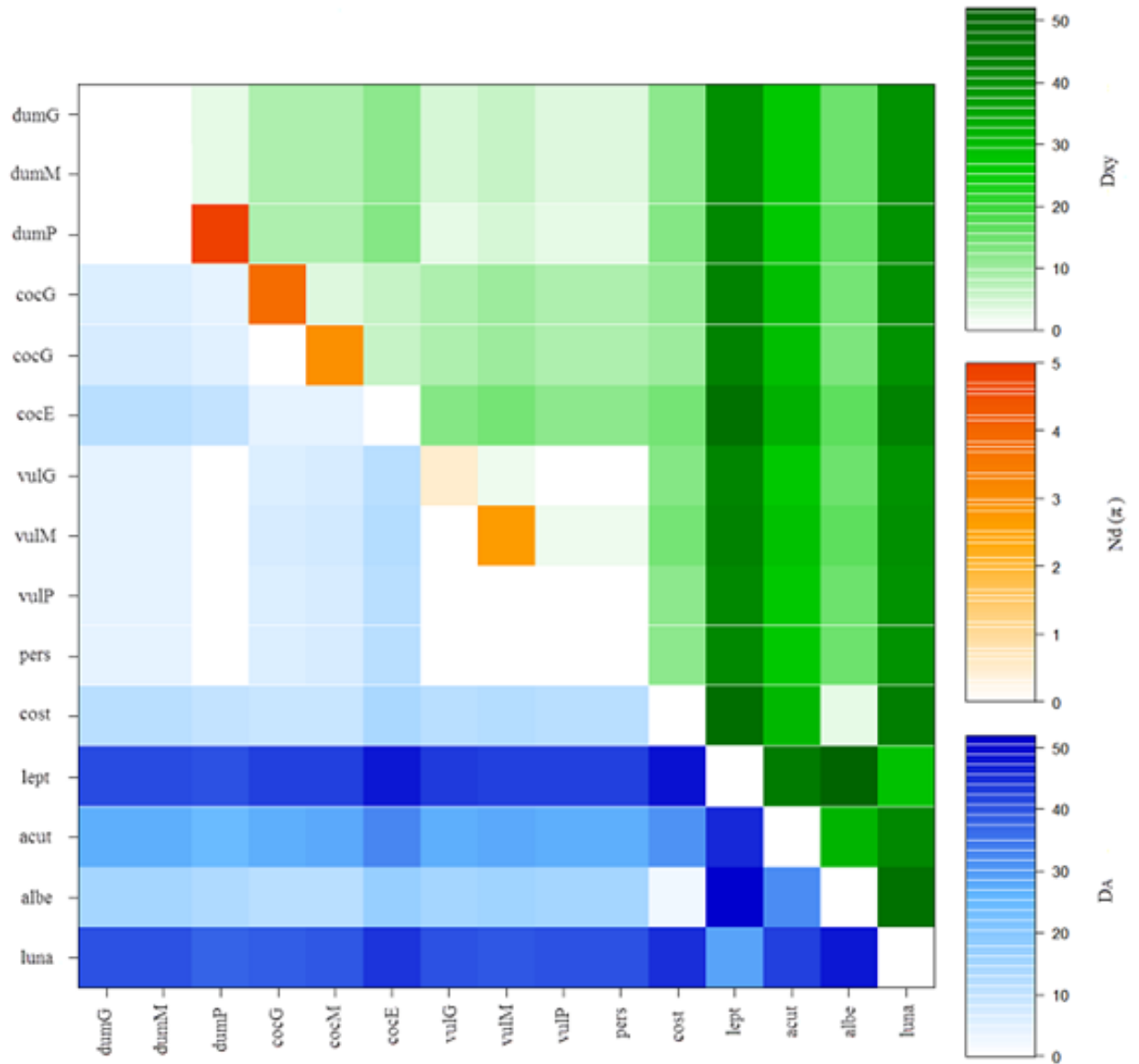


Figure 2.1 Summary of pairwise differences of ITS sequences (A) and cpDNA sequences (B). Showing: π nucleotide diversity (Nd) orange in diagonal, nucleotide differences between accessions (Dxy) green above diagonal and Nei's genetic distance between accessions (DA) blue below the diagonal. Information is given for: accessions of wild *P. dumosus* from Guatemala (dumG), Mexico (dumM), and cultivated *P. dumosus* from Peru (dumP). Accessions of wild *P. coccineus* from Guatemala (cocG), Mexico (cocM) and cultivar *P. coccineus* from Ecuador (cocE). Accessions of wild *P. vulgaris* from Guatemala (vulG), Mexico (vulM), Peru (vulP). *P. persistentus* (pers). *P. costaricensis* (cost). *P. leptostachyus* (lept). *P. acutifolius* (acut), *P. albescens* (albe) and *P. lunatus* (luna).

2.3.3 What are the similarities and divergences between *P. dumosus* and the other species in the bean secondary gene pool?

Similarities and divergences were evaluated by looking at shared haplotypes and nucleotide divergences between species. A phylogenetic assessment of genetic distances and relationships between species is presented in Chapter 3.

Nuclear DNA: The majority of haplotypes at the ITS were species specific (Table 2.3). There were no shared haplotypes between wild accessions of *P. dumosus* and the other species, although the molecular divergences between the haplotypes of wild *P. dumosus*, *P. coccineus* and *P. vulgaris* were low (Figure 2.2a). Among all species used in this study, only three of the ITS haplotypes were shared between two or more species: the **i7** haplotype was present in *P. vulgaris* and *P. persistentus*, the **i5** presented in *P. vulgaris* and in the cultivar accessions of *P. dumosus*, the **i3** presented in *P. albescens* and *P. costaricensis*. The haplotypes found among all species are listed in Appendix I, A2. The haplotypes with the most divergent sequences were the **i1**, **i2** and the **i3** presented in *P. leptostachyus*, *P. acutifolius* and *P. albescens*, respectively. Evaluation of ITS1 and ITS2 separately (data not shown) did not show different results of haplotypes conformations.

Results from nucleotide divergences are presented in Figure 2.1a. Dxy (green above the diagonal) showed less divergences when wild *P. dumosus* (dumG) was compared to wild *P. coccineus* (cocG, cocM) than when *P. dumosus* was compared to wild *P. vulgaris* (vulG,vulM,vulP). There were also small divergences between, *P. dumosus*

and *P. costaricensis*. There were large (2%) divergences between the *P. dumosus* wild and the cultivar form from Peru (dumP), this divergence was greater than the divergences between cultivar *P. dumosus* and *P. vulgaris*. The results from P-distances (Table 2.4) show divergences between wild *P. dumosus* and wild *P. coccineus* as 3% and divergences between wild *P. dumosus* and *P. vulgaris* were 5.3%. Wild *P. dumosus* differed by 2.7% from its cultivar form. Divergences between *P. dumosus* and *P. costaricensis* were 3.4%.

The average difference (D_A) (Figure 2.1a) shows that overall species divergences between *P. dumosus*, *P. vulgaris* and *P. coccineus* were almost at the same level (2.7%), this indicates that the ITS region of the three species is highly similar in nucleotide composition. Of all species included in the analysis *P. leptostachyus* was the most divergent. The overall divergences for all species were 4%.

Figure 2.2.A

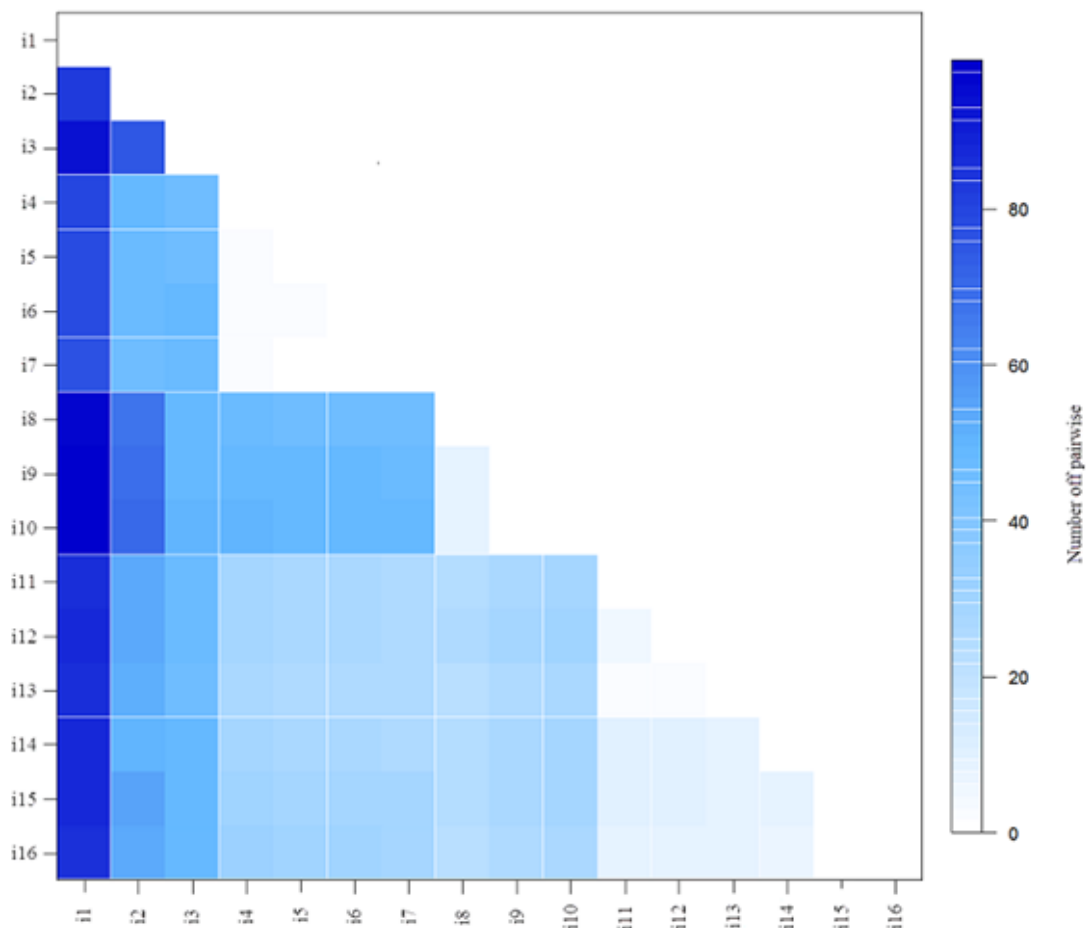


Figure 2.2B

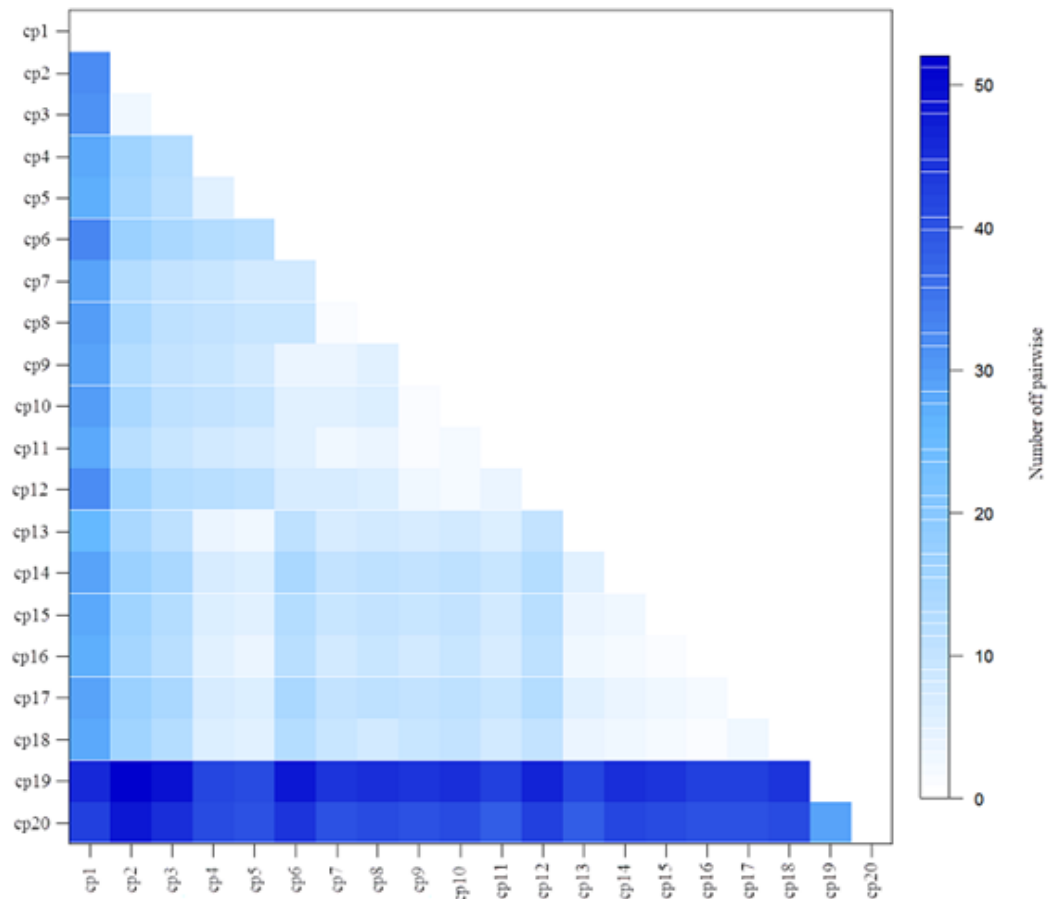


Figure 2.2 Divergences between haplotypes observed at the ITS sequences (A) and haplotypes observed at the cpDNA sequences (B).

Table 2.4 Genetic Distances (P-distance) calculated for sequences of ITS and cpDNA and ITS loci of the species of the bean GP2 and some close relatives. The overall genetic distance and its standard deviation are presented.

Species group	ITS	cpDNA
<i>P. dumosus</i> wild - <i>P. vulgaris</i>	0.053	0.003
<i>P. dumosus</i> wild - <i>P. coccineus</i>	0.034	0.006
<i>P. dumosus</i> wild - <i>P. costaricensis</i>	0.034	0.005
<i>P. dumosus</i> wild - <i>P. albescens</i>	0.059	0.010
<i>P. dumosus</i> wild - <i>P. acutifolius</i>	0.085	0.019
<i>P. dumosus</i> wild - <i>P. dumosus</i> cultivar	0.027	0.001
<i>P. coccineus</i> wild - <i>P. vulgaris</i>	0.034	0.006
<i>P. coccineus</i> wild - <i>P. costaricensis</i>	0.040	0.006
<i>P. coccineus</i> wild - <i>P. albescens</i>	0.051	0.009
<i>P. coccineus</i> wild - <i>P. acutifolius</i>	0.076	0.020
<i>P. coccineus</i> wild - <i>P. dumosus</i> cultivar	0.031	0.006
<i>P. vulgaris</i> - <i>P. costaricensis</i>	0.053	0.006
<i>P. vulgaris</i> - <i>P. albescens</i>	0.053	0.011
<i>P. vulgaris</i> - <i>P. acutifolius</i>	0.077	0.019
<i>P. vulgaris</i> - <i>P. dumosus</i> cultivar	0.027	0.003
<i>P. vulgaris</i> - <i>P. leptostachyus</i>	0.101	0.030
<i>P. persistentus</i> - <i>P. dumosus</i> wild	0.052	0.003
<i>P. persistentus</i> - <i>P. coccineus</i>	0.032	0.006
<i>P. persistentus</i> - <i>P. vulgaris</i>	0.001	0.001
<i>P. persistentus</i> - <i>P. acutifolius</i>	0.076	0.019
<i>P. persistentus</i> - <i>P. leptostachyus</i>	0.100	0.030
Overall genetic divergences	0.038	0.008
Standard deviation	0.004	0.001

Chloroplast DNA: Initially haplotypes were counted for each locus, in order to track any association between loci and shared haplotypes between species. Subsequent analyses of divergences were made by using the concatenated matrix, as regions of the chloroplast are all part of a single chromosome. The outline of shared haplotypes from the cpDNA is illustrated in Figure 2.3. The haplotypes were set for each locus and only haplotypes found in wild accessions were included in the figure the inclusion of *P. persistentus* in Figure 2.3 is discussed later.

Haplotypes found in individual loci: The cpDNA of *P. dumosus* presented five haplotypes; *P. coccineus* presented eighteen haplotypes and *P. vulgaris* presented twelve haplotypes. The fewer number of haplotypes in *P. dumosus* was expected since as seen in the previous diversity analysis, the level of polymorphism of this species was lower than the polymorphism of *P. coccineus* and *P. vulgaris*. Two private or unique haplotypes were found in the *trnL* intron of *P. dumosus*. The haplotype **ti4** displayed in all accessions except for the accession G35758 (from Sacatepéquez-Guatemala) which presented the second haplotype, **ti6**. A third unique haplotype was found in the *trnT-trnL* spacer, haplotype **ts8**, and it is found in all accessions *P. dumosus*.

The junctions involving *P. dumosus* show that this species shares two haplotypes. One of the shared haplotypes is found in the locus *rps14-psaB* spacer, haplotype **rs7**. This haplotype is shared only by the species of the GP2 and by *P. persistentus*. The results indicate that *rps14-psaB* spacer is a locus highly conserved through the species of the GP2. The **rs7** haplotype is not only present in all four species; it is also present in most

of accessions across the species. Only the accessions G24358, G12853-*P. vulgaris* and G35650-*P. coccineus*, have different haplotypes for this locus; those accessions present the haplotype rs9, rs10 and rs6 respectively. The haplotype rs7 is also the only haplotype shared between *P. coccineus* and *P. vulgaris*. The other shared haplotype was found in the locus *accD-psaI* spacer, haplotype as3. All accessions of *P. dumosus* presented this haplotype, as well as accessions G36206 and G35837A from *P. coccineus* and the accession G40604-*P. costaricensis*. Using the program Network 4.5163 [129] Median-joining networks were draw to represent the frequency of haplotypes for each loci Figure 2.4.

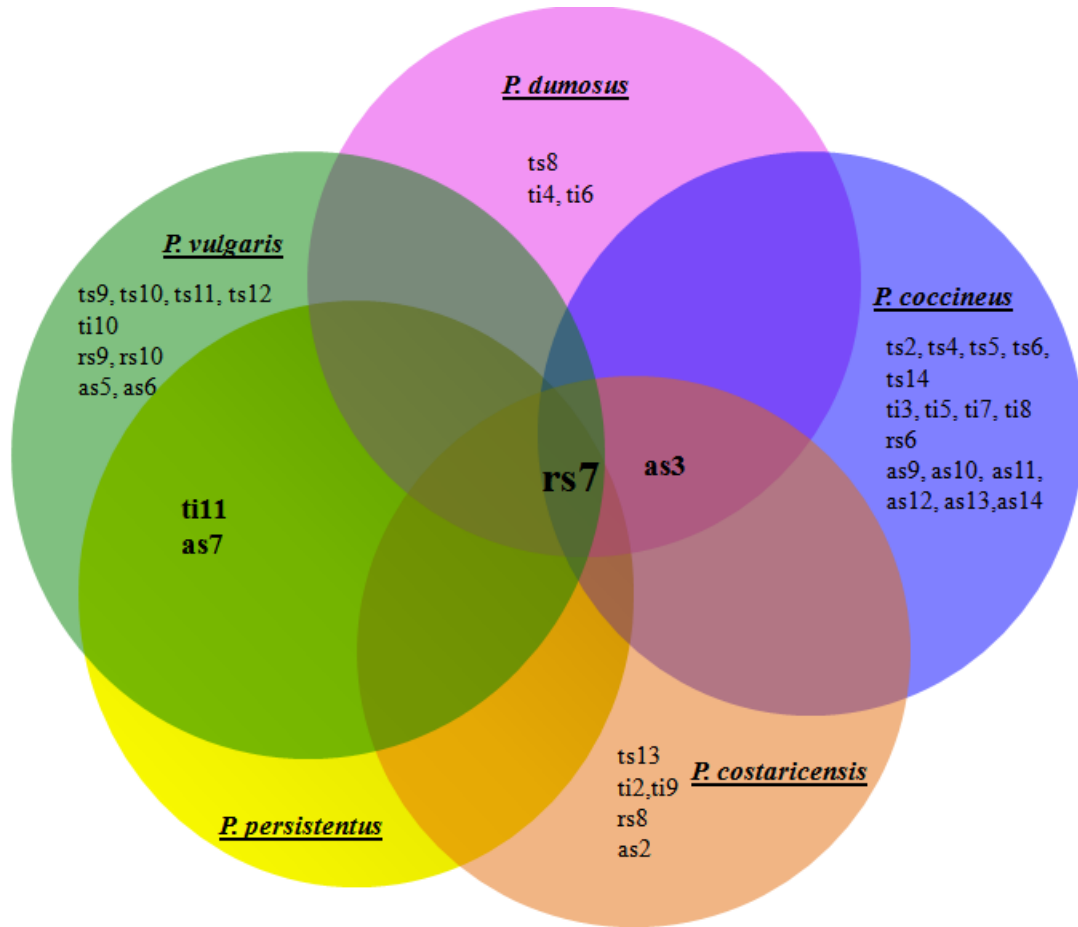


Figure 2.3 cpDNA junctions of common haplotypes between bean species. The letters of each haplotype indicates the loci to which it belongs: ts= *trnT-trnL* spacer, ti= *trnL* intron, rs=*rps14-psaB* spacer, as= *accD-psaI* spacer. Haplotype rs7 is present in all species, as3 is present in *P. dumosus*, *P. coccineus*, and *P. costaricensis*. All species except *P. persistentus* present unique haplotypes.

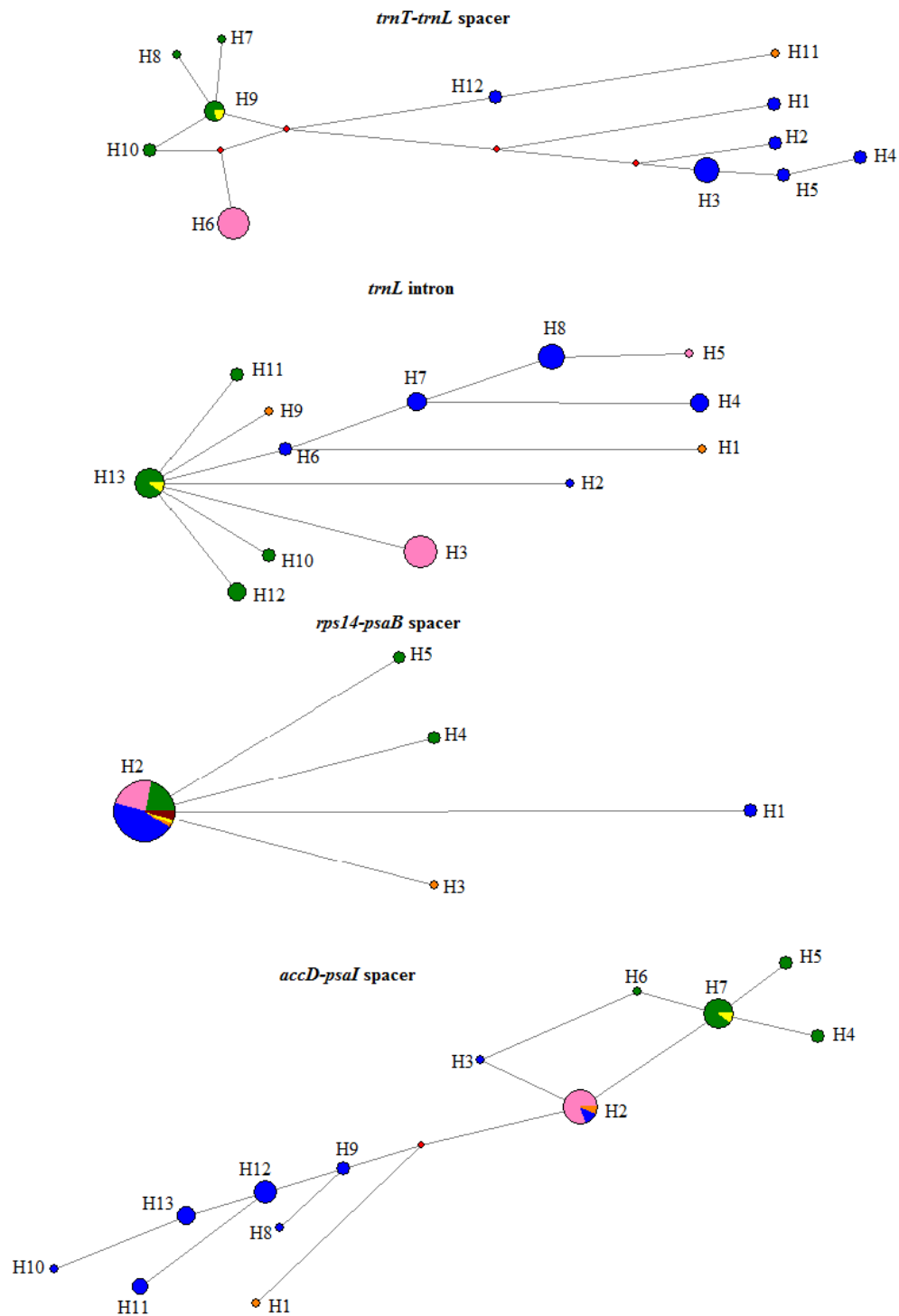


Figure 2.4 Median-joining networks representing cpDNA haplotype frequencies. Each pie represents one haplotype, the area of the pie is proportional to haplotype frequency within the wild samples of the GP2. Pink=*P. dumosus*, blue=*P. coccineus*, green=*P. vulgaris*, orange=*P. costaricensis*, yellow=*P. persistentus*.

Chloroplast DNA divergences: overall low divergences were detected among the cpDNA sequences (Figure 2.2b). Divergences were calculated using the concatenated matrix of the cpDNA loci. In total 20 haplotypes were identified from the matrix (Appendix I, A3). Between those, the most diverged haplotypes were the **cp19** presented in *P. leptostachyus* and **cp20** presented in *P. lunatus*.

Contrary to what was observed in the ITS results, nucleotide divergences D_{xy} of cpDNA showed less divergence between *P. dumosus* and *P. vulgaris* than between *P. dumosus* and *P. coccineus* (Figure 2.1b). There was a slightly greater difference between wild *P. dumosus* and *P. vulgaris* from Guatemala than between wild *P. dumosus* and *P. vulgaris* from Mexico.

Pairwise divergences were 0.03% for *P. dumosus*–*P. vulgaris* and 0.6% for *P. dumosus*–*P. coccineus*. Divergences between *P. dumosus*–*P. costaricensis* were also low (0.5%). Interestingly the divergences between *P. coccineus*–*P. vulgaris* were at the same level (0.6%) as divergences between *P. coccineus*–*P. costaricensis* and *P. vulgaris*–*P. costaricensis* (Table 2.4). Almost no divergences were observed between *P. vulgaris* and the cultivar *P. dumosus* from Peru.

The D_A for cpDNA sequences shows that there are very low divergences among the species of the GP2 in general; the average genetic divergence for all species was 0.8%. This indicates that, based on the cpDNA, the bean GP2 diverges little.

2.4 Discussion

2.4.1 Wild *P. dumosus* displayed a low level of genetic diversity.

This research investigated the diversity of several wild accessions of *P. dumosus*, through multi-locus markers in both nuclear and chloroplast genomes. The results in this research Chapter indicate that nucleotide diversity for ITS and cpDNA regions in wild *P. dumosus* is very low and is the lowest among the species of the bean GP2 (Table 2.3). The fact that wild forms of close relatives to *P. dumosus* do display nucleotide polymorphism for the same locus (*P. coccineus* was highly diverse and *P. vulgaris* was fairly diverse) indicates that the results are not underestimated by locus conservation.

Low nucleotide diversity in wild populations is often associated with factors such as: mating system, geographical distribution and gene flow [130]. Considering that *P. dumosus* is predominantly allogamous [131] the mating system must not be the main cause of the low polymorphism observed here. Instead other aspects of the biology of this species may have a more significant contribution to its low polymorphism. A potential explanation is the recent speciation of *P. dumosus*, associated with a constricted distribution and selection under environmental restrictions.

Recent species: the diversification of the species in the GP2 occurred relatively recently. Species of the group originated less than 2 million years ago [73] and the present species diversity of this group is related to the formation of the Central America high mountains.

As explained in Chapter 1, it is still unclear if *P. dumosus* originated from a hybridisation event or diverged from a basal clade. Based on the divergences scenario, it is likely that *P. dumosus* was the latest species that seceded from the basal group from which *P. vulgaris* was first separated followed by *P. coccineus* varieties. Delgado-Salinas et al. [85] have indicated that separations in the coccineus group are still ongoing. This is confirmed by our results from AMOVA that showed strong genetic structure at the accession level of *P. coccineus*. On the other hand, if the hybridisation hypothesis is considered, it implies that *P. dumosus* is the youngest of the group since its origin would come from hybridisation that involves at least two, if not three, of the older species (*P. vulgaris*, *P. coccineus* and perhaps *P. costaricensis*). In either case *P. dumosus* is a relatively recent species and it is likely that not enough time has passed for this species to accumulate polymorphism in the sequences of ITS and cpDNA. When species are recent, there will be low divergences between the sequences of each species. In a phylogenetic analysis this low divergence will generate low resolution for tree branches and increases polytomies. The following chapter of this thesis details a phylogenetic analysis where these aspects are evaluated.

Reduced polymorphism could also be a cause of genetic erosion, this is when the diversity of a species is reduced due to the disappearing of genotypes and/or alleles being

lost. Genetic erosion can occur by multiple causes [8]. Of interest in this research is that genetic erosion can occur through the maintenance of ex-situ collections. In germplasm collections genotypes and alleles can be lost or reduced during many stages in the preparation, sub-sampling, exchange, storage and regeneration of seeds [132]. Notwithstanding the lower diversity of *P. dumosus* compared to *P. vulgaris* and to *P. coccineus* has also being detected by studies carried out with in-situ populations of landraces materials [76,133]. Deterioration of habitats, fragmentation and restricted distribution can also generate and enhance genetic erosion.

Restricted distribution: Wild *P. dumosus* has a very narrow distribution [74] being found only in central and southern Guatemala [51]. It is unclear what the causes for the restricted distribution of the ancestral forms are. Bermejo et al. [7] suggested that the ancestral forms were reduced due to human intervention. Reduction of populations may have occurred in the past since feral, weedy and cultivar forms are found in a wide range of locations from Central to South America. In combination with a recent divergence, restricted distribution could have highly affected the level of diversity of wild *P. dumosus*. Often populations under restricted distribution have a low molecular diversity which is likely due to selection under limiting environmental conditions [130]. As previously described in Chapter 1, wild *P. dumosus* indeed grows under stringent environmental conditions (high altitude and cool wet areas with deep organic soil). Selection and adaptation to these conditions has necessarily played a key role in the reduction of polymorphism among the population.

Sample size cannot be excluded as a factor that may have contributed to the observed low diversity in wild *P. dumosus*. This research evaluated materials from all but one locality where wild *P. dumosus* has been collected [74]. Even though sample size could influence the estimates in this research, the accessions evaluated here are representative of the distribution range of the wild forms; therefore underestimation due to sampling would be minor.

Finally our results showed polymorphic variations only in the accessions G35877-locality Sololá and G35758-locality Sacatepéquez. Previous analysis of seed protein variation [74] identifies several protein patterns among samples of those two accessions. Both accessions are included in the core collection at CIAT germplasm bank [134] which indicates that those are representative of most polymorphism found among wild *P. dumosus*. Since there are no more known occurrences of wild *P. dumosus* in other areas than the ones evaluated here, we suggest that exploration for germplasm collections should be considered. Small gaps of potential wild populations have been identified in the eastern Chiapas and Alta Verapaz in Guatemala [41]. In addition, a wide exploration should be made in the localities of Sololá and Sacatepéquez where this research detected some nucleotide variations.

It is expected that hybrid population have a high level of genetic diversity because of the admixture of the parental gene pools [93,135]. The low level of polymorphism shown by wild *P. dumosus* does not indicate that this species is a hybrid. However, it is possible that most diversity of wild forms has been lost due to environmental constrictions, as only a small number of samples of ancestral wild *P. dumosus* are

available. It is not possible for us to know if in the past there were other areas with higher levels of diversity.

2.4.2 Do variation from cpDNA and nuclear ITS reveal that there has been hybridisation between *P. dumosus*, *P. vulgaris* and *P. coccineus*?

Hybridisation has been suggested as an explanation for the discrepancy between the chloroplast and the nrDNA genome of *P. dumosus*. My results confirm such a discrepancy; ITS sequences of *P. dumosus* are more similar to *P. coccineus* and cpDNA sequences of *P. dumosus* are more similar to *P. vulgaris*. One of the objectives in this research was to find evidence of hybridisation between *P. dumosus* and its close relatives.

Within a species, the repeats found in the ITS sequence can reveal the merging of two ITS copies within a single genome depending on three patterns. (1) Whether the different ITS repeats are maintained without recombination or homogenization [136], allowing some ITS copies to be traced back to the two parental species. (2) Different repeats undergo some degree of homogenization and recombination that generates chimeric ITS that combines motifs from both parents [137,138]. (3) ITS copies from one parental lineage may have been lost through the process of concerted evolution and one repeat becomes dominant within the new genome this can occur by mechanisms of unequal crossing over and gene conversion, these mechanisms in

conjunction are known as concerted evolution [116,137,139]. It is known that concerted evolution at the ribosomal of other legumes species e.g. *Glycine max*, has been rapid and complete enough to render the ribosomal repeats as a unique chromosome [85]. Given the high homogeneity of ITS sequences within *P. dumosus*, is apparently that concerted evolution is operating, which could also explain the lack of additive polymorphism with *P. vulgaris*. The analysis of genetic distances and nucleotide variation indicates that homogenisation in *P. dumosus* may have occurred bias to *P. coccineus*. In diploid species and homoploid hybrids, this process can occurred fast and erase the signal of one of the parentals in a few generations [140], as it has been demonstrated in several groups e.g. *Senecio* [141], *Gilia* [142] *Heliosperma* [143], *Gossypium* [137] *Paeonia* [144]. It is possible that other regions of the nrDNA of *P. dumosus* also shows the bias to *P. coccineus* genome, as later demonstrated in Chapter 4.

At the cpDNA, two shared haplotypes were found, however, those may not indicate hybridisation but rather the detection of ancestral haplotypes. The haplotype rs7 was found in all species of the bean GP2, which indicates the conserved nature of the locus *rps14-psaB* and suggests that rs7 is likely an ancestral haplotype. Most likely this haplotype is associated with the split of the *Vulgaris* group, perhaps from an extinct or unknown ancestral progenitor. One recently published analysis of cpSSR [84] also reported evidences of ancestral polymorphism conserved among *P. vulgaris* and *P. coccineus* accessions associated with the lineage separation of *P. vulgaris*.

Under hybridisation it is expected that all sympatric accessions share multiple haplotypes or a mix of the parental haplotypes [91]. It is doubtful that the haplotype as3 (locus *accD-psaI* spacer) indicates an event of hybridisation. This is because as3 is present only in one accessions 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. The presence of the haplotype as3 is likely a case where two sequences have the same polymorphism by convergence. However this is not to discard the possibility of this being a case of an ancestral haplotype. The discontinuity to observe this haplotype on more accessions perhaps is due to a process of incomplete lineage sorting, loss of variation or missing sampling. Similar discontinuity with ancestral haplotypes from cpDNA have been observed by Chacon et al. [57] among *P. dumosus*, *P. costaricensis* and *P. vulgaris*. Although the shared haplotypes found here do not directly suggesting events of hybridisation they are providing information about the common ancestry and diversification of the species.

2.4.3 The case of *Phaseolus persistentus*

The species *P. persistentus* was initially classified in the Falcati section [51] since morphological characters of this species are similar to those observed in *P. leptostachyus* (type species of the Falcati section). However studies with ITS [73] suggested that the species belongs to the *Phaseoli* group and is close to *P. vulgaris*. Gaur [72], has indicated the need for reconfirmation of the assignation of this species

to either group since morphological characters are widely different from the *Phaseoli* group.

Our analysis of sequences of cpDNA and ITS did not find evidence that indicates that *P. persistentus* could be closer to the falcate group than to the *Phaseoli* group. The results indicate that divergences between *P. persistentus* and *P. leptostachyus* were larger than between *P. persistentus* and any other species. Also the level of divergences between *P. persistentus* and the set of species that belongs to the falcate group were similar to the level divergences among the species of the group. In contrast, the level of divergences between *P. persistentus* and *P. leptostachyus* were similar to the level of divergences between any of the species from the *Phaseoli* group and *P. leptostachyus*.

At this point it is valid to question whether or not *P. persistentus* is a true species as it lacks specific variation in sequences that showed strong species specific variation among bean species. *P. persistentus* is not only highly related to *P. vulgaris*; it also presented the same nucleotide polymorphism and the same haplotypes as those observed in *P. vulgaris* (figure 2.3). Overall our data indicates that it is likely that *P. persistentus* is an accession of *P. vulgaris*. Unfortunately, only one specimen of this species has being observed which limits experimentation.

2.5 Next Steps

The data from this chapter contributed to the evaluation of the diversity of wild *P. dumosus* and assessed evidence of hybridisation coming from nucleotide polymorphism. Once genetic diversity is examined, a fundamental next step is to perform a phylogenetic analysis to better understand the relationship under evolutionary models and phylogenetic inferences. This will provide information not only of nucleotide changes but also information from conservative regions. Phylogenetic analysis among species of the GP2 applying different inferences and evolutionary models is performed in the following Chapter 3.

The findings in this chapter bring up the question of whether other regions of the nuclear genome of *P. dumosus* perhaps present more polymorphism. It is possible that introgression could be detected in other variable regions of the nuclear genome. Such analysis is performed in Chapter 4, which also addresses a comparison between wild and cultivar accession.

3 Phylogenetic relationships between the common bean secondary gene pool

3.1 Introduction

As established in Chapter 1, a greater understanding of the within species genomic variability and the genetic discontinuities between species of the common bean GP2 will facilitate the breakdown of hybrid incompatibilities, required for wide interspecific crosses for crop improvement.

This thesis focuses on the study of the molecular evolution of wild *P. dumosus*, the least studied cultivated bean species. In Chapter 2 it was established that *P. dumosus* molecular diversity is low and confirmed that *P. dumosus* is highly homogeneous. It was also confirmed that levels of molecular divergence are very low between *P. dumosus* and its closest relatives, *P. coccineus* and *P. vulgaris*.

It is still unclear if this species originated from a hybridisation event that involved *P. coccineus*, *P. vulgaris* or perhaps *P. costaricensis* [51] or was the product of ongoing divergences within the *P. coccineus* group. The hypothesis of hybridisation emerged as an explanation for the intermediate morphology of *P. dumosus* that combines characteristics from *P. vulgaris* but also from *P. coccineus* [79]. This hypothesis has been supported by the incongruent relationships derived from the nuclear and chloroplast genome of *P. dumosus*. Evaluation of the chloroplast genome have indicated that *P. vulgaris* is *P. dumosus*' closest relative but examination of the nuclear genome indicates that *P. coccineus* is *P. dumosus* closest relative [19,73,75,81–85].

Although the above studies have contributed to exposing the complexity of the relationships among bean species, to date many questions about the origin and relationship of *P. dumosus* remain unresolved. Particularly, it is necessary to confirm the discrepancies between the nuclear and the chloroplast genomes through a wide analysis of wild representatives of the three species. All previous analyses used either a nuclear or a chloroplast marker with only a small numbers of individuals of wild *P. dumosus*. One study [73], presented a phylogeny for the genus *Phaseolus* by combining a nuclear and a chloroplast marker but only a few representatives of *P. dumosus* were included and the discrepancies between phylogenetic signals from the two genomes were not presented. A phylogenetic analysis that includes several wild accessions of *P. dumosus*, *P. coccineus* and *P. vulgaris* is hence lacking. Moreover, as discrepancies between gene trees and species trees can be due to many causes including lineage sorting, introgression, ancient hybridisation, data errors or a combination of all of these, further confirmation of the nature of the discrepancies is required. Also, the hypothesising of past hybridisation events by genomic discrepancies need to be evaluated using different genomic regions and a wide number of individuals [91,145].

The general aim of this thesis chapter is to provide a detailed illustration of the evolutionary relationships between wild forms of these three species. The chloroplast sequences of *accD-psaI* spacer, *trnT-trnL* spacer, *trnL* intron and *rps14-psaB* spacer, and sequences of the ITS region generated in Chapter 2, were used for a phylogenetic reconstruction. The methods of Maximum likelihood and Bayesian methods were used for phylogenetic analysis conducted to confirm the divergence of relationships observed by the nuclear and chloroplast genome of wild *P. dumosus*. Additionally,

recently developed techniques were also used for the identification of reticulation events. Even if divergence between gene trees from independent loci can suggest a case of hybridisation, events of reticulation cannot be reconstructed directly by tree-building methods, as those are designed to resolve any conflicting phylogenetic signal by displaying the most supported relationships in a bifurcation manner [145,146]. The use of hybridisation networks is a better tool to investigate possible reticulation events since it can display conflicting phylogenetic signals and several patterns of relationships [145]. My final objective in this chapter was to use a phylogenetic hybridisation network to identify putative events of reticulation that could explain the divergence between the two genomes and provide further evidence of the origin of *P. dumosus*.

3.2 Phylogeny reconstruction in wild bean

Phylogenetic trees (gene trees) were drawn to illustrate the relationship between *P. dumosus*, *P. vulgaris* and *P. coccineus*, based on sequences of nuclear ITS and chloroplast. Trees from each genome were inferred separately. Subsequently, a combined analysis using phylogenetic networks was performed. Figure 3.1 presents an overview of the procedures, analysis and parameters used to derive the phylogenetic trees. Described below are the methods and parameters used for the phylogenetic analysis: alignments, models of character evolution, methods of inference and probabilistic testing.

3.2.1 Alignments and a priori analysis

Gene trees were built using the set of sequences of wild beans that were described in Chapter 2. For the purpose of phylogenetic analysis, the sequences were first re-aligned using a progressive alignment algorithm with the Muscle aligner program [118]. The alignments were depurated using the Gblocks[©] program [147]. The depuration step eliminates poorly aligned nucleotide positions by selecting blocks of positions that are the most conserved and informative for a phylogeny.

For part of the phylogenetic analysis the alignments of the cpDNA were used as a matrix of concatenated loci or partitions (alignments of individual loci). In each analysis it is stated whether concatenated or partition alignment was used. The concatenated matrix of cpDNA included the sequences of the loci *accD-psaI* spacer, *trnT-trnL* spacer, *trnL intron* and *rps14-psaB* spacer and the matrix was built using the program Mesquite Version 2.75 [148]. Because the phylogenetic signal from the concatenated data can hide the signal of any individual locus, before the application of more complex parameters and before concatenation (in the case of the cpDNA) *a priori* exploratory phylogenetic analysis was conducted. This analysis allowed observation of the phylogenetic signal generated from each individual locus.

The prior analysis consisted of the elaboration of a gene tree for each independent locus. These trees were generated with the program MEGA 5 [149] applying the Neighbor Joining (NJ) method [126]. The NJ method was chosen for the exploratory analysis because this method demands low computational resources and trees are generated quickly.

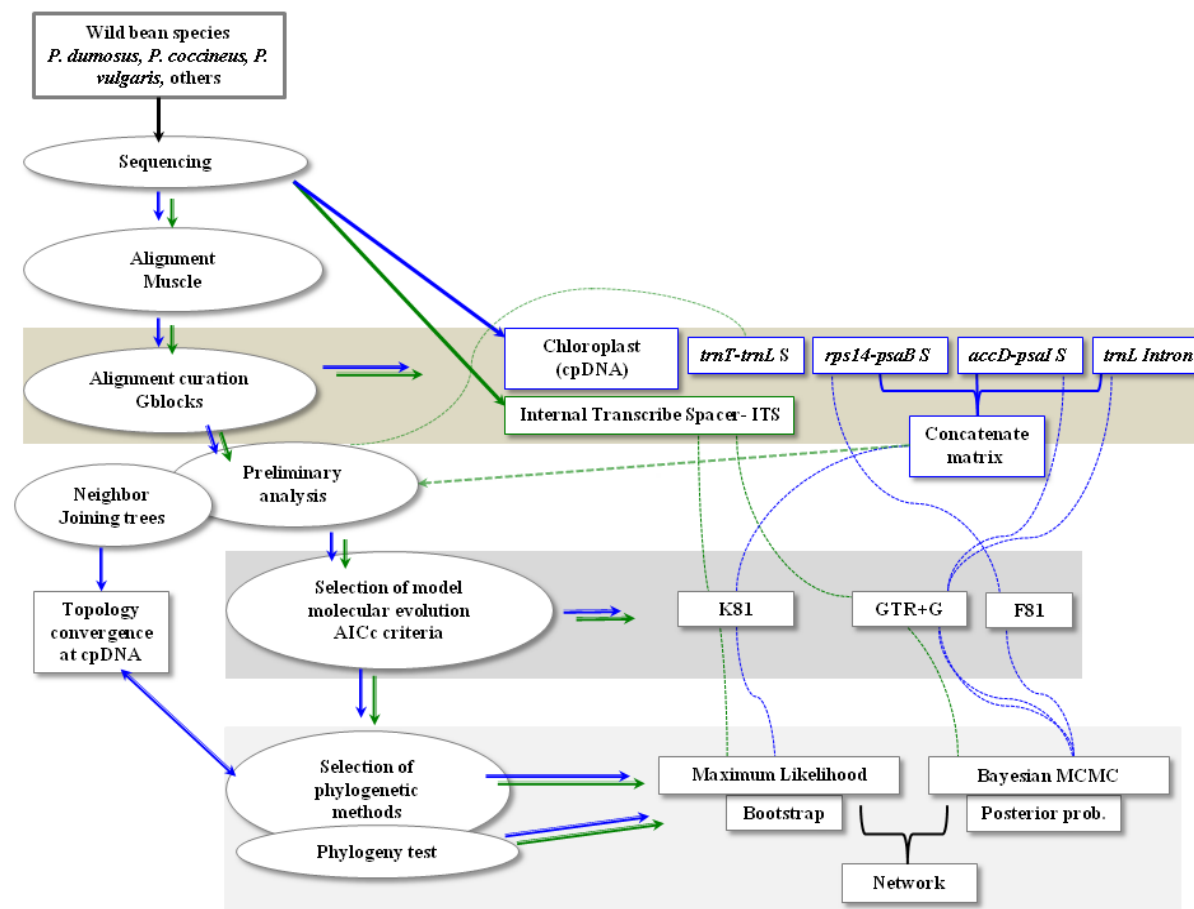


Figure 3.1 Illustration of methods and parameters used in the construction of phylogenetic trees using sequences of cpDNA (blue lines) and ITS (green lines) for wild relatives of the common bean GP2. Dashed lines indicate parameters or analysis applied to partitions, solid lines indicate parameters and analysis applied to all sequences

3.2.3 Methods used to infer the phylogenetic relationship of *P. dumosus* and its close relatives

To infer the phylogenetic relationship of the wild beans, a model of evolution was applied to the sequences under analysis (cpDNA and ITS). This model of character evolution ('mce') was used to estimate the rates of change at the sequence positions. The mce set the parameters which allowed estimation of the probabilities of nucleotide substitutions in the sequence as explained elsewhere [150].

The selection of the model of character evolution for the cpDNA and ITS sequences was made using the program jModelTest [151,152]. The mce was selected as the best fit model for each partition (i.e. each locus) and also for the entire concatenated matrix (for cpDNA). The selection was made under the Akaike Information Criteria (AICc) [153]. This criteria determines the relative likelihood of two or more models fitting the data. The selected models were as follows: The TPM (Three Parameter Model) [154], was selected as the best fit model for the ITS, the *trnT-trnL spacer*, *trnL intron*, and the *accD-psaI* spacer. The model F81 [155] was selected as the best fit model for the *rps14-psaB* spacer. For the whole concatenated matrix the best fit model was the GTR +G (General Time Reversible) [156]. Figure 3.1 gives an illustration of the way that the models were evaluated in each phylogenetic reconstruction.

The phylogenetic relationship of wild *P. dumosus* and its closest relatives was established by using two methods of phylogenetic inferences, first the **Maximum Likelihood** method and second the **Bayesian** method. Finally using the information generated from the phylogenetic trees, a **phylogenetic hybridisation network** was

used to search for evidence that indicates reticulation and hybridisation among the bean species.

3.2.3.1 The Maximum Likelihood (ML) method

The ML method [155] is based on the statistical methodology of deriving estimates. The method estimates the likelihood that a known sequence would have evolved on a tree for which a structure and an evolutionary model are given [157]. The method chooses the tree which maximizes the likelihood of observing the given set of data in the tree. Using the cpDNA sequences and the ITS sequences of wild beans, two ML trees were inferred, trees were calculated and drawn using MEGA5 [149]. As the ML method does not analyse the data by partitions, the full concatenated cpDNA sequence was analysed as a unit. The tree with the highest likelihood was chosen under the GTR model of evolution with a gamma distribution with six categories. For the ITS sequences the tree with the highest likelihood was chosen under the GTR model with a gamma distribution with four categories. The final structures were made by a Nearest-Neighbor-Interchange rearrangement (NNI) this is a heuristic to improve the likelihood of a tree. In Mega NNI is calculated, by rearranging trees at the "neighbor" level (swap their subtrees) and choosing the "neighbor" tree with the higher likelihood [158].

Both ITS and cpDNA phylogenies were tested by the bootstrap test [159]. In MEGA5, the bootstrap test is calculated by randomly re-sampling or realigning the nucleotide on the original alignment and constructing a randomised sequence data set that differs

by some small random changes from the real alignment. Each new alignment is used to construct a new tree, the topology of each of these trees is compared against the topology of the original tree. The randomized sequences are generated by sampling columns from the original alignment and placing them in a random order, but conserving the same length of nucleotides. Some columns can be sampled more than one time and some columns can be excluded. The tree constructed with the randomized sequence may or may not look like the original tree. If the phylogenetic signal is strong in the sequences, the topologies generated during bootstrapping should differ little from the original tree [157]. To conduct the bootstrapping test the number of randomised sequences to build is fixed i.e 1000 samples. The bootstrap value of a branch in the original tree is determined by the percentage of replicate trees in which the associated taxa in the branch are clustered together. Bootstraps values above 70% are generally accepted as strong support for a clade [159].

The bootstrap percentage was calculated from a thousand replicates for the ML phylogeny made with the ITS sequences of wild beans and ten thousand replicates for the ML phylogeny made with the cpDNA of wild beans.

3.2.3.2 Bayesian Markov chain Monte Carlo (MCMC) method

The Bayesian method also applies the maximum likelihood algorithm, but unlike the ML method, the Bayesian method searches for a set of possible trees that can fulfil the data, instead of one unique tree [160]. The likelihood in a Bayesian inference is used to estimate how accurate the hypothesis (tree or relationship) explains the observed

data. The Bayesian inference is constructed by choosing the most probable tree. The probability of the tree is a function of both the likelihood and the prior probability.

The prior probability and posterior probability are the two fundamental components of a Bayesian analysis. The prior probability is the original probability, the probability assigned when there is no previous knowledge of which is the relationship between the taxa. The posterior probability measures the likelihood that a tree is the true tree given the model, the priors and the data. The estimation of the posterior probability (PP) requires the integration of several parameters: topology, branch lengths and the parameters of the model of character evolution, amongst others [157,161,162].

PP is mathematically defined by the Bayes theorem as:

$$PP(A|B) = PP(A)PP(B|A)/PP(B)$$

Where:

A=data (matrix of sequence alignments)

and

B=parameters of the model (topology, branch lengths, substitution model).

Analytically it is not possible to estimate the PP and integrate all parameters even for a small number of taxa [163]. However it is possible to calculate the posterior probability of a tree by using the Markov Chain Monte Carlo (MCMC) simulation technique. This technique is used to generate a chain of phylogenetic trees that are sampled according to their likelihood. The method allows assigning of parameters to each partition of the data which should lead to better phylogenetic estimations

[164,165]. Making a Bayesian inference with the data from wild bean allows the evaluation of each locus of the cpDNA under its estimated model of character evolution, and is therefore expected to generate a better fitted phylogeny.

The Bayesian (MCMC) trees for wild beans were created with the program Mr Bayes version 3.2 [160,166]. The priors of the model were set as follows: For the chloroplast sequences three partitions were established. Partition one corresponds to the *trnL* intron sequences, which was analysed under the GTR + G model. Partition two corresponds to the *rps14-psaB* spacer sequences, which was analysed under the F81 model. Partition three corresponds to the *accD-psaI* spacer, which was analysed under the GTR model.

A discrete gamma 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. Note that the GTR model applied to partitions one and three was a replacement for the TPM model that is not implemented in Mr Bayes. The ITS region was analysed under the GTR model, with a discrete gamma distribution and four discrete categories. The clades support in Bayesian MCMC phylogenies is given by the proportion of the MCMC sampled trees that contains a particular split. The MCMC calculates and labels each split by their posterior probability.

For phylogenetic inferences with ITS and cpDNA sequences of wild bean the simulations of the MCMC were run under the following parameters: For both ITS and cpDNA sequences, the searches for tree topologies were performed in two parallel

runs. Each run was performed with four Markov chain settings (three “heated” chains and one “cold” chain). Tree searches started from a random tree and 25% of the sampled trees were used for the initial phase (burn-in) when the samples still influenced by the starting point. For the inference of the cpDNA phylogeny, ten million generations were tested and tree samplings were saved every 1000th generation. For the inference of the ITS phylogeny one million generations were tested and tree samplings were saved every 500th generation.

The values of the standard deviation were used as criteria of convergence of the two parallel runs, this is: when the two runs converge into a stationary distribution the average standard deviation of the splits frequencies are close to zero, meaning that the two trees become increasingly similar. The distribution of the log likelihood values and the values of the PSRF (potential scale reduction factor) were also checked. In Mr Bayes the log likelihood values are plotted against the generations. The plot should show a random distribution of the values without a tendency to increase or decrease. The PSRF represents the variance within runs compared against the variance between runs, as the runs converge the variance should approach 1.

3.2.3.3 Consensus networks to represent incongruence among phylogenetic trees of the common bean secondary gene pool

In order to explain the divergence between the chloroplast and the nuclear genome of *P. dumosus*, a hybridisation network was generated. This phylogenetic approach

represents the relationship between two gene trees in a simple model that shows the reticulate evolutionary history of each tree [106]. The hybridisation networks allow merging of the topologies from cpDNA and ITS into a network that proposes hybridisation events as explanations for the divergence found in each topology.

The hybridisation network was computed using the program Dendroscope 3.1 [167]. The Networks were computed using the ML trees for the cpDNA and ITS of wild bean. Branches with BS<60% were collapsed, both trees were rooted at the *P. acutifolius* branch and bean accessions that were not represented in both phylogenies were not included in this analysis.

3.3 Results

3.3.1 What are the wild bean relationships when phylogenetic inference is made from sequences of chloroplast?

The tree topologies from cpDNA sequences inferred by the Bayesian and ML approaches were largely congruent. The optimal trees for the Bayesian and ML analysis are shown in Figure 3.2.A. Bayesian and ML topologies represented the major relationship as two major clades (clade 1 and clade 2). The major relationships were strongly supported by the Bayesian topology (PP>95%) and highly supported in the ML tree topology (BS> 80%- 99%).

Clade 1 was formed by a polytomy of wild *P. dumosus* and *P. vulgaris*. There was no further resolution among the accessions of *P. dumosus*, which is in agreement with the results from the molecular diversity analysis detailed in Chapter 2. Three accessions of *P. vulgaris* resolved to branches that were well supported in the Bayesian topology (PP>99%) but not in the ML (BS<60%) topology. One accession of *P. coccineus* (G36206) and one accession of *P. costaricensis* (G40604) were also included in clade 1 Figure 3.2A. **Clade 2** was formed by subclade *P. coccineus* and a sister subclade comprising of 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 the accession G35650-Guatemala. The divergence in this accession may be due to the fact that it lacks haplotype rs7, as observed in Chapter 2.

3. Phylogenetic relationships between the common bean secondary gene pool

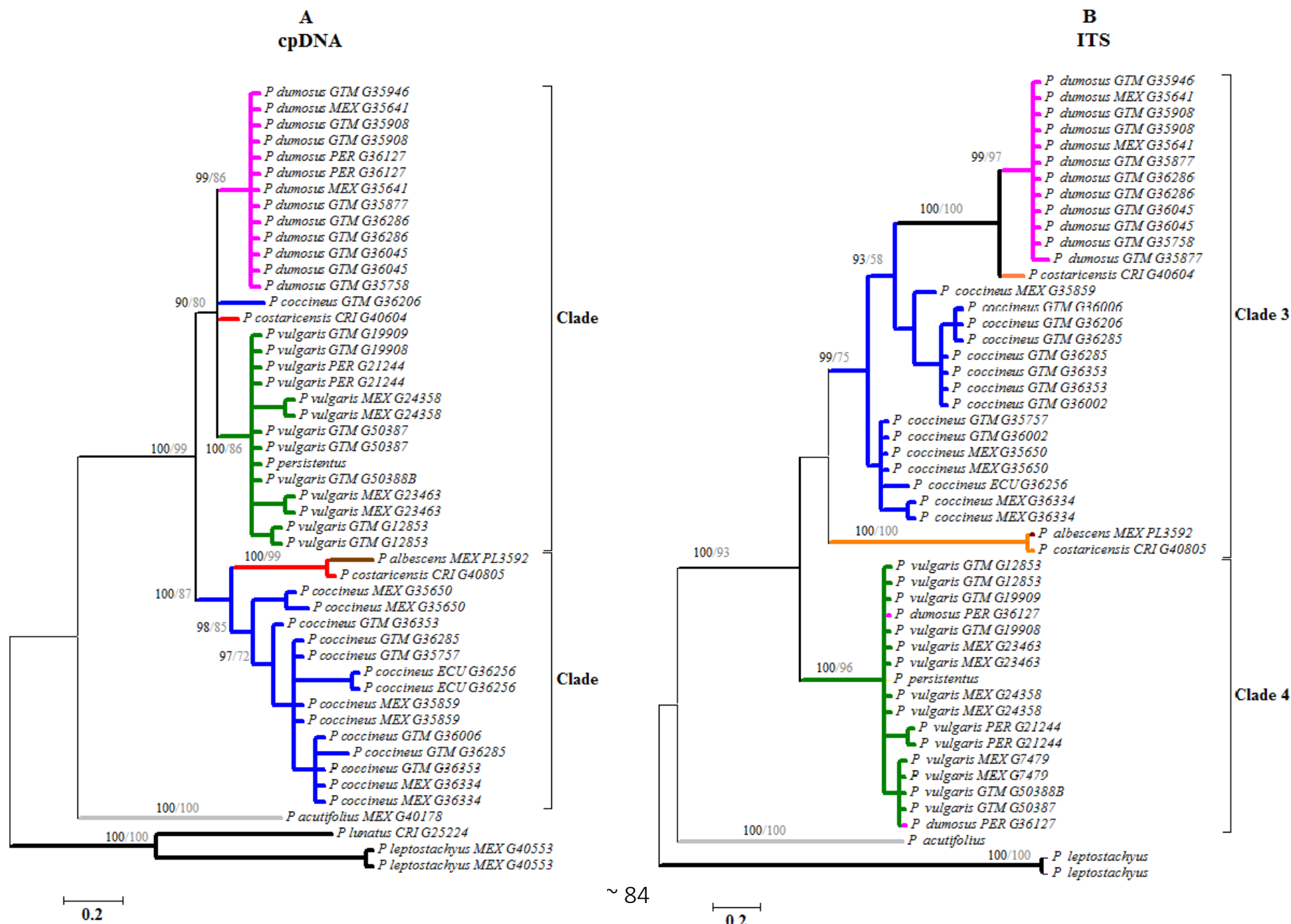


Figure 3.2 cpDNA and ITS phylogenies for wild accessions of the common bean GP2. Trees were generated from sequences of the cpDNA (A) and the ITS region (B). Phylogenetic inferences were made by Bayesian MCMC and ML analysis. Statistics of the branches are posterior probabilities >60% (number in black) and bootstraps values >60% (number in grey). Major clades are identified by numbers 1 to 4. Branch lengths are proportional to number of nucleotide substitutions (scales represent 0.2 substitution).

3.3.1 What are the wild bean relationships when phylogenetic inference is made from sequences of the ITS?

Similar to that observed in the cpDNA analysis, there was convergence in topologies of the Bayesian and ML analyses derived from sequences of the ITS of wild beans Figure 3.2.B. The major relationships were also represented by two clades (clade 3 and clade 4). Clades were strongly supported by both methods (BS>93%, PP>95%); A subclade of *P. coccineus* and one accession of *P. costaricensis* were located as the sister clades of *P. dumosus* in **clade 3**. This is in contrast to the results from the cpDNA analysis where *P. vulgaris* appears as the sister subclade of *P. dumosus*. There was no resolution among *P. dumosus* accessions, whereas *P. coccineus* accessions resolved partially, and the *P. coccineus* subclade that was closer to *P. dumosus* was well supported (BS=94, PP=100). This subclade was mainly formed by *P. coccineus* accessions from Guatemala, whilst accessions from Mexico and Ecuador did not resolve from the basal node of the clade.

Clade 4 was largely composed by *P. vulgaris*. There were two separations among the accessions of this species. The two cultivated accessions of *P. dumosus* and the accession of *P. persistentus* were also included in this clade.

Overall the phylogenetic analysis conducted here shows that wild *P. dumosus* as a monophyletic group is highly supported. Based on nuclear-ITS sequences *P. dumosus* is closer to *P. coccineus*, and more specifically to *P. coccineus* accessions from Guatemala. Based on the cpDNA, *P. dumosus*' closest relative is *P. vulgaris* and there is no indication of any *P. vulgaris* accessions being closer to *P. dumosus* than others. The phylogenetic analysis also shows that there are differences between the accessions of *P. costaricensis*. Both phylogenies (cpDNA and ITS) show that one accession (G40604) is close if not a sister to *P. dumosus* and the second accession (G40805) appears in an intermediate position between the two major clades.

3.3.2 Phylogenetic network analysis, can the divergences between cpDNA and ITS topologies be explained by reticulations?

A phylogenetic network was calculated to display the phylogenetic conflict between the nuclear and the chloroplast genome as reticulations that may have occurred between the species under study. The hybridisation network (Figure 3.3) and the galled network (Appendix II, B1), showed three splits representing the relationships among the bean GP2. Each corresponded to a species group, and overall the representations of splits were in agreement with the tree topologies (ML and Bayesian trees). The network identified 21 putative events of hybridisation, 16 of which were reticulation events within species and five were reticulation events of putative hybridisation across species. A reticulation node was generated as an explanation of the divergence between the chloroplast and the nuclear genome of *P. dumosus*.

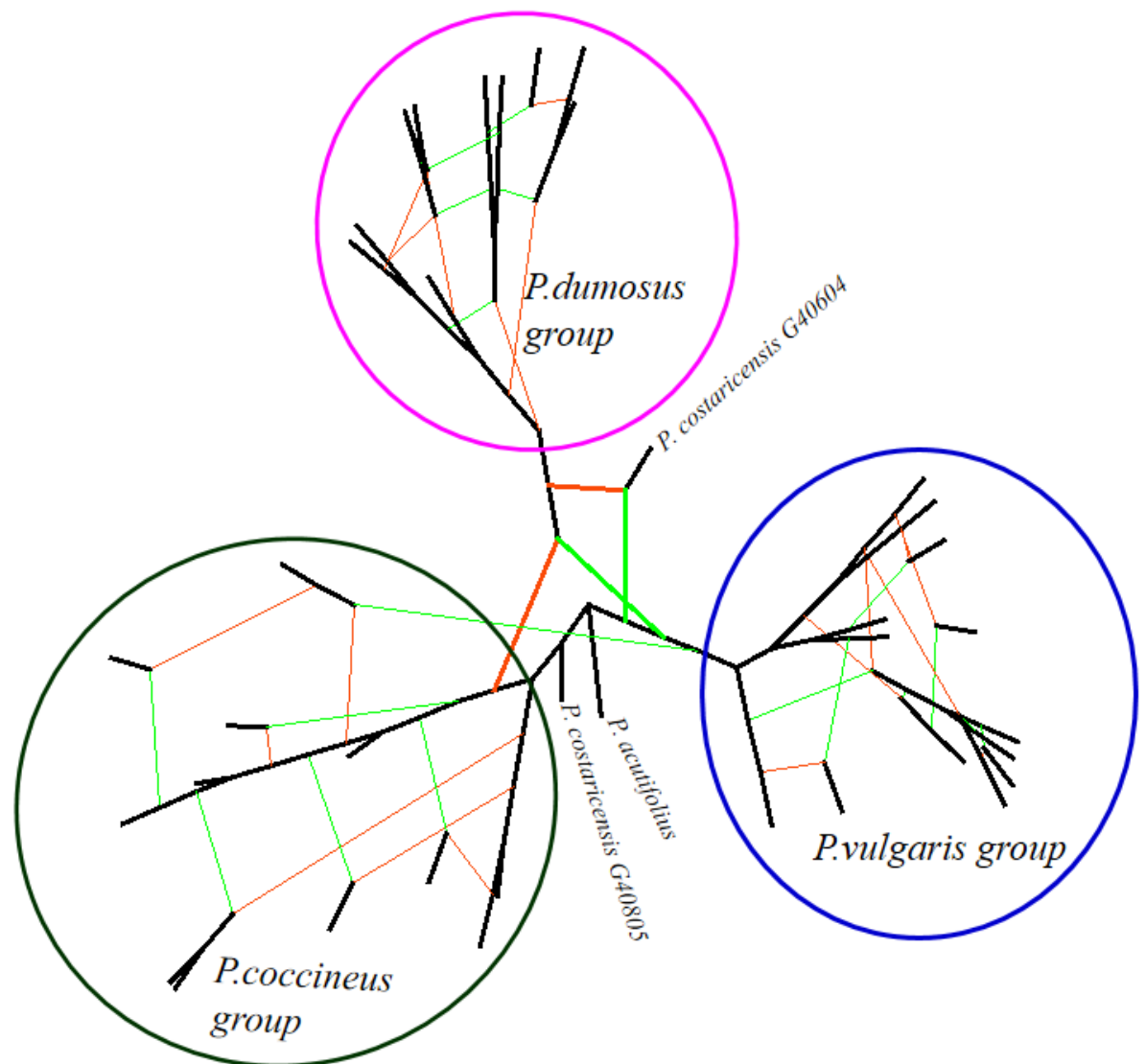


Figure 3.3 Hybridisation network showing the hypothetical reticulation events among species of the common bean GP2. The network was generated from 60% BS consensus trees of ML analyses for sequences of cpDNA and ITS. Reticulations from the chloroplast genome are indicated by green lines, those from the nuclear genome are represented by orange lines. Strong lines indicate reticulations between species and fine lines indicate reticulations within species. Accession names are omitted for clarity.

3.4 Discussion

3.4.1 The phylogenetic relationships of the common bean secondary gene pool do not follow a tree-like shape

The reconstructed phylogeny generated in this chapter confirmed the divergence in the relationships derived from the chloroplast genome and the relationships derived from the nuclear genome of *P. dumosus*. Even when different evolutionary models were applied, both the phylogenetic methods (ML and Bayesian) generated similar tree topologies and well supported clades. This provides strong evidence of the accuracy of the relationships that were inferred.

Overall the results from independent phylogeny shows that *P. dumosus* is closer to *P. vulgaris* when analysed by the chloroplast genome, in agreement with previous analyses of cpDNA using restriction enzymes [56, 82, 168] and cpSSRs [83,84]. On the other hand the results from the nuclear genome indicate that *P. dumosus* is closer to *P. coccineus* in agreement with previous restriction enzymes and ITS studies [73,81,85].

Under the hypothesis that the genomic divergences in *P. dumosus* are due to hybridisation, the hybridisation network indicates that when the information from both genomes (chloroplast and nucleus) is combined, the relationships between *P. dumosus*, *P. coccineus* and *P. vulgaris* do not follow a bifurcate tree-like organisation,

but rather a net of splits (Figure 3.3) in which the complexity of the relationship occurring between accessions of each species is represented by the many reticulations.

In a tree-like figure, conflicting phylogenetic signals are forced to resolve in either one or the other side of the bifurcation and complex interactions are not represented. The conciliation of the phylogenetic signals from cpDNA and ITS of *P. dumosus* into one tree-like figure, resolved to agree with either the cpDNA or the ITS as presented in Delgado-Salinas et al. [73]. But a tree topology mask important elements of the true relationships. In the case of the relationships between *P. dumosus*, *P. coccineus* and *P. vulgaris*, the splits presented here are likely to be a better representation of the interactions that have occurred among those species. It allows putative reticulations that may have occurred among those species to be detected as well as proposing intermediate forms between each species group.

3.4.2 Hybridisation may have given origin to *P. dumosus*

The hybridisation network (Figure 3.3) proposes that *P. dumosus* is a hybrid between a lineage of ancient *P. coccineus* (the split from Guatemala) and an ancient lineage on the *P. vulgaris* line. The location of the split indicates that the hypothetical hybridisation event may have occurred shortly after the split from *P. coccineus*.

However the network does not prove a hybridisation event between *P. coccineus* and *P. vulgaris* giving rise to the origin of *P. dumosus*. Indeed the fact that the reticulation is associated with ancient lineages indicates that the hypothesis needs to consider more

biological evidence, including evaluation of additional ancestral lines particularly in the *P. vulgaris* line. Unfortunately no fossil pollen or other fossil records of the *Phaseolus* species that could help in researching this hypothesis, have yet been found, as explained in Chapter 2. Here we have evaluated the currently available material of wild *P. dumosus*.

Overall our results from traditional phylogeny and hybridisation networks show a strong indication that *P. dumosus* originated from a hybridisation event. Following the Gottlieb [169] criteria to validate a hybrid event, the current evidence pointing to a *P. dumosus* hybrid origin accounts for 75% of these criteria. Summarising the evidence as follows: An intermediate morphology exists between *P. coccineus* and *P. vulgaris* [7,170]. Ancient forms of *P. dumosus* are found only in Guatemala and are in sympatry with ancient forms of its putative parental species [74]. Artificial crosses can produce fertile offspring [51]. Molecular evidence indicates the incongruence between the relationships observed from analysis of nuclear and chloroplast genomes [73,81,82, 85,131,168]. This incongruence has been reconfirmed by evaluation of all available accessions of wild *P. dumosus*. The putative lineage of *P. coccineus* and *P. vulgaris* that hybridised to form *P. dumosus* has been identified (this research).

As discussed in Chapter 2 concerted evolution can be one of the mechanisms that could explain the lack of signal from *P. vulgaris* in the nrDNA of *P. dumosus*. It is possible that *P. dumosus* is a recent hybrid under the effect of concerted evolution in which intralocus repeats of the ITS had undergone homogenization towards the *P. coccineus* genome. As consequence of repeats homogenisation, the phylogenetic

inferences from the ITS are misleading, presenting only the relationship of *P. dumosus* with *P. coccineus*.

Although the evidence exposed above does suggest that *P. dumosus* originated from a hybridisation event, hybridisation is difficult to distinguish from lineage sorting, this is the retention of ancestral polymorphisms through multiple speciation events. Lineage sorting could also explain the pattern of relationships at the cpDNA of *P. dumosus*, however by lineage sorting is unlikely to find a pattern of multiple shared polymorphisms between *P. dumosus* and *P. coccineus* at the nrDNA. A multi-locus analysis of the nrDNA of *P. dumosus* will provide further evidences to better discern between these processes.

4 Genetic structure and relationships of *P. dumosus* inferred from whole genome DArT analysis

4.1 Introduction

This thesis research has been orientated to generate information about the patterns of relationships among the common bean GP2 with particular focus on the species *Phaseolus dumosus*. As has been remarked throughout this research, further knowledge is required about the variation and putative hybrid origin of *P. dumosus*, which holds a valuable resource of agronomic traits that are lacking in domesticated varieties of the common bean [38]. In the preceding Chapter 2 and Chapter 3 an analysis of the molecular variation and relationships arising from sequences of the cpDNA and the nuclear ITS region was carried out. Based on this analysis the discrepancies between the phylogenetic signals derived from the chloroplast and the nuclear genome were confirmed and a lack of diversity was detected among wild types of *P. dumosus*. To further support the foregoing results and truly understand the possibilities of interspecific introgression for crop improvement, a multi-locus population level study was required. Likewise it was necessary to compare the diversity between landraces, cultivars and wild types of *P. dumosus* and other taxa of interest. This will facilitate a better assessment of *P. dumosus* diversity and more conclusive evidence of the patterns of relationships and speciation processes within the *Phaseolus* clade. For this aim a Diversity Arrays Technology (DArT) array was developed for *P. dumosus*.

4.1.1 DArT markers: uses, advantages and disadvantages

A DArT array provides a whole genome screening of polymorphism, via a high-throughput microarray platform that generates hundreds to thousands of polymorphic markers in a single assay. The polymorphism in DArT markers is of the dominant type, which indicates presence or absence of alleles. DArT detects molecular polymorphism (SNPs and INDELs) resulting from the polymorphism found at the restriction site generated by the reduction of a pooled DNA representation [112,171]. The DNA restriction fragments are cloned, randomly isolated and arrayed in a solid phase slide. The array is then hybridised with a reduced genomic DNA.

Compared with sequencing technologies, DArT markers is a low cost technology that has the potential to source a large number of polymorphism, providing useful information on genetic diversity and for phylogenetic analysis [172]. DArT markers have been proven to be highly efficient in identifying DNA polymorphism and characterising germplasm without previous knowledge of the DNA sequence. This makes the method applicable to all species regardless of how much DNA sequence information is available for that species [171,173]. Compared to other molecular markers such as AFLPs, RAPDs, and SSRs, DArT markers overcome the difficulties in correlating bands on gels with allelic variants by utilizing hybridisation-based methods and solid state surfaces [174]. It is a high throughput, quick, and highly reproducible method. It is cost-effective, with an estimated cost per data point tenfold lower than SSRs markers [175].

A main disadvantages of using DArT is that the markers are primarily dominant (present or absent), which limits its application. The development of the array can be laborious as it involves several steps, including preparation of genomic representation for the target species, cloning, data management and analysis. Currently genotyping with DArT markers does not exist as commercially independent technology as the arrays are only constructed by the team that developed technique through their economical commercial service.

DArT arrays have been developed for *P. vulgaris* and it proved to be effective for analysis of genetic diversity within this specie [176]. DArT markers have been used for analysis of genetic diversity, linkage mapping and in population structure of germplasm collections in other legumes: mungbean (*Vigna radiata* L) and soybean (*Glycine max* Merr) [177], chickpea (*Cicer arietinum* L) [178] and in several other crop species such as cassava (*Manihot esculenta*) [175], winter rye (*Secale cereale* L) [174], potato (*Solanum tuberosum* L) [179], and common wheat (*Triticum aestivum* L) [180].

This chapter describes the development and use of DArT markers to assess the diversity, population structure and phylogenetic relationships of *P. dumosus* and its closest relatives. The aims of this thesis chapter is to use genome-wide DArT markers for: 1. Profiling *P. dumosus* germplasm (accessions) and searching for differential levels of genetic structure by applying a fine scale genetic diversity analysis. 2. Conduct a structure analysis between *P. dumosus* and its closest relatives, looking for evidence that suggest past hybridisation events. 3. Confirm and complement previous

phylogenetic analysis and patterns of relationships inferred from loci of the nuclear genome.

4.2 Materials and Methods

4.2.1 Plant material and DNA sets

Figure 4.1 shows the distribution of the bean genotypes used in this research. It includes the subset of samples used in Chapter 2 and Chapter 3 plus seventeen accessions of landrace and cultivated varieties, mainly of *P. dumosus* that were obtained from the United States Department of Agriculture (USDA). The seed germination, leaf harvesting and DNA extraction from plant materials were carried out as described in Chapter 2.

4. Genetic structure and relationships of *P. dumosus* inferred from whole genome DArT analysis.

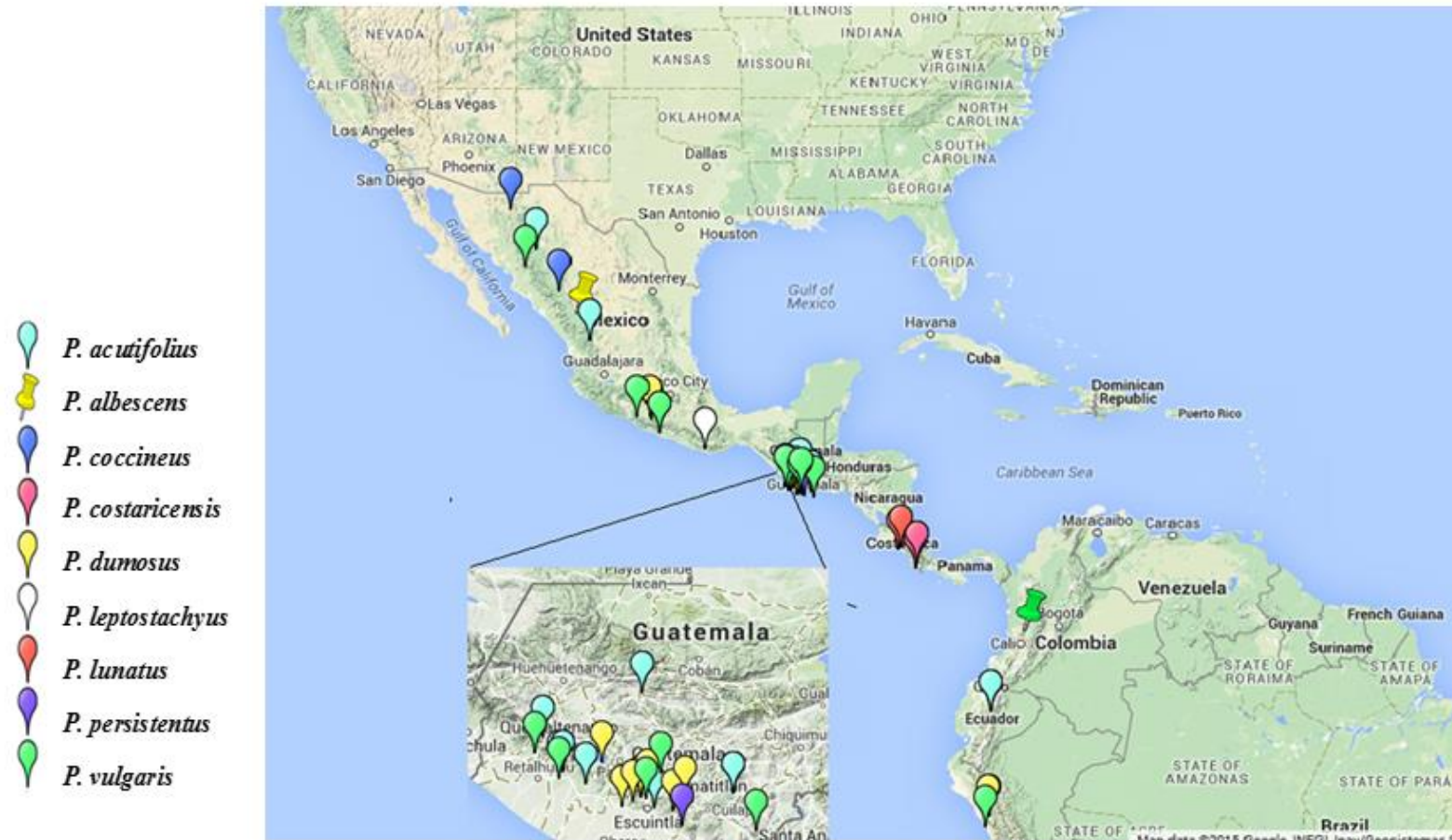


Figure 4.1 Geographic distribution of species of the GP2. (Based on google map data, Google, 2015)

The quality of the DNA was tested by digesting it with the restriction enzymes *HindIII* and *PstI* (New England Biolabs) following the buffer conditions suggested by the manufacturers. Digested and non-digested DNA were loaded side by side and run on an electrophoresis gel at 1 % agarose. DNA was visualised under UV light using Gel Genius (Syngene Bio Imaging System, BioSciences). The DNA concentration was quantified on the NanoDrop ND-1000 spectrophotometer (Labtech, UK) recording the absorbance at A230, A260 and A280. Approximately 1 µg of pure DNA from each bean sample was pooled into a skirted V-bottom 96-well plate.

The final representation of DNA for the genotyping for *P. dumosus* and its close relatives consisted of 186 samples and 63 genotypes (Appendix III, C1) distributed into two 96-well plates. Plate 1 contained wild types and plate 2 contained primarily landrace and cultivar types. Both plates were sent to DArT P/L (<http://www.diversityarrays.com/index.html>) for array development and genotyping.

4.2.2 Parameters of the DArT array constructed for genotyping of the GP2

Figure 4.2 gives an illustration of the steps involved during the array construction and samples characterisation. Genomic libraries were built for each species representation and the *PstI/BstNI* system was used as a method of reduction of complexity. Libraries were made in parallel with the construction of the libraries for the DArT array of the common bean [176,181]. Table 4.1 presents the number of clones obtained for the

GP2 array. The genotyping was conducted by hybridising the representative DNA to DArT array, 4208 markers were identify as polymorphic.

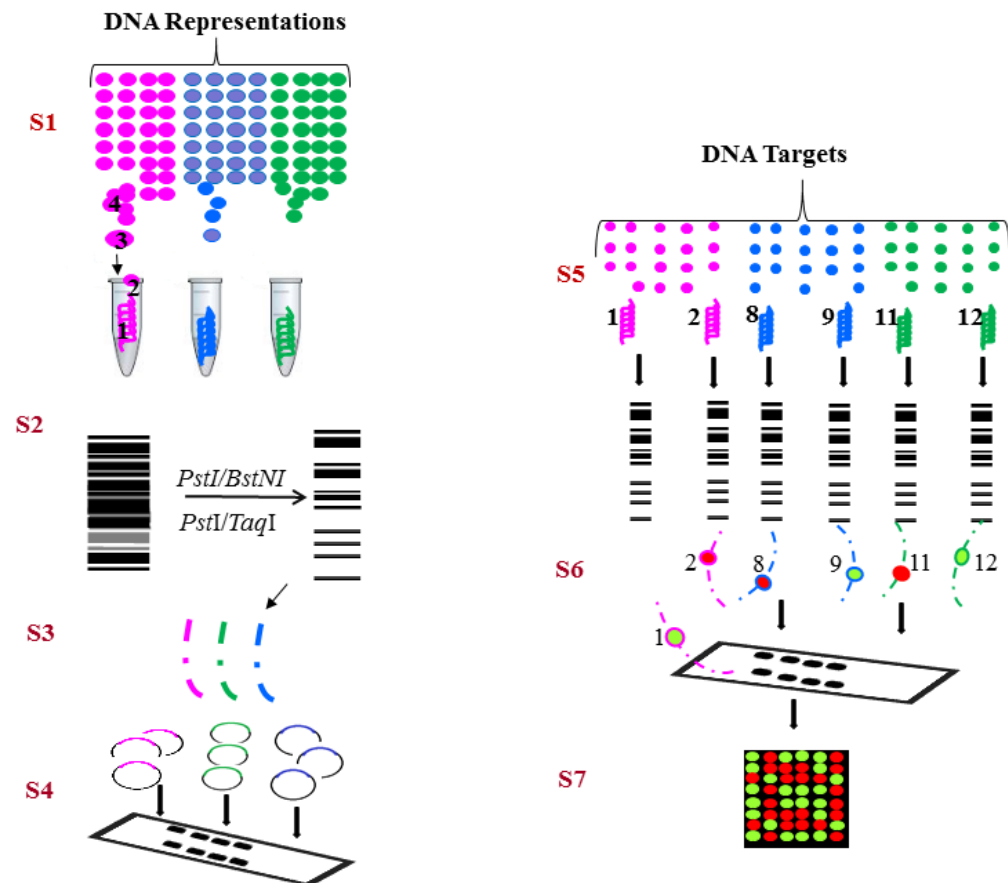


Figure 4.2 Illustration of methods for array construction and characterization with DArT. (S1) DNA representations of the bean species were tested for QA and pooled together. (S2) Genome complexity reduction; the pooled DNA was cut with a combination of *PstI/TaqI* and *PstI/BstNI* and ligated to a *PstI* adapter, the variable fragments generated were amplified by a selective PCR. (S3) The fragments from each representation were cloned and the inserts were amplified by PCR. (S4) The amplified products from cloning, were dried and spotted onto glass slides that constituted the array. (S5) A second set of DNA from each sample to characterise, was reduced using the same complexity reduction method used to generate the array. (S6) After reduction, each fragment generated for each sample, was labelled with a green or red fluorescent dye and hybridised to the array slides, (S7) Slides were scanned to detect fluorescent signal emitted from the hybridised fragments. (Figure is based on Jaccoud et al. [111]).

Table 4.1 Libraries of the GP2 available in DArT technology

Specie	Method of reduction	Number of clones generated
<i>P. vulgaris</i>	PstI/BstNI	2304
<i>P. coccineus</i>	PstI/BstNI	1536
<i>P. dumosus</i>	PstI/BstNI	1536
<i>P. costaricensis</i>	PstI/BstNI	768
<i>P. albescens</i>	PstI/BstNI	768
<i>P. lunatus</i>	PstI/BstNI	768

The report generated from the DArT genotyping provides information on present, absent, missing or ambiguous characteristics for each analysed genotype. It also includes information about the quality and accuracy of the data in the array. This information was used to purge the data of ambiguous or unclear markers, prior to statistical and genetic structure analysis.

4.2.2.1 DArT Array quality analysis

The performance of the DArT markers in wild bean genotyping was measured by the values of the Call Rate, the Discordance, the P-value, and the Polymorphism Information Content. These are parameters that provide information about the polymorphism, consistency and reliability displayed by each marker in the array. Each parameter and the margins that were used are described below.

The Call Rate of a marker represents the percentage of samples that were consistently scored by the marker; it is a measure of reproducibility that indicates the consistency

of a marker. Complementarily, the discordance indicates the reproducibility of the score as evaluated by technical replicates. As Quality Control (QC) within the *P. dumosus* and close relatives array, only markers that displayed call rate >95% and discordance = 0 were selected for further analysis. The P-value in the DArT array is based on an analysis of variance (ANOVA). It represents the discriminatory power of the marker; P-value indicates how well the marker can characterise the samples by separating presences and absences. A P-value >0.75 was used as QC criteria for selection of markers within the GP2 array.

The PIC was used to explore the overall polymorphism and distribution of variation of the markers in the *P. dumosus* array. PIC measures the distribution of the scores of a marker. With a maximum value of 0.5 (when a marker scores 50% presence and 50% absence), the PIC value indicates how informative are the markers in the array.

4.2.3 Data analysis

4.2.3.1 Analysis of genetic diversity of *P. dumosus*

The DArT genotyping outputs a matrix of presence/absence. This matrix was used to investigate the genetic structure between the genotypes of *P. dumosus*, *P. coccineus* and *P. vulgaris*. Overall the aim of the following analysis was to research the genetic diversity among *P. dumosus* genotypes. Additionally an ancestry analysis was performed, to evaluate if genome-wide polymorphism shows evidence of past hybridisation events between *P. dumosus* and its closest relatives. Relationships

between species and accessions were investigated by two clustering methods: model-based clustering and genetic distance-based clustering.

Genetic distance based clustering: The Genetic distance between samples was represented by Principal Coordinates analysis (PCoA). This is a multivariate analysis that finds and plots the major patterns within a multivariate data set (e.g. data from DArT array with multiple loci and multiple samples). The procedure aims to reveal essential patterns of the variation and the genetic relationship among the data, without altering the data itself. The analysis synthesizes the data from a mass of variables into a set of compound axes, the first axis will explain most of the variation, then the second and so on [182]. The PCoA analysis was used to outline the genetic structure within and between species of bean.

Using the bean genotyping, a pairwise Euclidian genetic distance (GD) matrix was calculated. The GD matrix was used as the input data for the PCoA analysis. Both the GD matrix and the PCoA were calculated using the program GenAlEx 6.5[183] using the option for covariance matrix with data standardisation. The calculated GD matrix from the DArT genotyping was also used to perform an analysis of molecular variance (AMOVA) among and within bean species. Within the *P. dumosus* genotypes AMOVA analyses were performed using GenAlEx. This program calculates an analogue of F_{ST} (as explained in Chapter 2) called Φ_{PT} which allows estimation of population genetic differentiation when data from dominant markers are used, such being the case for DArT markers. An estimation of the molecular variance among and

within all species was also performed using the program Arlequin 3.5 [123] which allows the hierarchical organisation of the samples.

4.2.3.2 Analysis of ancestry and relationship of wild *P. dumosus* and the common bean secondary gene pool assessed by DArT markers

The DArT genotyping data was used to outline the relationships among *P. dumosus*, *P. coccineus* and *P. vulgaris*. For this purpose genetic distances were obtained by computing the Jaccard dissimilarity [184]. A dendrogram was constructed using the Weighted Neighbor-Joining method implemented in the software DARwin 5 [185], in this method as in neighbor joining, two taxa are joined in each iteration. However, the weighted criterion for choosing a pair of taxa to join takes into account that errors in distance estimates [126,186]. The significance of each node was evaluated by bootstrapping with 1000 replications.

Since the relationship between *P. dumosus* and its close relatives has been identified as complex, it is better represented by a phylogenetic network than by a traditional tree diagram (Chapter 3). To unveil conflicting phylogenetic signals potentially caused by reticulation events, a phylogenetic network was made using the genetic distance matrix for the wild genotypes. The network was generated in Splits Tree 4 [187] using the NeighbourNet uncorrected distances and the Equal Angle splits transformation method. The fitting of the network to the data is given by the Least Square Fit value (LS fit), which indicates how adequately the network represents the phylogenetic signal of the data.

Model based clustering analysis was carried out using the program STRUCTURE version 2.3.4 [188,189]. The program executes a Bayesian analysis of genetic structure, applying a Markov chain Monte Carlo (MCMC) algorithm (see Chapter 3 for further explanation). Overall, the analysis represents the genetic subdivision found among the data by assigning individuals to a cluster or partition.

The number of partitions or populations (K) is specified in the model: each population is characterized by a set of allele frequencies at each locus and do not have any pre-label. The model assumes that within populations the loci are at Hardy-Weinberg equilibrium, hence samples are assigned to a population in order to maximise the equilibrium. Each sample is assigned to a population; samples with admixed genotypes are assigned to more than one population. The simulation also provides values of the estimated membership or ancestry (Q) for each sample. For the analysis of wild beans, an admixture model was used, as this model allows individuals to have ancestry from different populations. Allelic frequencies were set as correlated among populations. K was set from one to ten and the simulation was run with a hundred thousand iterations for the starting period and one million iterations for the MCMC.

The test for the optimal K , Evanno test Δk [190] and the probability for the best K [188,189] were applied to infer the optimal K value. This determines the K that maximizes the estimated log-likelihood ($\ln(\Pr(X|K))$) of the data between successive runs. Both tests were made with the programs CLUMPP [191] and Structure Harvester [192]. A plot was generated representing the estimated K and Q values for each sample. In the plot, each K is represented by a colour and each sample in the data is

representing by a vertical line. Each line is colour coded according to the Q fraction of the sample for each inferred K cluster. It is expected that the bean analysis will show clusters in which most of the samples from one species group together. But also it is expected to observe patterns of admixture between clusters, these are mixed genotypes (represented by lines with more than one colour in the plot), representing an admixed ancestry.

4.4 Results

4.4.1 DArT markers generated for the screening of the common bean GP2

A total of 4208 polymorphic markers were detected among the clones of the common bean secondary gene pool. The average call rate (reliability of the scores) was 97.76% for Plate 1 and 97.97% for plate 2. The vast majority (99%) of the markers presented a frequency higher than five percent. For genotyping, the initial set of 186 samples was reduced to 152 since some samples failed on DNA quality. A matrix was built representing the presence vs absence of the marker in each sample. After applying the QC criteria, 742 markers were selected for genetic diversity analysis.

4.4.2 Genetic diversity of a germplasm collection of wild *P. dumosus* assessed by nuclear markers

The analysis of the polymorphism and genetic diversity was carried out with data of the 742 markers selected by applying the QC criteria. Table 4.2 present the average diversity for *P. dumosus*, *P. coccineus* and *P. vulgaris* based on the 742 DArT markers.

Table 4.2 Average gene diversity of *P. dumosus*, *P. coccineus* and *P. vulgaris* based on 742 polymorphic DArT markers.

Popuation	N	Na	Ne	I	h	% P
<i>P. coccineus</i>	50	1.862±0.014	1.330±0.010	0.353±0.007	0.218±0.005	0.885
<i>P. dumosus</i>	43	0.574±0.026	1.048±0.006	0.0461±0.005	0.029±0.004	0.132
<i>P.vulgaris</i>	35	1.462±0.018	1.235±0.012	0.2192±0.010	0.142±0.007	0.464

N = Number of samples

Na = Number of different alleles

Ne = Number of Effective Alleles = $1 / (p^2 + q^2)$

I = Shannon's Information Index = $-1 * (p * \ln(p) + q * \ln(q))$

h = Diversity = $1 - (p^2 + q^2)$, for haploid binary data p = Band Freq. and q = 1 - p.

P=Percentage of polymorphic loci.

Showing average± Standard deviation.

Figure 4.3 represents the PCoA plot showing pairwise genetic diversity within the accessions of *P. dumosus*. The variance of the three principal coordinates accounted for 72.10% of the total variation (49.34%, 15.24%, and 7.52% respectively). This variation generated a three group distribution that separates the wild accession G35758-Sacatepequez on one side of the plot and wild G36286 and landraces PI195389 and PI317563 on the other side. All the other accessions were grouped in a central position. The results from PCoA were complemented and re-confirmed by the

results of molecular variation analysis. The AMOVA (Appendix III, C2) showed a low level (15%) and low significance ($F_{ST}=0.15385$, $P=0.01271\pm0.00338$) for the variation between categories wild, landrace and cultivar. It indicates that most (85%) of the variation in *P. dumosus* corresponds to polymorphisms present within the categories.

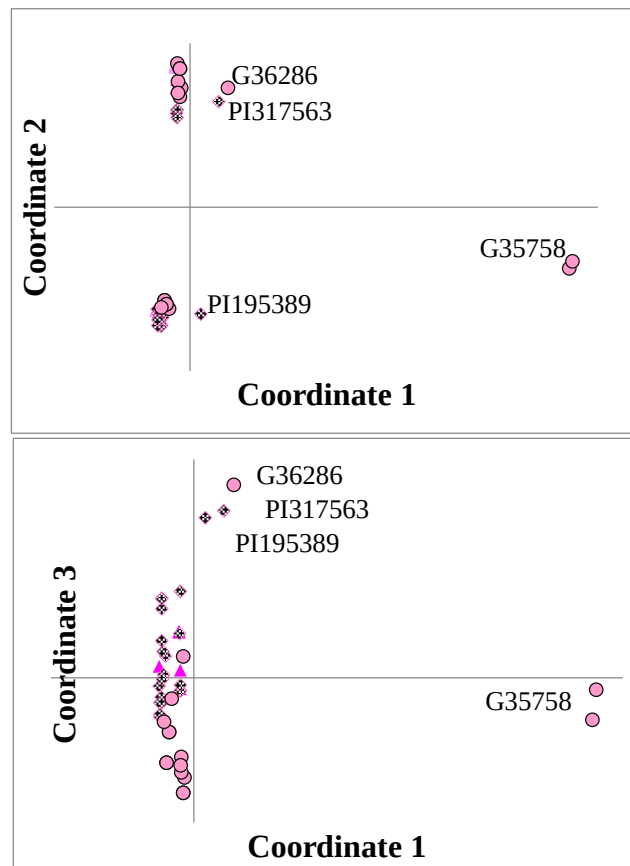


Figure 4.3 Principal Coordinates analysis of 742 DArT markers for *P. dumosus*. Showing 72.10% of variation found among the first three components (49.34%, 15.24%, and 7.52% respectively)

● = wild accessions, ◆ = landraces, ▲ = cultivars.

4.4.3 Analysis of relationships and ancestry of *P. dumosus* using whole genome DArT markers

Whole genome genotyping made with DArT markers was used to evaluate the genetic relationships among the species *P. dumosus*, *P. vulgaris* and *P. coccineus*. A pairwise genetic distance was calculated using the DArT scores for *P. dumosus*, *P. coccineus* and *P. vulgaris*. It included wild, landrace and cultivar genotypes for each species. Wild genotypes for the related species *P. costaricensis*, *P. albescens* and *P. acutifolius* were also included. The PCoA analysis generated from the genetic distance matrix shows 72.52% of the total variation explained by the first three coordinates (47.81%, 20.85%, and 3.86%, respectively). The plot in Figure 4.4 shows the variation distributed in the fourth main cluster; basically a group for each species and a fourth group formed by the hybrids and the cultivars of *P. vulgaris*. The accessions of *P. costaricensis* were beside the *P. dumosus* group and *P. albescens* were very near to the group of *P. coccineus*.

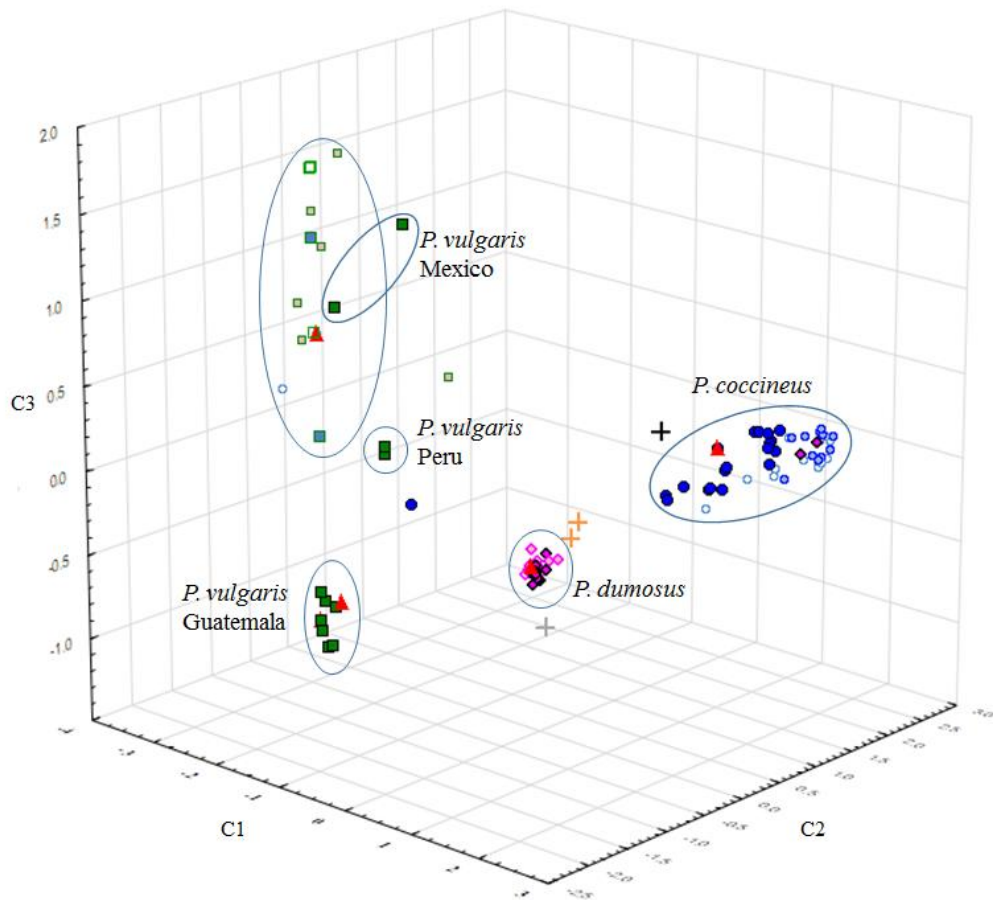


Figure 4.4 Plot from principal coordinates analysis, showing the genetic distance among species of common bean GP2 using 704 DArT markers. The plot represents 72.52% of the molecular variation in coordinate 1 (C1) 47.81%, coordinate 2 (C2) 20.85%, and coordinate 3 (C3) 3.86%.

P. dumosus: wild (◆) landrace (◇) cultivar (◇)

P. coccineus: wild (●) landrace (●) cultivar (○)

P. vulgaris: wild (■) cultivar (□)

P. costaricensis (+)

P. albescens (+)

P. acutifolius (+)

Hybrids: *P. vulgaris* x *P. dumosus* (□), *P. vulgaris* x *P. coccineus* (■)

Controls (▲)

There were no accessions of *P. dumosus* clustering with the *P. vulgaris* group. Differences in the distributions of the cultivar and wild genotypes were observed for *P. coccineus* and *P. vulgaris* but not for *P. dumosus*, for which the two genotypes were more tightly grouped in one cluster, as expected from the diversity analysis. The separation between Mesoamerican and South America genotypes was also observed only among *P. vulgaris* genotypes, with the wild genotypes from Peru separated from those from Mexico and Guatemala.

4.4.3.1 Structure analysis

In accordance with the probability for the best K, using the median values of $\log_e$ of likelihood value $L(k)$, the K with the highest probability to best represent the data was $K=8$ (Appendix III, C3). According with the Δk (Evanno test) there is a bimodal distribution of delta Δk , with a first peak when $K=6$ and a second peak when $K=8$ (Appendix III, C4). This bimodal K indicates that there is a secondary structure among the data from *P. vulgaris* and *P. coccineus*. It also indicates that the optimal number of partitions is perhaps higher than $K=8$. However the intention of this research was not to estimate the real K but rather search for evidence of admixture among the *P. dumosus* data. Therefore, the focus from here is on the ancestry values found along the runs, Figure 4.5. In the ancestry analysis that included all species of GP2, when K was set to two populations, all the accessions of *P. dumosus* where assigned to a single cluster. This cluster also includes *P. coccineus*, *P. albescens* and *P. costaricensis*. The second cluster included all representatives from *P. vulgaris*. For $K=3$ the analysis

separates all but two accessions of *P. dumosus* into its own cluster, the second and third cluster were formed by *P. coccineus* and *P. vulgaris* respectively. At K=4 the simulation added the cluster that separated *P. acutifolius* into its own cluster, and from K=5 to K=10 the separations corresponded to substructures within *P. coccineus* and *P. vulgaris* (Figure 4.5). A secondary analysis made with only *P. dumosus*, *P. vulgaris* and *P. coccineus* showed similar results for K=2 and K=3.

The overall analysis shows strong separation between species, the cultivated hybrids were also differentiated. There was no admixed within *P. dumosus* samples, *P. coccineus* was the group with higher admixture, with 14% of the samples showing some level of admix. Interestingly, some of the admixture is with *P. dumosus*. Among *P. vulgaris*, 3% of the samples were admixed. Additionally, for *P. vulgaris*, the structure analyses differentiated among and between genotypes from Guatemala and the genotypes from Mexico. For *P. vulgaris* the analysis also distinguished the genotypes from Peru (G21244). Along the STRUCTURE runs the alpha values were constantly close to zero which indicates little admixture and strong structure among the data.

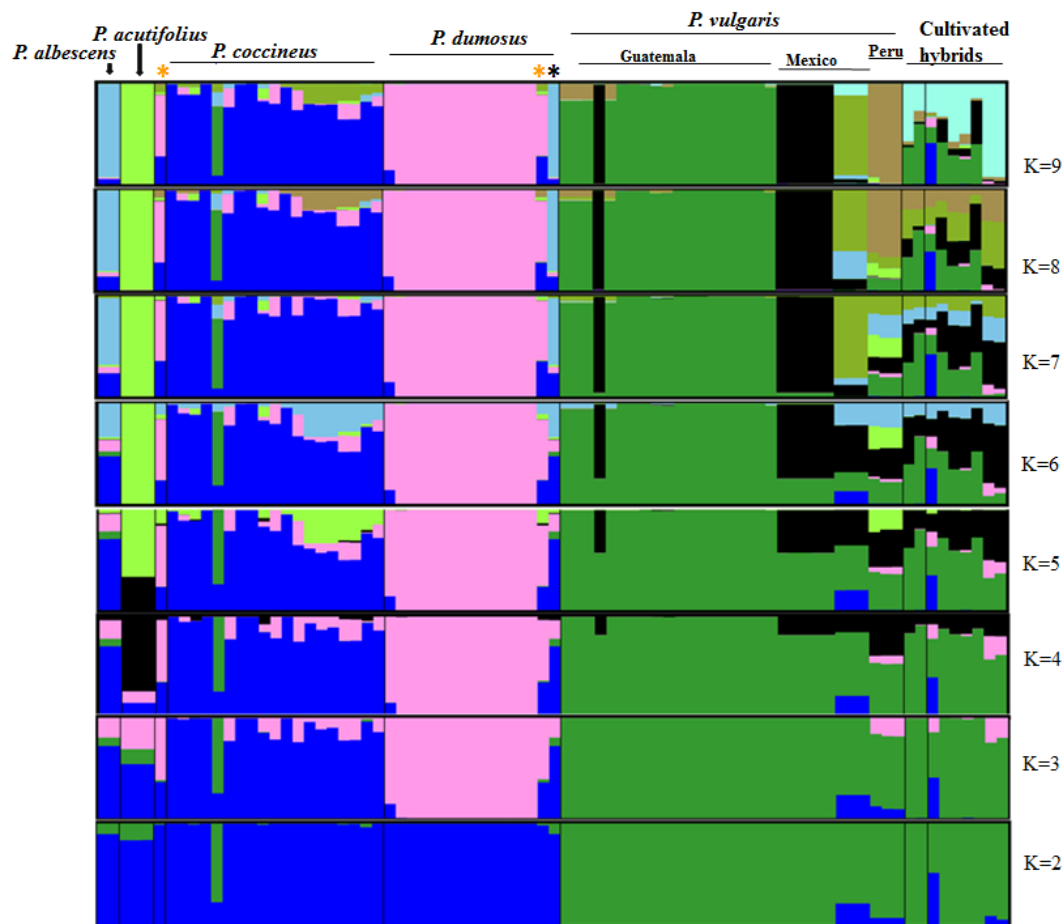


Figure 4.5 Structure analysis of wild bean accessions. The level of ancestry was inferred by a Bayesian model, each colour represents one population, each accession is represented by a vertical bar, and the length of each coloured segment in each vertical bar represents the proportion contributed by ancestral populations. *=*P. costaricensis*, **=*P. albescens*

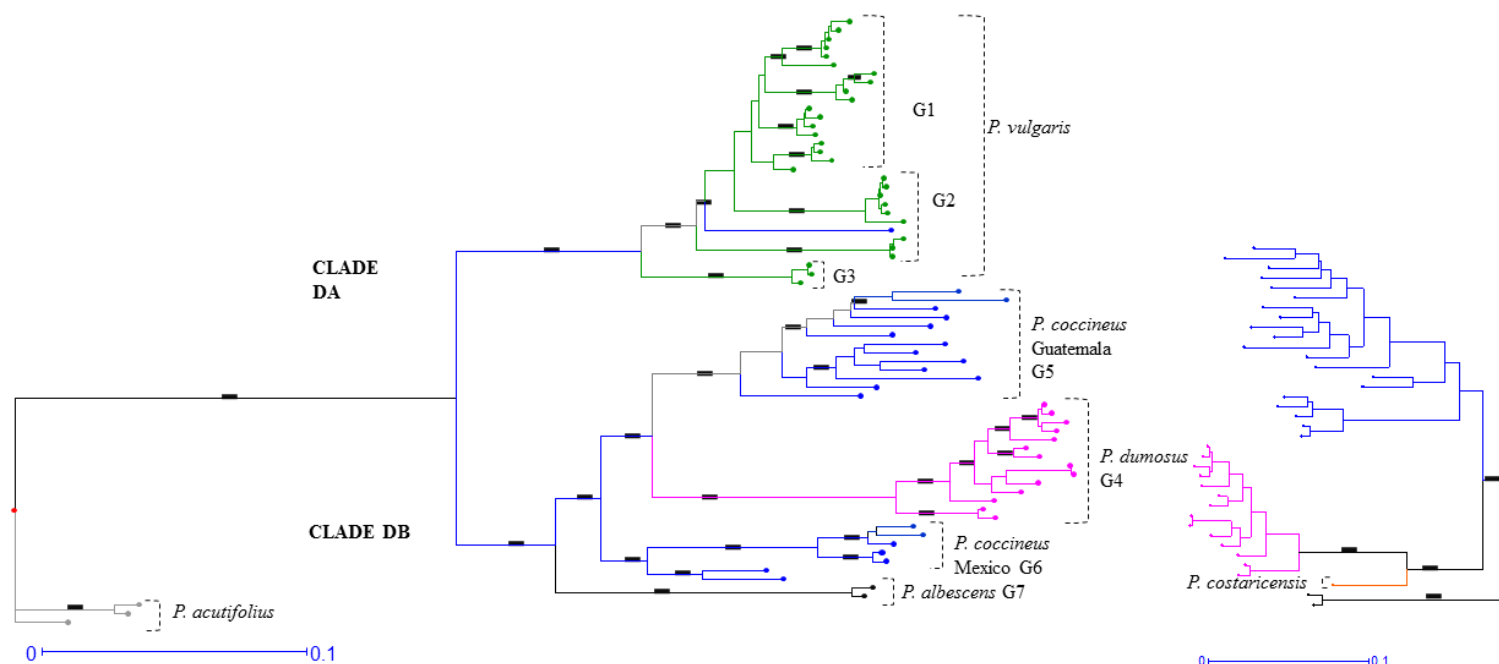


Figure 4.7A

Figure 4.7B

Figure 4.6 Weighted Neighbor-Joining Dendrogram for wild germplasm accessions of the common bean secondary gene pool. The tree was outlined from Jaccard's genetic distance calculated for polymorphisms in 4,208 DArT markers. (A) The two major clades (DA & DB), and their internal subclades, are shown as follows: *P. vulgaris* (green lines) from Guatemala (G1), Mexico (G2) Peru (G3); *P. dumosus* (pink lines) (G4); *P. coccineus* (blue lines) from Guatemala (G5), Mexico (G6) and *P. albescens* (black lines) (G7). Out-group: *P. acutifolius* (grey lines). Dark bars indicate branches supported by bootstrap values greater than 90%. (B) Topology of the clade DB including *P. costaricensis* (Orange line). *Cultivated hybrids were not included in this analysis.

The phylogenetic analysis complemented the findings in the PCoA and the ancestry analysis; the cladogram in Figure 4.6 shows the major relationships represented by two clades (clade DA and clade DB), both of which were strongly supported (BS=100%). Representatives of *P. acutifolius* were used as the out-group clade.

The clade DA was formed by the accessions of *P. vulgaris*, the clade DB groups the representatives of *P. dumosus*, *P. coccineus* and *P. albescens*. The topology shows *P. dumosus* to be within *P. coccineus* which could suggest that *P. dumosus* derived from *P. coccineus* Guatemala with no involvement from *P. coccineus* Mexico or *P. vulgaris*. However it was observed that the inclusion of *P. costaricensis* (accession G40604) generated a different topology (figure 4.6 B) in which *P. dumosus* separates from *P. coccineus* clustering together with *P. costaricensis*.

The phylogenetic signal of the data was well represented by the NeighborNet (LS fit =99.83%), as shown in Figure 4.7. The analysis based on the 4,208 DArT markers outlined two main clusters: one cluster formed by *P. vulgaris* accessions; and a second formed by *P. dumosus*, *P. coccineus*, *P. costaricensis* and *P. albescens*. Within this cluster there are two levels of sub-clusters: one representing each species and one representing groups of accessions within each species. Conflicting signals or reticulations (represented by “boxes” on the base of the nodes) were detected both between and within the clusters.

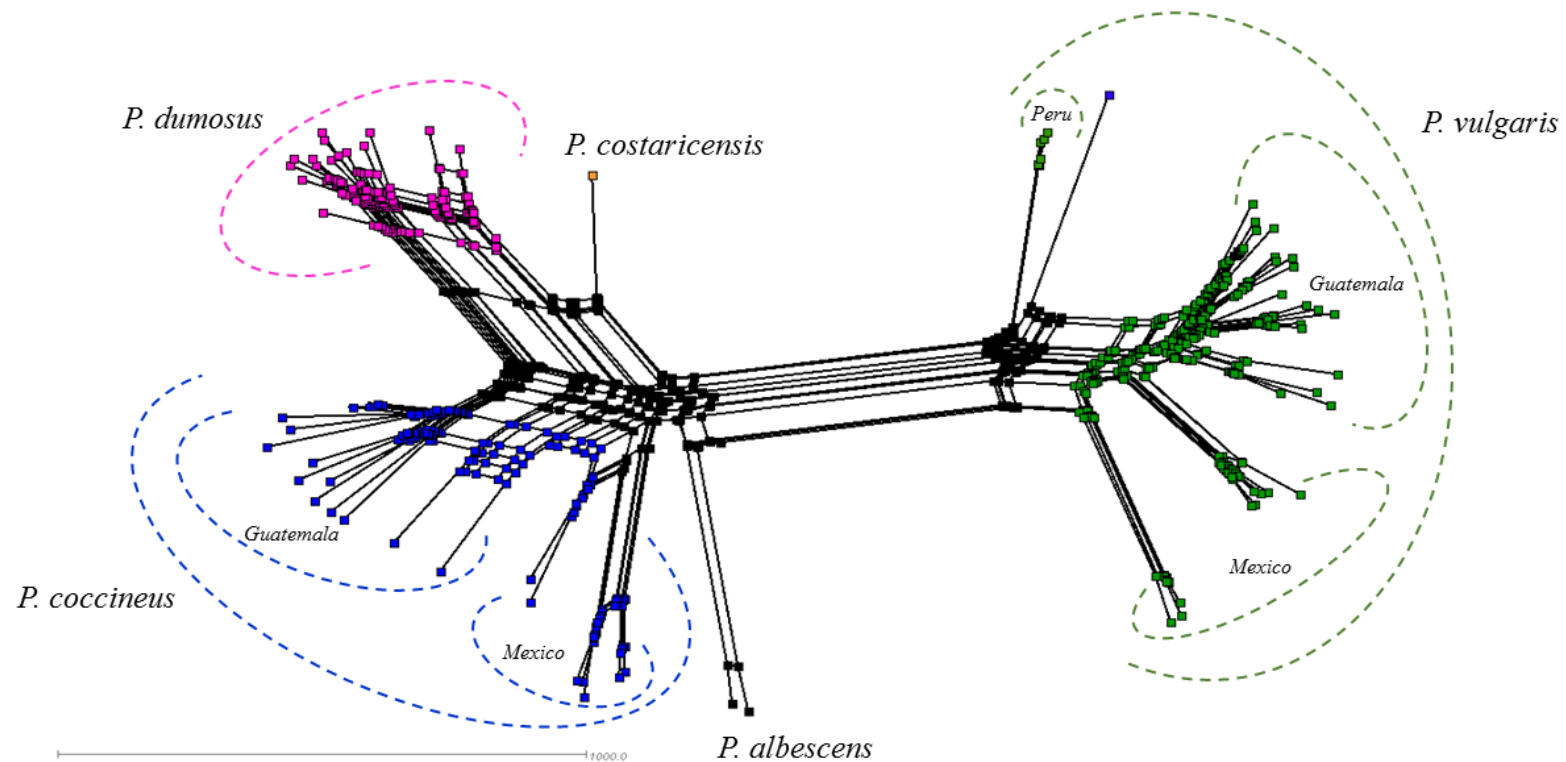


Figure 4.7 Unrooted network for wild germplasm accessions of the common bean secondary gene pool. The network was produced from genetic dissimilarity, applying the Neighbour Net algorithm. *cultivar hybrids were no included in the analysis.

4. 5 Discussion.

4.5.1 Whole genome molecular markers reveal a similar level of diversity for *P. dumosus* compared to diversity inferred by cpDNA and ITS

A main objective of this thesis research has been to assess the genetic diversity among a germplasm collection of wild *P. dumosus*, in order to establish the variation of the collection and generate information for breeding and conservation programmes. In this chapter whole-genome DArT molecular markers were developed and evaluated, with the objective of profiling a germplasm collection of wild *P. dumosus* and other wild representatives of the common bean GP2. After the QC, the number of polymorphic markers found in the array are comparable to those found in hop (*Humulus lupulus* L) DArT array for (730 markers) [193].

Using the polymorphisms from DArT markers the analysis performed through this chapter found lower genetic diversity of *P. dumosus* compared to the genetic diversity of *P. coccineus* and *P. vulgaris* (average gene diversity = 0.029, 0.219, and 0.142 respectively) likewise genetic structure within *P. dumosus* was also the least, the majority of the samples were grouped in a single cluster (figures 4.4, 4.5). These results, are in agreement with the results found by Soleri et al. [133] where an evaluation of nuclear and chloroplast SSRs among farmers varieties of *P. coccineus*, *P. dumosus* and *P. vulgaris* found that *P. dumosus* was the least diverse of the three species. Differences in habitat, breeding system, domestication process and historical

dispersion are aspects associated with the differences in genetic diversity and structure of these three species. While wild *P. coccineus* and *P. vulgaris* are widely distributed in Central America (for *P. vulgaris* also South America), wild *P. dumosus* grows in a more restricted area only in Guatemala. It is well known that *P. vulgaris* was domesticated the earliest and has undergone intense distribution, at least two areas of high diversification are point to the centre of origin for this species and two domestication centres had being identified [34,61,194]. *P. coccineus* is known for be a species with a broad diversity and habitat plasticity, capable of colonizing xeric environments [32,51]. It is widely cultivated in Central America, South America and Europe. As mentioned in Chapter 1, *P. dumosus* is considered to be a more recent specie, being less cultivated and less dispersed than *P. coccineus* and *P. vulgaris*. It is possible that intraspecific diversification has not arisen yet for this species as may have occurred for *P. coccineus* and *P. vulgaris*. Cultivar *P. dumosus* is mainly grown in highlands and temperate locations. In fact farmers recognize its seeds, as a "cold land bean" [74,76,133]. Farmers preferences for either *P. vulgaris* or *P. coccineus* [76] may have also contribute to a less diversification of landraces and cultivars of *P. dumosus*. The mating system is another differential aspect between this species, *P. vulgaris* is mainly self-crossings, while *P. coccineus* and *P. dumosus* are outcrossing species. Outcrossing may have contribute to keep *P. dumosus* diversity despite its restricted habitat, and I point this out because considering the sample size and distribution of the sampling used in this analysis, *P. dumosus* shows a moderate level of diversity.

Considering the sample size and number of markers used in other studies with other species that have displayed low levels of diversity with DArT markers e.g. pigeonpea (*Cajanus cajan* Mill) [196], *Cicer* spp. [197], and hop (*Humulus lupulus* L.) [193] *P. dumosus* diversity is fairly high.

Considering the habitat conditions in which *P. dumosus* grows it is clear that this is a highly valuable genetic resource for the improvement of common bean. Following this research, an analysis of genetic diversity of in-situ populations of *P. dumosus* using DArT markers may be a good step to further characterise this specie as a genetic resource for bean improvement. Currently the germplasm collection of wild *P. dumosus* includes only nine accessions, these accessions represent a valuable source of diversity, efforts should be made to increase the numbers of wild representatives of *P. dumosus*. This research underpins the need for new field explorations to look for remaining wild populations and extensions of the *P. dumosus* germplasm collection.

4.5.2 Does the genome-wide analysis give evidence of hybridisation of *P. dumosus*?

Beyond the analysis of diversity of wild *P. dumosus*, another objective of this research has been to confirm previously outlined patterns of relationships, and generate new information that contributes to a better understanding of species boundaries among the common bean GP2. Previous analysis from ITS sequences (Chapter 3) had found that *P. dumosus* is closer to *P. coccineus* than to *P. vulgaris*. However, an analysis that

included wild representations of the three species and multiple loci on the nuclear genome was necessary to provide assurance that the observed pattern is associated with the species history and not with the locus history. Due to the characteristic of being generated from a whole genome microarray, the variation detected by the DArT markers can be considered as combining polymorphisms derived from a whole genome screen and so provides information on the variation of more than seven hundred loci. Likewise, the signal from the data is strongly confirmed by the consistency of results when several analytical approaches were used.

In this chapter, the genome profiling of wild beans made with DArT markers robustly confirmed the signal from the nuclear genome of *P. dumosus*. The structure and the PCoA analysis clearly differentiated each species in a different cluster generating a similar patterns of relationships as the one outlined by the ITS locus, with *P. dumosus* closer to *P. coccineus* than to *P. vulgaris*.

The inclusion or exclusion of *P. costaricensis* generates a conflicting signal in the cladogram (Figure 4.6), which suggest that *P. dumosus* derived from *P. coccineus* Guatemala with little or no involvement of *P. coccineus* Mexico, or *P. vulgaris*. The outline of relationships is also seen in the network (Figure 4.7), this pattern of relationships is similar to the one found by Piñero and Eguiarte [81] which led them to conclude that *P. dumosus* was a variety of *P. coccineus*. The inclusion of *P. costaricensis* in the analysis decreases the bootstrap values of the *P. coccineus* node (Figure 4.6B) and separates *P. coccineus* in one branch and *P. costaricensis* and *P. dumosus* in another branch. This is in agreement with the results shown by Delgado-

Salinas et al. [73]. This research confirms that *P. costaricensis* has likely been involved in the evolution or backcrossing events (or both) in the history of *P. dumosus*. Overall the findings in this chapter support the hypothesis of a reticulate origin for *P. dumosus*, a hybridisation event that involved *P. coccineus*, *P. vulgaris* and perhaps backcrossing events to *P. coccineus* (as pollen donor), and in which *P. costaricensis* was also involved, as proposed by Llaca et al. [82] and Schmit et al. [56].

Overall, the dense net-like structure in the network displayed a considerable number of interactions within and between species of GP2. These findings indicate the density and complexity of the interactions among species of the bean GP2.

5 Conclusions and future directions

5.1 Diversity and hybrid origin of *P. dumosus*

The origin and diversity of the year-long bean, *Phaseolus dumosus* has been the subject of study within this PhD thesis research. The fine scale analysis applied in this study permitted the maximisation of the differences among the samples helping to better outline the variation among accessions and species. In this analysis a moderate level of genetic diversity and low differentiation was found within the germplasm collections of wild *P. dumosus*. This indicates the valuable of the collection not only as a representative of the bean variation but also as a source for genetic improvement of common bean. Since the actual germplasm samples were collected almost four decades ago, it should be investigated whether the diversity observed across wild, cultivated and landrace genotypes of *P. dumosus* still present in-situ populations or whether has already disappeared due to habitat erosion. It is recommended that new systematic field collections be carried out to increase the wild germplasm representatives of this species and also to assess the extent of *in-situ* diversity across the range of the species.

Resolving relationships among closely related and recently diverged lineages, can represent a challenge for the finding variable markers with the capacity to differentiate at lower structural levels. There is the additional challenge of associating the biological explanation to situations of incongruences between phylogenies, since several different evolutionary processes can generate topological discordance among gene trees. Hybridisation is among these evolutionary processes.

Hybridisation from discordance between phylogenies from different genomes or genes requires further biological support. To clarify the origin of *P. dumosus* this study contributes new evidence that reinforces the hypothesis of hybridisation. By using genome-wide DArT molecular markers, sequences of ITS and sequences of several cpDNA loci, this study confirms the conflicting phylogenetic signal between the cpDNA and the nrDNA of *P. dumosus*. The findings from the analysis carried out here strongly support the hybridisation hypothesis. Other evolutionary processes as incomplete lineage sorting could generate a similar pattern of relationships and conflicting signals in phylogenetic analysis. However lineage sorting or ancestral variation could explain the patterns observed at the cpDNA of *P. dumosus* but would not explain the extensive markers shared at the nuclear genome between *P. dumosus* and *P. coccineus*.

The lack of shared polymorphism between *P. vulgaris* and *P. dumosus* in the nuclear genome is in agreement with the hypothesis that concerted evolution and genome homogenisation may have occurred after hybridisation. Further evidence for hybridisation was given by the intermedia position of *P. dumosus* in the PCoA analysis. All these evidences are supported by the morphology of *P. dumosus* that displayed characters associated to *P. coccineus*: semi tuberous root, larger seed and predominantly outcrossing and with *P. vulgaris*: epigeal cotyledons, introrse stigma, as well as a growing habits.

It is possible that backcrossing to *P. coccineus* may have also occurred in a lower proportion during the speciation of *P. dumosus*. The outline of relationships by network, leads to the interpretation that *P. dumosus* originated from a hybridisation

event between *P. coccineus* and an early taxon of the *P. vulgaris* clade, and the position of *P. costaricensis* suggests the possibility of this species acting as a mediator in backcrossing.

5.2 Remarks on the common bean secondary gene pool

The analysis performed here detected the existence of ancestral cpDNA haplotypes in all species of the GP2-secondary gene pool (Chapter 2), and also verified that *P. dumosus* is the closest species to *P. vulgaris*.

This research found that *P. persistentus* is more widely divergent from *P. leptostachyus* than any other species tested. The molecular analysis carried out here showed no differentiation from *P. vulgaris*. Hence, based on molecular analysis, it is likely that *P. persistentus* is a morphologically-divergent accession of *P. vulgaris*.

This research also confirmed the wide diversity of *P. coccineus* with separations in the clade still ongoing, and many genetic differences between *P. coccineus* accessions were observed. Further novel traits are therefore likely to be present among unregistered populations and may be of value in future breeding programs if they can be collected. Among accessions of *P. costaricensis* and *P. albescens* there was also variation; both species appear to be closer to *P. coccineus* than to *P. vulgaris*. As was recommended in the case of *P. dumosus*, analysis of *in-situ* diversity (and efforts to

increase germplasm representation) should be considered for these species. This would allow deeper analysis of genetic diversity to be carried out on both species.

This research confirms that wild types of the GP2 carry significant genetic variation which make of these species a valuable genetic resource for common bean improvement and for food security. Based on the close relationship among these species is clear that strategies of conservation for common bean may have to include the species of the GP2. The gap analysis from the GP2 of common bean identifies more than 98% of wild relatives missing in germplasm representations [41]. Recent germplasm collections carried out in Mexico and Costa Rica have reported strong reduction in growing areas and populations of wild bean [45, 69]. It is possible that diversity that is represented in germplasm collection of this group is not represented in the in-situ populations any longer, but also many areas of potential populations of wild beans having yet been explored.

In this research, I have conducted a first analysis of wild cultivar and landraces types of *P. dumosus* using whole genome markers, and I have generated a platform of molecular markers to specifically evaluate the species of the GP2. A first Analysis that should proceed to this research is a more extensive evaluation of cultivar and landraces of *P. dumosus* from Central and South America including recent collected germplasm from Costa Rica [198], this could be a good starting point to know the genetic diversity of current populations.

5.3 Contribution to bean breeding and future directions

This research has contributed to generating a platform to evaluate thousands of markers among the species that comprise the common bean secondary gene pool.

Aside from the utility of the markers to characterise a wide number of bean accessions in a cost-effective way, a future direction from this research could be the use of the DArT array for identification of genes associated with species barriers. This could be done by identifying the clones that showed the greatest variation between species of the GP2. Clones that showed a constant variation between species but were conservative within species, could be a good starting point to identify sequences associated with reproductive isolation, without the need for generating *de novo* sequencing for each species. For example sequences of these clones can be compared with the sequence of the common bean genome [35] as a starting point for gene discovery.

The utility of DArT markers developed for breeding programs has been demonstrated in routine analysis of gene discovery, screening of breeding lines, and mapping of novel traits, improve covering and enhance existing maps. Examples are in several crops species including oats (*Avena sativa* L.) [199], banana (*Musa* sp.) [200] and carrot (*Daucus carota* L) [201]. Thudi et al. [177] have recently shown the utility of DArT for establishing marker-trait associations (MTAs) for drought and heat tolerance in chickpea.

For bean breeding programs, the DArT array can be used for routine analysis of gene discovery among wild relatives of the common bean, characterisation and screening

of breeding lines and mapping of novel traits e.g. traits for tolerance for abiotic stress in *P. coccineus*. These markers can even be used to evaluate *P. acutifolius* which is specie of interest for improve tolerance to abiotic stress, as it grows in more arid environments. The DArT markers generated in this research can also be used to improve covering and enhance existing maps of *P. vulgaris* and *P. coccineus*.

Finally the DArT array is a direct contribution to bean improvement since the markers generated in this research make it possible to identify variable accessions and admixture among species such as *P. dumosus*, *P. albescens* and *P. costaricensis*, for which a wide pool of markers was not previously available.

References

- 1 Darwin CR. **Variation of animals and plants under domestication.** London: John Murray 1868.
- 2 Charlesworth B, Charlesworth D. **Darwin and genetics.** *Genetics* 2009;**183**:757–66.
- 3 Rasmusson DC, Phillips RL. **Plant breeding progress and genetic diversity from de novo variation and elevated epistasis.** *Crop Sci* 1997;**37**:303–10.
- 4 Nevo E. "1 4 **Genomic diversity in nature and domestication.**" **Plant diversity and evolution:** In: Genotypic and phenotypic variation in higher plants eds Henry R.J. CABI 2005
- 5 Schoen DJ, Brown AH. **Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers.** *Proc Natl Acad Sci* 1993;**90**:10623–7.
- 6 Mohammadi SA, Prasanna BM. **Analysis of genetic diversity in crop plants statistical tools and considerations.** *Crop Sci* 2003;**43**:1235.
- 7 Bermejo J, León Jp. **Neglected crops: 1492 from a different perspective.** Food and Agriculture Organization of the United Nations 1994.
- 8 van de Wouw M, Kik C, van Hintum T, van Treuren R, Visser B. **Genetic erosion in crops: concept, research results and challenges.** *Plant Genet Resour* 2010;**8**:1–15.
- 9 Singh D, Jackson G, Hunter D, Fullerton R, Lebot V, Taylor M, et al. **Taro leaf blight a threat to food security.** *Agriculture* 2012;**2**:182–203.
- 10 Tatum LA. **The southern corn leaf blight epidemic.** *Science* 1971;**171**:1113–6.
- 11 Padmanab.Sy. **Great bengal famine.** *Annu Rev Phytopathol* 1973;**11**:11–26.
- 12 Ploetz RC. **Fusarium wilt of banana is caused by several pathogens referred to as fusarium oxysporum f. sp. cubense.** *Phytopathology* 2006;**96**:653–6.
- 13 Rao VR, Hodgkin T. **Genetic diversity and conservation and utilization of plant genetic resources.** *Plant Cell Tissue Organ Cult* 2002;**68**:1–19.
- 14 Maxted, N., Ford-Lloyd, B. V., Kell, S. P., Iriondo, J. M., Dulloo, M. E., Turok, J. **Crop wild relative conservation and use.** Birmingham: CABI 2007.
- 15 Hajjar R, Hodgkin T. **The use of wild relatives in crop improvement: a survey of developments over the last 20 years.** *Euphytica* 2007;**156**:1–13.

- 17 Spillane C, Gepts P. **Evolutionary and genetic perspectives on the dynamics of crop gene pools.** In: Cooper HD, Spillane C, Hodgkin T (eds), Broadening the genetic base of crop production. CABI 2001;25–70.
- 18 Gepts P. A **Phylogenetic and genomic analysis of crop germplasm: a necessary condition for its rational conservation and use.** In: Gustafson JP, ed. *Genomes*. Springer US 2000;163–81.
- 19 Spataro G, Tiranti B, Arcaleni P, Bellucci E, Attene G, Papa R, et al. **Genetic diversity and structure of a worldwide collection of *Phaseolus coccineus* L.** *Theor Appl Genet* 2011;**122**:1281–91.
- 20 Gepts P, Aragão FJL, Barros E, Blair MW, Brondani R, Broughton W, et al. **Genomics of *Phaseolus* beans, a major source of dietary protein and micronutrients in the tropics.** *Genomics Trop Crop plants* 2008;113–43.
- 21 Jansa J, Bationo A, Frossard E, Rao IM. **Options for improving plant nutrition to increase common bean productivity in Africa fighting poverty in Sub-Saharan Africa: the multiple roles of legumes in integrated soil fertility management.** In: Bationo A, Waswa B, Okeyo JM, Maina F, Kihara J, Mokuwye U, eds. . Springer Netherlands 2011;201–40.
- 22 Ramasawmy S. **Integrating food security information in national statistical systems: experiences, achievements, challenges.** Rome Food and Agriculture Organization of the United Nations 2012.
- 23 Blair MW. **Mineral biofortification strategies for food staples: The example of common bean.** *J Agric Food Chem* 2013;**61**:8287–94.
- 24 **Emergency market mapping and analysis (EMMA), county Wajir, Kenya.** Mercy Corps 2011.
- 25 Beebe S, Rao I, Blair M, Acosta J. **Phenotyping common beans for adaptation to drought.** *Front Physiol* 2013;**4**:35
- 26 Dita MA, Rispaill N, Prats E, Rubiales D, Singh KB. **Biotechnology approaches to overcome biotic and abiotic stress constraints in legumes.** *Euphytica* 2006;**147**:1–24.
- 27 Miklas P, Kelly J, Beebe S, Blair M. **Common bean breeding for resistance against biotic and abiotic stresses: From classical to MAS breeding.** *Euphytica* 2006;**147**:105–31.
- 28 Beaver JS, Osorno JM. Achievements and limitations of contemporary common bean breeding using conventional and molecular approaches. *Euphytica* 2009;**168**:145–75.
- 29 Beebe SE, Rao IM, Devi MJ, Polania J. **Common beans, biodiversity, and multiple stresses: Challenges of drought resistance in tropical soils.** *Crop Pasture Sci* 2014; **65**:667-675
- 30 Singh SP. **Broadening the genetic base of common bean cultivars.** *Crop Sci*

- 2001;**41**:1659–75.
- 31 Hernández-López VM, Vargas-Vázquez MLP, Muruaga-Martínez JS, Hernández-Delgado S, Mayek-Pérez N. **Origen, domesticación y diversificación del frijol común: Avances y perspectivas.** *Rev Fitotec Mex* 2013;**36**:95–104.
 - 32 Porch T, Beaver J, Debouck D, Jackson S, Kelly J, Dempewolf H. **Use of wild relatives and closely related species to adapt common bean to climate change.** *Agronomy* 2013;**3**:433–61.
 - 33 Singh SP, Gepts P, Debouck DG. **Races of common bean (*Phaseolus vulgaris*, Fabaceae).** *Econ Bot* 1991;**45**:379–96.
 - 34 Bitocchi E, Nanni L, Bellucci E, Rossi M, Giardini A, Zeuli PS, et al. **Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data.** *Proc Natl Acad Sci* 2012;**109**:788–96.
 - 35 Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. **A reference genome for common bean and genome-wide analysis of dual domestications.** *Nat Genet* 2014;**46**:707–13.
 - 36 Roorkiwal M, Coyne CJ, Maxted N, Nelson MN, Ambrose MJ, Saxena RK, et al. **Legume crops phylogeny and genetic diversity for science and breeding.** *CRC Crit Rev Plant Sci* 2014;**34**:43–104.
 - 37 Asfaw A1, Blair MW, Almekinders C. **Genetic diversity and population structure of common bean (*Phaseolus vulgaris* L.) landraces from the East African highlands.** *Theor Appl Genet* 2009;**120**:1-12.
 - 38 Beebe S, Idupulapati Rao I, Mukankusi C, Buruchara R. **Improving resource use efficiency and reducing risk of common bean production in Africa, Latin America, and the Caribbean. Eco-efficiency: From vision to reality.** Centro Internacional de Agricultura Tropical (CIAT) 2012;117-134.
 - 39 Baudoin JP, Camarena F, Lobo M. **Improving *Phaseolus* genotypes for multiple cropping systems.** *Euphytica* 1997;**96**:115–23.
 - 40 Shi C, Navabi A, Yu K. **Association mapping of common bacterial blight resistance QTL in Ontario bean breeding populations.** *BMC Plant Biol* 2011;**11**:52.
 - 41 Ramírez-Villegas J, Khoury C, Jarvis A, Debouck DG, Guarino L. **A gap analysis methodology for collecting crop genepools: a case study with *Phaseolus* beans.** *PLoS One* 2010;**5**:10
 - 42 Acosta-Gallegos JA, Kelly JD, Gepts P. **Pre-breeding in common bean and use of genetic diversity from wild germplasm.** *Crop Sci* 2007;**47**:S3: 44 .
 - 43 Sharma S, Upadhyaya HD, Varshney RK, Gowda CLL. **Pre-breeding for diversification of primary gene pool and genetic enhancement of grain legumes.** *Front Plant Sci* 2013;**4**:309.
 - 44 Acosta-Díaz E, Hernández-Torres I, Amador-Ramírez MD. **Collection and**

- characterization of wild species of Phaseolus (Fabaceae) in north eastern Mexico for ex situ conservation.** *Plant Ecology and Evolution* 2015;**148**:119–27.
- 45 Araya Villalobos R, González Ugalde W, Camacho Chacón F, Sánchez Trejos P, Debouck DG, Villalobos RA, et al. **Observations on the geographic distribution, ecology and conservation status of several Phaseolus bean species in Costa Rica.** *Genet Resour Crop Evol* 2001;**48**:221–32.
 - 46 Schwartz HF, Singh SP. **Breeding common bean for resistance to white Mold: a review.** *Crop Sci* 2013;**53**:1832–44.
 - 47 Beaver JS, Perea CGM, Osorno JM, Ferwerda FH, Miklas PN. **Registration of bean golden yellow mosaic virus resistant dry bean germplasm.** *Crop Sci* 2005;**45**:2126–2127.
 - 48 Busogoro JP, Jijakli MH, Lepoivre P. **Identification of a novel source of resistance to angular leaf spot disease of common bean within the secondary gene pool.** *Plant Breed* 1999;**118**:417–23.
 - 49 Mahuku GS, Jara CE, Cajiao C, Beebe S. **Sources of resistance to Colletotrichum lindemuthianum in the secondary gene pool of Phaseolus vulgaris and in crosses of primary and secondary gene pools.** *Plant Dis* 2002;**86**:1383–7.
 - 50 Marquez ML, Terán H, Singh SP. **Selecting common bean with genes of different evolutionary origins for resistance to Xanthomonas campestris pv. phaseoli.** *Crop Sci* 2007;**47**:1367–74.
 - 51 Freytag GF, Debouck DG. **Taxonomy, distribution, and ecology of the genus Phaseolus (leguminosae-papilionoideae) in North America, Mexico and Central America.** Texas: BRIT 2002.
 - 52 Singh S, Schwartz H. Review: **Breeding common bean for resistance to insect pests and nematodes.** *Can J Plant Sci* 2010;**91**:239–50.
 - 53 Singh SP, Teran H, Schwartz HF, Otto K, Lema M. **Introgressing white mold resistance from Phaseolus species of the secondary gene pool into common bean.** *Crop Sci* 2009;**49**:1629.
 - 54 Graham PH, Ranalli P. **Common bean (Phaseolus vulgaris L).** *F. Crop. Res* 1997;**53**:131–46.
 - 55 Miranda Colín S. **Origen de Phaseolus vulgaris L. Frijol Común).** *Agronomia Tropical* 1967;**2**:191–205.
 - 56 Schmit V, Du Jardin P, Boudoin JP, Debouck DG. **Use of chloroplast DNA polymorphisms for the phylogenetic study of seven Phaseolus taxa including P. vulgaris and P. coccineus.** *Theor Appl Genet* 1993;**87**:506–16.
 - 57 Chacón S MI, Pickersgill B, Debouck DG, Arias JS. **Phylogeographic analysis of the chloroplast DNA variation in wild common bean (Phaseolus vulgaris L.) in the Americas.** *Plant Syst Evol* 2007;**266**:175–95.

- 58 Blair MW, Soler A, Cortés AJ. **Diversification and population structure in Common Beans (*Phaseolus vulgaris* L.)**. *PLoS One* 2012;7:e49488.
- 59 Rodriguez M, Rau D, Bitocchi E, Bellucci E, Biagetti E, Carboni A, et al. **Landscape genetics , adaptive diversity and population structure in *Phaseolus vulgaris***. *New Phytol* 2015.
- 60 Science R, American M, America M, America M. **RFLP diversity of common bean (*Phaseolus vulgaris*) in its centres of origin**. *Genome* 1994;37:2:256-63.
- 61 Gepts P. **Origin, domestication, and evolution of the common bean, *Phaseolus vulgaris***. In: CAB, ed. Voysest O, Van Schoonhoven A (eds.) *Common beans: research for crop improvement*. Oxon, UK 1991. 7–53.
- 62 Koenig R, Gepts P. **Alozyme diversity in wild *Phaseolus vulgaris* : further evidence for two major centers of genetic diversity**. *Theor Appl Genet.* 1989;6:809–17.
- 63 Kwak M, Gepts P. **Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae)**. 2009;979–92.
- 64 Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, et al. **Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes**. *New Phytol* 2013;197:300–13.
- 65 Mamidi S, Rossi M, Moghaddam SM, Annam D, Lee R, Papa R, et al. **Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L.** *Heredity (Edinb)* 2013;110:267–76.
- 66 Gepts P. *Phaseolus vulgaris* (Beans). Published Online First: 2001.
- 67 M. Santalla, A.B. Monteagudo, A.M. González, A.M. De Ron. **Agronomical and quality traits of runner bean germplasm and implications for breeding**. *Euphytica* 2004;135:205–15.
- 68 Sicard D, Nanni L, Porfiri O, Bulfon D, Papa R. **Genetic diversity of *Phaseolus vulgaris* L. and *P. coccineus* L. landraces in central Italy**. *Plant Breed* 2005;124:464–72.
- 69 Hernández-delgado S, Muruaga-martínez JS, Vargas-vázquez MLP, Martínez-mondragón J, Chávez-servia JL, Gill-langarica HR. **Advances in genetic diversity analysis of *Phaseolus* in Mexico**. *INTECH* 2015.
- 70 Rodiño AP, Lema M, Pérez-Barbeito M, Santalla M, De Ron AM. **Assessment of runner bean (*Phaseolus coccineus* L.) germplasm for tolerance to low temperature during early seedling growth**. *Euphytica* 2007;155:63–70.
- 71 Freytag GF, Debouck DG. *Phaseolus costaricensis*, a new wild bean species (Phaseolinae, Leguminosae) from Costa Rica and Panama, Central America. *Novon* 1996;6:157–63.
- 72 Gaur P. M., Mallikarjuna N., Knights T., Beebe S., Debouck D., Mejia A. et al.

- Gene introgression in grain legumes.** In: International conference on grain legumes: Quality improvement, value addition and trade (Kanpur) 2009.
- 73 Delgado-Salinas A, Bibler R, Lavin M. **Phylogeny of the genus *Phaseolus* (Leguminosae): a recent diversification in an ancient landscape.** *Syst Bot* 2006;**31**:779–91.
 - 74 Schmit V, Debouck D. **Observations on the origin of *Phaseolus polyanthus* Greenman.** *Econ Bot* 1991;**45**:345–64.
 - 75 Schmit V. **Etude de *Phaseolus polyanthus* Greenman et autres taxons du complexe *Phaseolus coccineus* L.** Thesis 1992.
 - 76 Worthington M, Soleri D, Aragón-Cuevas F, Gepts P. **Genetic composition and spatial distribution of farmer-managed *Phaseolus* bean plantings: An Example from a village in Oaxaca, Mexico.** *Crop Sci* 2012;**52**:172–1735
 - 77 Singh M, Upadhyaya HD, Bisht IS. **Genetic and genomic resources of grain legume improvement.** London; Waltham, MA: Elsevier 2013.
 - 78 Mercado-Ruaro P, Delgado-Salinas A. **Karyotypic studies on species of *Phaseolus* (Fabaceae: Phaseolinae).** *Am J Bot* 1998;**85**:1.
 - 79 Hernandez-Xolocotzi E. M-CS and PC. **El origen de *Phaseolus coccineus* L. darwinianus Hdz.-X. & Miranda-C. Sube nova.** *Rev la Soc Mex Hist Nat* 1959;**20**:99–121.
 - 80 Smartt J. **The possible status of *Phaseolus coccineus* L. SSP. darwinianus Hdz. X. et Miranda C. As a distinct species and cultigen of the genus *Phaseolus*.** *Euphytica* 1973;**22**:424–6.
 - 81 Piñero D, Eguiarte L. **The origin and biosystematic status of *Phaseolus coccineus* ssp. *polyanthus*: electrophoretic evidence.** *Euphytica* 1988;**37**:199–203.
 - 82 Llacá V, Delgado Salinas A, Gepts P. **Chloroplast DNA as an evolutionary marker in the *Phaseolus vulgaris* complex.** *Theor Appl Genet* 1994;**88**:646–52.
 - 83 Angioi SA, Desiderio F, Rau D, Bitocchi E, Attene G, Papa R. **Development and use of chloroplast microsatellites in *Phaseolus* spp. and other legumes.** *Plant Biol* 2009;**11**:598–612.
 - 84 Desiderio F, Bitocchi E, Bellucci E, Rau D, Rodriguez M, Attene G, et al. **Chloroplast microsatellite diversity in *Phaseolus vulgaris*.** *Front Plant Sci* 2013;**3**:312
 - 85 Delgado-Salinas A, Turley T, Richman A, Lavin M. **Phylogenetic analysis of the cultivated and wild species of *Phaseolus* (Fabaceae).** *Syst Bot* 1999;**24**:438–60.
 - 86 Debouck D. **Diversity in *Phaseolus* species in relation to the common bean.** In: Singh S, ed. common bean improvement in the twenty-first century. Springer Netherlands 1999:25–52.

- 87 Wendel J, Doyle J. **Phylogenetic incongruence: window into genome history and molecular evolution.** In: Soltis D, Soltis P, Doyle J, eds. *Molecular Systematics of Plants II*. Springer US 1998. 265–96.
- 88 Linder CR, Rieseberg LH. **Reconstructing patterns of reticulate evolution in plants.** *Am J Bot* 2004;**91**:1700–8.
- 89 Grant PR, Grant BR, Petren K. **Hybridisation in the recent past.** *Am Nat* 2005;**166**:56–67.
- 90 Sang T, Zhong Y. **Testing hybridisation hypotheses based on incongruent gene trees.** *Syst Biol* 2000;**49**:422–34.
- 91 McKinnon GE. **Reticulate evolution in higher plants, plant diversity and evolution.** In: Henry. RJ, ed. *Genotypic and phenotypic variation in higher plants*. Oxfordshire: CABI 2005; 81–96.
- 92 Funk VA. **Phylogenetic patterns and hybridisation.** *Ann Missouri Bot Gard* 1985;**72**:681–715.
- 93 Rieseberg LH, Baird SJ, Gardner KA. **Hybridisation, introgression, and linkage evolution.** *Plant Mol Biol* 2000;**42**:205–24.
- 94 Normark BB, Lanteri AA. **Incongruence between morphological and Mitochondrial-DNA characters suggests hybrid origins of parthenogenetic weevil lineages (Genus *Aramigus*).** *Syst Biol* 1998;**47**:475–94.
- 95 Kelly LJ, Leitch AR, Clarkson JJ, Hunter RB, Knapp S, Chase MW. **Intragenic recombination events and evidence for hybrid speciation in *Nicotiana* (Solanaceae).** *Mol Biol Evol* 2010;**27**:781–99.
- 96 Schumer M, Rosenthal GG, Andolfatto P. **How common is homoploid hybrid speciation?.** *Evolution* 2014;**68**:1553–60.
- 97 Jansen R, Ruhlman T. **Plastid genomes of seed plants.** In: Bock R, Knoop V, eds. *Genomics of chloroplasts and mitochondria*. Springer Netherlands 2012. 103–26.
- 98 Hegarty MJ, Hiscock SJ. **Hybrid speciation in plants: new insights from molecular studies.** *New Phytol* 2005;**165**:411–23.
- 99 Corriveau JL, Coleman AW. **Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species.** *Am J Bot* 1988;**75**:1443–58.
- 100 Schmit V, Du Jardin P, Baudoin J-P, Debouck DG. **Diversity studies of some *Phaseolus* taxa using chloroplast DNA as a molecular marker.** *Belgian Journal of Botany* 1992;**125**: 278-279.
- 101 Schlötterer C, Tautz D. **Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution.** *Curr Biol* 1994;**4**:777–83.
- 102 Sang T, Crawford DJ, Stuessy TF. **Documentation of reticulate evolution in**

- peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution.** *Proc Natl Acad Sci* 1995;**92**:6813–7.
- 103 Martinsen GD, Whitham TG, Turek RJ, Keim P. **Hybrid populations selectively filter gene introgression between species.** *Evolution (N Y)* 2009;**55**:1325–35.
 - 104 Städler T, Roselius K, Stephan W. **Genealogical footprints of speciation processes in wild tomatoes: Demography and evidence for historical gene flow.** *Evolution (N Y)* 2005;**59**:1268–79.
 - 105 Machado CA, Kliman RM, Markert JA, Hey J. **Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives.** *Mol Biol Evol* 2002;**19**:472–88.
 - 106 McBreen K, Lockhart PJ. **Reconstructing reticulate evolutionary histories of plants.** *Trends Plant Sci* 2006;**11**:398–404.
 - 107 Guggisberg A, Mansion G, Conti E. **Disentangling Reticulate Evolution in an Arctic-Alpine Polyploid Complex.** *Syst Biol* 2009;**58**:55–73.
 - 108 Holland BR, Huber KT, Moulton V, Lockhart PJ, et al. **Using consensus networks to visualize contradictory evidence for species phylogeny.** *Mol Biol Evol* 2004;**21**:1459–61.
 - 109 Schwarzer J, Misof B, Schliewen UK. **Speciation within genomic networks: a case study based on *Steatocranus* cichlids of the lower Congo rapids.** *J Evol Biol* 2012;**25**:138–48.
 - 110 Kole C. **Wild crop relatives: genomic and breeding resources: legume crops and forages.** Springer Berlin Heidelberg 2010.
 - 111 Jaccoud D, Peng K, Feinstein D, Kilian A. **Diversity arrays: a solid state technology for sequence information independent genotyping.** *Nucleic Acids Res* 2001;**29**:E25.
 - 112 Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinbolls A, et al. **Diversity Arrays Technology (DArT) for whole-genome profiling of barley.** *Proc Natl Acad Sci U S A* 2004;**101**:9915–20.
 - 113 Papa R, Nanni L, Sicard D, Rau D, Attene G. **The evolution of genetic diversity in *Phaseolus vulgaris* L.** In: T. J. Motley, N. Zerega, and H. Cross (Eds) *Darwin's harvest: new approaches to the origins, evolution, and conservation of crops.* NY: Columbia University Press 2006. 121–42.
 - 114 Hartl DL, Clark AG. **Principles of population genetics.** Massachusetts: Sinauer associates 2007.
 - 115 Leclerc M, Barriel V. **Low divergence in rDNA ITS sequences among five species of *Fucus* (Phaeophyceae) suggests a very recent radiation.** *J Mol Evol* 1998;**46**:115–120.
 - 116 Franzke A, Mummenhoff K. **Recent hybrid speciation in *Cardamine***

- (Brassicaceae) conversion of nuclear ribosomal ITS sequences in *statu nascendi*. *Theor Appl Genet* 1999;**98**:831–834
- 117 Afanador L, Haley S, Kelly JD. **Adoption of a ‘mini-prep’ DNA extraction method for RAPD marker analysis in common bean (*Phaseolus vulgaris* L.).** *Annu Rep Bean Improv Coop* 1993;**6**:10–11
 - 118 Edgar RC. **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004;**32**:1792–97.
 - 119 Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. **Jalview Version 2 a multiple sequence alignment editor and analysis workbench.** *Bioinformatics* 2009;**25**:1189–91.
 - 120 Nei M, Li W H. **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci USA* 1979;**76**:5269–73.
 - 121 Librado P, Rozas J. **DnaSP v5: A software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009;**25**:1451-1.
 - 122 Nei M, Kumar S. **Molecular Evolution and Phylogenetics.** New York: Oxford University Press 2000.
 - 123 Excoffier L, Lischer H. **Arlequin suite version 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.** *Mol Ecol Resour* 2010;**10**:564–7.
 - 124 Wright S. **The genetical structure of populations.** *Ann Eugen* 1949;**15**:323–54.
 - 125 Excoffier L, Smouse PE, Quattro JM. **Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data.** *Genetics* 1992;**131**:479–91.
 - 126 Saitou N, Nei M. **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987;**4**:406–25.
 - 127 Serrano-Serrano ML, Hernández-Torres J, Castillo-Villamizar G, Debouck DG, Chacón Sánchez MI. **Gene pools in wild Lima bean (*Phaseolus lunatus* L.) from the Americas: Evidences for an Andean origin and past migrations.** *Mol Phylogenet Evol* 2010;**54**:76–87.
 - 128 Perez-Vega E, Paneda A, Rodriguez-Suarez C, Campa A, Giraldez R, Ferreira JJ. **Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.).** *Theor Appl Genet* 2010;**120**:1367–80.
 - 129 Bandelt HJ, Forster P, Röhl A. **Median-joining networks for inferring intraspecific phylogenies.** *Mol Biol Evol* 1999;**16**:37–48.
 - 130 Barrett SCH, Kohn JR, Falk DA, Holsinger KE. **Genetic and evolutionary consequences of small population size in plants: implications for conservation.** *Genet Conserv rare plants* 1991;3–30.

- 131 Schmit V, Munoz JE, Du Jardin P, Baudoin JP, Debouck DG. **Phylogenetic studies of some Phaseolus taxa on the basis of chloroplast DNA polymorphisms.** CIAT 1994.
- 132 Dulloo ME, Hanson J, Jorge MA, Thormann I. **Regeneration guidelines: general guiding principles.** In: Dulloo M.E., Thormann I., Jorge M.A. and Hanson J. ed. Crop specific regeneration guidelines. CGIAR System-wide Genetic Resource Programme (SGRP) 1998.
- 133 Soleri D, Worthington M, Aragón-Cuevas F, Smith S, Gepts P. **Farmers' varietal identification in a reference sample of local Phaseolus species in the Sierra Juárez, Oaxaca, Mexico.** *Econ Bot* 2013;**67**:283–98.
- 134 Tohme, J., P. Jones, S. Beebe, and M. Iwanaga. **The combined use of agroecological and characterization data to establish the CIAT Phaseolus vulgaris core collection.** In Core collections of plant genetic resources T. Hodgkin, A.H.D. Brown, T.J.L. van Hintum and E.A.V. Morales, eds. Chichester, UK: John Wiley and Sons 1995.
- 135 Rieseberg LH, Ellstrand NC, Arnold DM. **What can molecular and morphological markers tell us about plant hybridisation?.** *CRC Crit Rev Plant Sci* 1993;**12**:213–41.
- 136 Ritland CE, Ritland K, Straus NA. **Variation in the ribosomal internal transcribed spacers (ITS1 and ITS2) among eight taxa of the Mimulus guttatus species complex.** *Mol Biol Evol* 1993;**10**:1273–88.
- 137 Wendel JF, Schnabel A, Seelanan T. **Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium).** *Proc Natl Acad Sci U S A* 1995;**92**:280–4.
- 138 van Houten WH, Scarlett N, Bachmann K. **Nuclear DNA markers of the Australian tetraploid Microseris scapigera and its North American diploid relatives.** *Theor Appl Genet* 1993;**87**:498–505.
- 139 Feliner G, Rosselló J. **Concerted evolution of multigene families and homoeologous recombination.** *Plant Genome Diversity* 2012;**1**:171–193
- 140 Kovarik A, Pires JC, Leitch AR, Lim KY, Sherwood AM, Matyasek R, et al. **Rapid concerted evolution of nuclear ribosomal DNA in two Tragopogon allopolyploids of recent and recurrent origin.** *Genetics* 2005;**169**:931–44.
- 141 Pelsner PB, Abbott RJ, Comes HP, Milton JJ, Möller M, Looseley Me, et al. **The genetic ghost of an invasion past: colonization and extinction revealed by historical hybridisation in Senecio.** *Mol Ecol* 2012;**21**:369–87.
- 142 Grant V. **The origin of a new species of Gilia in a hybridisation experiment.** *Genetics* 1966;**54**:1189–99.
- 143 Frajman B, Oxelman B. **Reticulate phylogenetics and phytogeographical structure of Heliosperma (Sileneae, Caryophyllaceae) inferred from chloroplast and nuclear DNA sequences.** *Mol Phylogenet Evol* 2007;**43**:140–55.

- 144 Pan J, Zhang D, Sang T. **Molecular phylogenetic evidence for the origin of a diploid hybrid of *Paeonia* (Paeoniaceae).** *Am J Bot* 2007;**94** :400–8.
- 145 Morrison. D. **Introduction to Phylogenetic Networks.** Uppsala, Sweden: RJR Productions 2011.
- 146 Vriesendorp B, Bakker FT. **Reconstructing patterns of reticulate evolution in angiosperms: what can we do?.** *Am. J. Bot* 2004;**91**:1700-1708
- 147 Castresana J. **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000;**17**:540–52.
- 148 Maddison WP and DRM. **Mesquite: a modular system for evolutionary analysis.** Version 2.75 <http://mesquiteproject.org>.
- 149 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. **MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011;**28**:2731–9.
- 150 Kelchner SA, Thomas MA. **Model use in phylogenetic: nine key questions.** *Trends Ecol Evol* 2007;**22**:87–94.
- 151 Posada D. jModelTest: **Phylogenetic model averaging.** *Mol Biol Evol* 2008;**25**:1253–6.
- 152 Darriba D, Taboada GL, Doallo R, Posada D. **jModelTest 2: more models, new heuristics and parallel computing.** *Nat Meth* 2012;**9**:772.
- 153 Akaike H. **Information theory and an extension of the maximum likelihood principle.** In: Springer Series in Statistics. New York: Springer 1998.
- 154 Kimura M. **Estimation of evolutionary distances between homologous nucleotide sequences.** *Proc Natl Acad Sci* 1981;**78**:454–8.
- 155 Felsenstein J. **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981;**17**:368–76.
- 156 Tavaré S. **Some probabilistic and statistical problems in the analysis of DNA sequences.** *Lect Math Life Sci* 1986;**17**:57–86.
- 157 Higgs PG, Attwood TK. **Bioinformatics and molecular evolution.** Wiley-Blackwell 2009.
- 158 Day WHE. **Properties of the nearest neighbor interchange metric for trees of small size.** *J Theor Biol* 1983;**101**:275–88.
- 159 Felsenstein J. **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution (N Y)* 1985;**783**–91.
- 160 Ronquist F, Huelsenbeck JP. MrBayes 3: **Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003;**19**:1572–4.
- 161 Huelsenbeck J, Rannala B. **Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution**

- models.** *Syst Biol* 2004;**53**:904–13.
- 162 Ronquist F, J. P. Huelsenbeck P v. d. M. **Bayesian phylogenetic analysis using MrBayes.** In: P. Lemey and Vandamme A. SM, ed. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge: Cambridge University Press 2009.
 - 163 Erixon P, Svennblad B, Britton T, Oxelman B. **Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics.** *Syst Biol* 2003;**52**:665–73.
 - 164 Brandley MC, Schmitz A, Reeder TW. **Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards.** *Syst Biol* 2005;**54**:373–90.
 - 165 Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. **Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference.** *Bioinformatics* 2004;**20**:407–15.
 - 166 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. **MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space.** *Syst Biol* 2012;**61**:539–42
 - 167 Huson DH, Scornavacca C. Dendroscope 3: **An interactive tool for rooted phylogenetic trees and networks.** *Syst Biol* 2012;**61**:1061–7.
 - 168 Vekemans X, Hardy O, Berken B, Fofana B, Baudoin J-P. **Use of PCR-RFLP on chloroplast DNA to investigate phylogenetic relationships in the genus Phaseolus.** *Biotechnol Agron Soc Environ* 1998;**2**:128–34.
 - 169 Gottlieb LD. **Levels of confidence in the analysis of hybridisation in plants.** *Ann Missouri Bot Gard* 1972;**59**:435.
 - 170 Colin SM. **Infiltración genética entre Phaseolus coccineus L. y Phaseolus vulgaris L.** Chipamgo: Colegio de Postgraduados, Escuela Nacional de Agricultura 1967.
 - 171 Huttner E, Wenzl P, Akbari M, Caig V, Carling J, Cayla C, et al. **Diversity arrays technology: a novel tool for harnessing the genetic potential of orphan crops.** In: Discovery to delivery: BioVision Alexandria 2004, Proceedings of the 2004 Conference of the world biological forum. CABI 2005. 145–55.
 - 172 Appleby N, Edwards D, Batley J. **New technologies for ultra-high throughput genotyping in plants.** *Methods Mol Biol* 2009;**513**:19–39.
 - 173 Gupta PK, Rustgi S, Mir RR. **Array-based high-throughput DNA markers for crop improvement.** *Heredity (Edinb)* 2008;**101**:5–18.
 - 174 Bolibok-Bragoszewska H, Heller-Uszyńska K, Wenzl P, Uszyński G, Kilian A, Rakoczy-Trojanowska M. **DArT markers for the rye genome - genetic diversity and mapping.** *BMC Genomics* 2009;**10**:578.
 - 175 Xia L, Peng K, Yang S, Wenzl P, de Vicente MC, Fregene M, et al. **DArT for**

- high-throughput genotyping of Cassava (*Manihot esculenta*) and its wild relatives.** *Theor Appl Genet* 2005;**110**:1092–8.
- 176 Briñez B, Blair M, Kilian A, Carbonell S, Chiorato A, Rubiano L. **A whole genome DArT assay to assess germplasm collection diversity in common beans.** *Mol Breed* 2012;**30**:181–93.
- 177 Vu TTH. **Use of diversity array technology (DArT) to identify QTLs for physiological traits in mungbean (*Vigna radiata*) and soybean (*Glycine max*).** Townsville:James Cook University 2013.
- 178 Thudi M, Upadhyaya HD, Rathore A, Gaur PM, Krishnamurthy L, Roorkiwal M, et al. **Genetic dissection of drought and heat tolerance in chickpea through genome-wide and candidate gene-based association mapping approaches.** *PLoS One* 2014;**9**.
- 179 Campbell R, Pont SDA, Morris JA, McKenzie G, Sharma SK, Hedley PE, et al. **Genome-wide QTL and bulked transcriptomic analysis reveals new candidate genes for the control of tuber carotenoid content in potato (*Solanum tuberosum* L.).** *Theor Appl Genet* 2014;**127**:1917–33.
- 180 Cui F, Fan X, Zhao C, Zhang W, Chen M, Ji J, et al. **A novel genetic map of wheat: utility for mapping QTL for yield under different nitrogen treatments.** *BMC Genet* 2014;**15**:57.
- 181 Briñez R. B. **‘Desenvolvimento da plataforma dart e mapeamento de locos associados com tolerância à seca em feijão (*Phaseolus vulgaris* L.)’** Thesis. Campinas: The Agronomic Institute (IAC) 2013.
- 182 Dytham C. **Choosing and using statistics: a biologist’s guide.** John Wiley & Sons 2011.
- 183 Peakall R, Smouse PE. **Genalex 6: genetic analysis in excel. population genetic software for teaching and research.** *Mol Ecol Notes* 2006;**6**:288–95.
- 184 Perrier X, Flori A, Bonnot F. **Data analysis methods.** *Genet Divers Cultiv Trop plants* 2003:43–76.
- 185 Perrier, X., Flori, A, Bonnot, F. **Data analysis methods.** In: Hamon, P., Seguin, M., Perrier, X., Glaszmann, J. C. Ed., Genetic diversity of cultivated tropical plants. Enfield, Montpellier: Science Publishers 2003.
- 186 Bruno WJ, Socci ND, Halpern AL. **Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction.** *Mol Biol Evol* 2000;**17**:189–97.
- 187 Huson DH, Bryant D. **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006;**23**:254–67.
- 188 Pritchard JK, Stephens M, Donnelly P. **Inference of population structure using multilocus genotype data.** *Genetics* 2000;**155**:945–59.
- 189 Falush D, Stephens M, Pritchard JK. **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.**

- Genetics* 2003;**164**:1567–87.
- 190 Evanno G, Regnaut S, Goudet J. **Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.** *Mol Ecol* 2005;**14**:2611–20.
 - 191 Jakobsson M, Rosenberg NA. **CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure.** *Bioinformatics* 2007;**23**:1801–6.
 - 192 Earl D, von Holdt B. **STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.** *Conserv Genet Resour* 2012;**4**:359–61.
 - 193 Howard EL, Whittock SP, Jakše J, Carling J, Matthews PD, Probasco G, et al. **High-throughput genotyping of hop (*Humulus lupulus* L.) utilising diversity arrays technology (DArT).** *Theor Appl Genet* 2011;**122**:1265–80.
 - 194 Kwak M, Kami JA, Gepts P. **The putative mesoamerican domestication center of *Phaseolus vulgaris* is located in the Lerma–Santiago basin of Mexico.** *Crop Sci* 2009;**49**:554.
 - 195 Brown CH, Clement CR, Epps P, Luedeling E, Wichmann S. **The paleobiolinguistics of the common bean (*Phaseolus vulgaris* L.).** *Ethnobiol Lett* 2014;**5**:104.
 - 196 Yang S, Pang W, Ash G, Harper J, Carling J, Wenzl P, et al. **Low level of genetic diversity in cultivated Pigeonpea compared to its wild relatives is revealed by diversity arrays technology.** *Theor Appl Genet* 2006;**113**:585–95.
 - 197 Roorkiwal M, von Wettberg EJ, Upadhyaya HD, Warschefsky E, Rathore A, Varshney RK. **Exploring germplasm diversity to understand the domestication process in *Cicer* spp. using SNP and DArT markers.** *PLoS One* 2014;**9**:1–10
 - 198 Araya-Villalobos, R. Martínez Umaña, K.; Debouck, Daniel G. **How many bean species are in Costa Rica today?.** Universidad de Costa Rica; Centro Internacional de Agricultura Tropical (CIAT). Annual Meeting Guatemala 2015.
 - 199 Hagmann E, Post L von, Post R von, Eklund M, Larsson C-T, Tuvešson S, et al. **QTL mapping of powdery mildew resistance in oats using DArT markers.** 5th EWAC-Eucarpia Cereals Section Conference 2012.
 - 200 Amorim EP, Vilarinhos AD, Cohen KO, Amorim VBO, dos Santos-Serejo JA, Silva SOE, et al. **Genetic diversity of carotenoid-rich bananas evaluated by Diversity Arrays Technology (DArT).** *Genet Mol Biol* 2009;**32**:96–103.
 - 201 Grzebelus D, Iorizzo M, Senalik D, Ellison S, Cavagnaro P, Macko-Podgorni A, et al. **Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers.** *Mol Breed* 2014;**33**:625–37.

Appendices

Appendix I

A1. Seeds identification and plant growing conditions



A2 Haplotypes identified for each accession of bean.

Species	Accession	ITS	<i>trnT-trnL</i> Spacer	<i>trnL</i> intron	<i>rps14-psaB</i> Spacer	<i>accD-psaI</i> Spacer	Concatenate cpDNA*
<i>P. acutifolius</i>	G40286	--	--	ti1	rs4	--	--
<i>P. acutifolius</i>	G40178	i2	--	ti1	rs3	as1	Hap 1
<i>P. albescens</i>	PL3592	i3	ts7	ti2	rs5	as2	Hap 2
<i>P. coccineus</i>	G35837A	--	--	ti5	rs7	as3	--
<i>P. coccineus</i>	G36006	i15	ts4	ti8	rs7	as14	Hap 10
<i>P. coccineus</i>	G36334	i11	ts6	ti8	rs7	as14	Hap 10
<i>P. coccineus</i>	G36353	i16	ts4	ti8	rs7	as10	Hap 11
<i>P. coccineus</i>	G36285	i16	ts4	ti8	rs7	as11	Hap 12
<i>P. coccineus</i>	G36206	i15	ts14	ti3	rs7	as3	Hap 4
<i>P. coccineus</i>	G36256	i12	ts3	ti5	rs7	as12	Hap 6
<i>P. coccineus</i>	G35650	i13	ts2	ti7	rs6	as9	Hap 7
<i>P. coccineus</i>	G36002	i13	ts4	ti8	rs7	as9	Hap 8
<i>P. coccineus</i>	G35757	i13	ts4	ti8	rs7	as13	Hap 9
<i>P. coccineus</i>	G35859	i14	ts5	ti8	rs7	as13	Hap 9
<i>P. dumosus</i>	G35758	i9	ts8	ti6	rs7	as3	Hap 18
<i>P. dumosus</i>	G35946	i9	--	ti4	rs7	as3	Hap 5
<i>P. dumosus</i>	G36127	i5	ts8	ti4	rs7	as3	Hap 5
<i>P. dumosus</i>	G35641	i9	ts8	ti4	rs7	as3	Hap 5
<i>P. dumosus</i>	G35908	i9	ts8	ti4	rs7	as3	Hap 5
<i>P. dumosus</i>	G36045	i9	ts8	ti4	rs7	as3	Hap 5
<i>P. dumosus</i>	G36286	i9	ts8	ti4	rs7	as3	Hap 5
<i>P. dumosus</i>	G35877	i10	ts8	ti4	rs7	as3	Hap 5
<i>P. lunatus</i>	G25224	--	ts1	ti13	rs2	as16	Hap 20
<i>P. lunatus</i>	G25226	--	ts1	ti13	rs2	as16	Hap 20
<i>P. vulgaris</i>	G7479	i5	ts9	ti11	rs7	--	--
<i>P. vulgaris</i>	G23463	i7	ts12	ti10	rs7	as5	Hap 14
<i>P. vulgaris</i>	G12853	i7	ts10	ti11	rs10	as7	Hap 15
<i>P. vulgaris</i>	G50387	i5	--	ti11	rs7	as7	Hap 16
<i>P. vulgaris</i>	G50388B	i5	--	ti11	rs7	as7	Hap 16
<i>P. vulgaris</i>	G19908	i7	ts11	ti11	rs7	as7	Hap 16
<i>P. vulgaris</i>	G19909	--	ts11	ti11	rs7	as7	Hap 16
<i>P. vulgaris</i>	G21244	i4	ts11	ti11	rs7	as7	Hap 16
<i>P. vulgaris</i>	G24358	i6	--	ti11	rs9	as6	Hap 17
<i>P. costaricensis</i>	G40604	i8	ts13	ti9	rs8	as3	Hap 13
<i>P. costaricensis</i>	G40805	i3	--	ti2	rs7	as2	Hap 3
<i>P. leptostachyus</i>	G40553	i1	--	ti12	rs1	as15	Hap 19
<i>P. persistentus</i>	DQ4457	i7	--	ti11	rs7	as7	Hap 16

[illegible]

[illegible]

Consensus sequence A C G A G A A G G C G C C T G T T G C T T T C A T A G T T A A G C G T A T C A G C G C G G G C A C G A G C T

Haplotypes

Hap i1	T.	A.	T.	.	C	A	A.	A.	.	C.	.	.	G	C	G.	.	.	A	A.	.	C	A.	A	G	C	C	T.	C.	.	.	.	.	.	.	.	.	T.
Hap i2	C	G.	.	.	.	.	C.	A.	.	T.	.	.	.	A.	.	.	.	.	.	.	.	.	.	A	G.	.	.	.	T.	.	A.	.	G.	A.	A.	C.	
Hap i3	.	.	G.	G	G.	C.	T.	.	A.	.	C	C.	.	.	.	G	C.	.	C.	.	T.	.	.	.	.	.	.	.	.	.	.	T.	.	.	C.		
Hap i4	.	.	.	.	.	.	.	.	T.	.	.	.	.	.	.	G.	.	.	.	.	.	.	A.	.	.	.	.	.	.	.	T.	.	.	.	.	.	
Hap i5	.	.	.	.	.	.	.	.	T.	.	.	.	.	.	.	C.	.	.	.	.	.	.	A.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Hap i6	.	.	.	.	.	.	.	.	T.	.	.	.	T.	.	.	.	.	.	.	.	.	.	A.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Hap i7	.	.	.	.	.	.	.	.	T.	.	.	.	.	.	.	.	.	.	.	.	.	.	A.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Hap i8	.	.	.	G.	.	.	.	.	C.	.	C.	.	.	.	.	C.	.	.	.	.	.	.	.	.	.	.	.	T.	.	A.	.	.	.	.	C.		
Hap i9	.	.	.	G.	.	.	.	.	C.	.	C.	.	.	.	.	C.	.	.	.	.	.	.	.	.	.	.	.	.	.	A.	.	.	.	.	C.		
Hap i10	.	.	.	G.	.	.	.	.	C.	.	C.	.	.	.	.	C.	.	.	.	.	.	.	.	.	.	.	.	.	.	A.	.	.	.	.	C.		
Hap i11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T.	.	.	.	.	.	.	.	.	T.	.	.	A.	.	.	.	.	.	.		
Hap i12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G.	.	.	.	.	.	.	.	.	.	A.	A.	.	.	.	.	.	.	.	
Hap i13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G.	.	.	.	.	.	.	.	.	.	.	A.	.	.	.	.	.	.	.	
Hap i14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G.	.	C.		
Hap i15	.	.	.	.	.	.	.	.	T.	.	.	.	.	.	.	C.	.	.	.	.	.	.	.	G.	.	.	.	.	.	.	.	G.	.	.	.	.	
Hap i16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C.	.	.	.	.	.	.	.	G.	.	.	.	.	.	.	.	G.	.	.	.	.	

Haplotypes generate from a concatenated matrix of the chloroplast sequences: trnT-trnL spacer, rps14-psaB spacer and accD-psaI spacer.

[illegible]

Consensus sequence C - - GAGT- - G- G- - - TGG- - AATAA- T- - - GC- - TT- AG- CA- A- GCC-

Haplotype

Hap1 TGGTCTCAAATAACCCTTCGCCGCGGGGCTTAGTGCTTTGAGTCATTGA

Hap 2 A A

Hap 3. - - - - , - - - - , - - - - , - - - - , - - - - , - - - - , - - - - , - - - - , - - - - .

Hap 4 G A

Hap 5 T

Hap 6 . - T . - - - - - G - - - - -

Hap 7

Hap 8

Hap 9 - - - - -

Hap 10

Hap 11:

Jan 12 - - - - -

Hap 13 - - - - C - - - - -

Hap 14 - - - - - C - - - - -

Hap 15:A.....

Hap 16

Hap 17

Нап 18

Нар 19 - - - - - A - - - - - G - G - C - C - C - - - -

Нар 20 - T - - - - - T - - - - A - - - - - C - - - - - C

[illegible]

[illegible]

Consensus sequence - G G A T - - A G A - - C T T - C G A C T G - - - A G T G C G T - C C T G - - - - A - A - - - G A

Haplotype

Hap 1 GTAGCTCGAGTAGGGATAGAGACAGCACCTACTAAGTCAGATGTCACTC

Hap 2 - - - - - . - . - - . . - - . - . . . - - - - . - - - - - - -

Hap 3 - - - - - . - . - - . . - - . - . . . - - - - . - - - - - - -

Hap 4 - - - - - . - . - - - . - - - . - . . - - - - - . . . - - - - - . . - - - - - . . .

Hap 5 - - - - . - . - - . - - . - . . - - - - . . .

Hap 6 - - - . - - - . - . - - - . . - - . - . . . - - - A . - - - - . . - - G - - - - - . .

[illegible][illegible][illegible]

Hap 10 - - - . - - - . - . - - . . - - . - . . . - - . - - - - . - - - . - - - . .

Hap 11 - - - . - - - . - . - - . . - - . - . . . - - . - - - - . . - - - - - . .

Hap 12 - - - . - - - . - . - - . . - - . - . . . - - - . - - . . . - - - . A. - - - . .

Hap 13 - - - - - . - . - - - . - - . - . . . - - - - - . . . - - . . - - - - - . .

Hap 14 - - - - - A - - - - -

Hap 15 - - - - - . - . - - - . - - . - . . . - - - - - . . . - - . . - - - - - . .

Hap 16 - - - - - , - , - - - , - - - , . . . - - - - - , . . . - - - - - , .

Hap 17 - - - - - , - , - - - , - - - , . . . - - - - - , . . . - - - - - A - - - - - .

[illegible]

Hap 19 - . . C . - A - . . C - - . . - . . - T G - - G - - - - - G - - - - - A . .

Hap 20 A . . C . C - - . . - G A . . C . . - . . - - - - - - - . - - - A - - - - - C - - .

A4 AMOVA derived from sequences of ITS and cpDNA.

AMOVA for sequences of ITS.

Source of variation	d. f.	Sum of squares	Variance components	Percentage of variation	Fixation index	P-value
Among populations of wild <i>P. dumosus</i> (Guatemala and Mexico)	1	0.017	-0.022	-32.35		
Within population wild <i>P. dumosus</i>	10	0.9	0.09	132.35		
Total	11	0.917	0.068		FST : -0.32353	1
Among populations of wild <i>P. vulgaris</i> (Guatemala, Mexico and Peru)	1	0.333	-0.011	-2.86		
Within population of wild <i>P. vulgaris</i>	10	4	0.4	102.86		
Total	11	4.333	0.389		FST : -0.02857	0.7351
Among populations of wild <i>P. coccineus</i> (Guatemala and Mexico)	1	12.127	1.57	43.52		
Within populations of wild <i>P. coccineus</i>	12	24.444	2.037	56.48		
Total	13	36.571	3.607		FST : 0.43519	0.0127
Among groups (wild <i>P. dumosus</i> , wild <i>P. coccineus</i> and wild <i>P. vulgaris</i>)	2	418.287	15.284	90.65	FCT : 0.90653	0.00684
Among populations within groups	4	16.144	0.68655	4.07	FSC : 0.43569	0.00293
Within populations	33	29.344	0.88923	5.27	FST : 0.94726	0
Total	39	463.775	16.85931			

AMOVA for sequences of *trnL* intron of wild bean.

Source of variation	d. f.	Sum of squares	Variance components	Percentage of variation	Fixation index	P-value
Among groups (wild <i>P. dumosus</i> , wild <i>P. coccineus</i> and wild <i>P. vulgaris</i>)	2	39.545	1.30353	78.38	FCT: 0.78385	0.00587
Among populations within groups	4	1.318	-0.00729	-0.44	FSC: -0.02027	0.39101
Within populations	38	13.936	0.36675	22.05	FST: 0.77947	0
Total	44	54.8	1.66299			

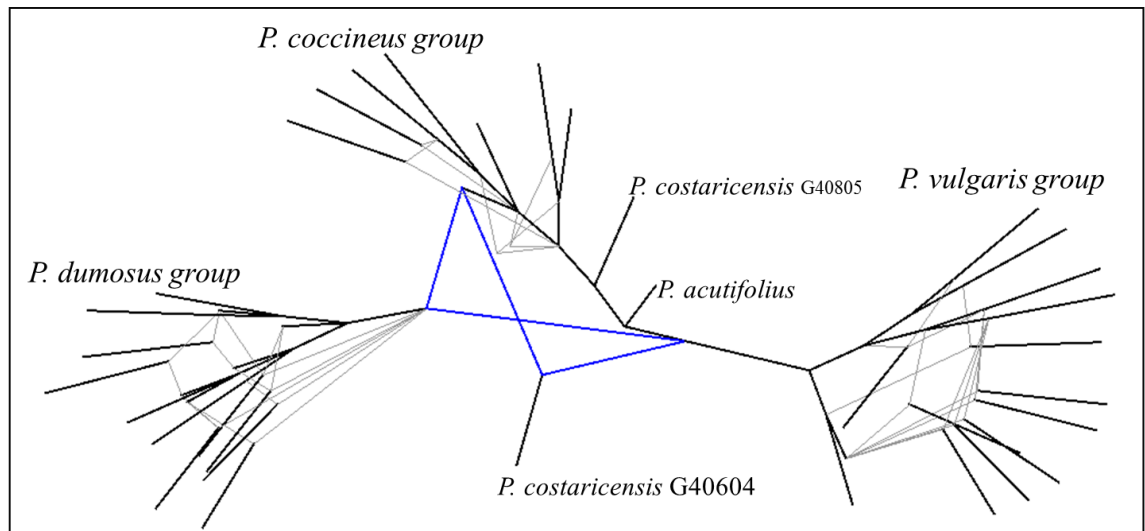
**AMOVA made using a concatenated matrix of the chloroplast sequences:
trnT-trnL spacer, *rps14-psaB* spacer and *accD-psaI* spacer.**

Source of variation	d. f.	Sum of squares	Variance components	Percentage of variation	Fixation index	P-value
Among groups (wild <i>P. dumosus</i> , wild <i>P. coccineus</i> and wild <i>P. vulgaris</i>)	2	73.139	2.87186	76.4	FCT: 0.76396	0.01075
Among populations within groups	4	4.434	0.0659	1.75	FSC: 0.07426	0.11535
Within populations	30	24.643	0.82143	21.85	FST: 0.78149	0
Total	36	102.216	3.75919			

In all the AMOVA analysis the group level conformed by accessions of the same species. The population level corresponds to accessions from the same country. **d.f.** indicates degrees of freedom. Significance level =0.05

Appendix II

B1. Galled consensus network



The topology summarizes incongruities between the ITS and the cpDNA. Blue lines represent reticulations.

Appendix III

C1 Accessions lists

List of accessions representatives of the species on the common bean secondary gene pool, biological status and Germplasm bank source.

(*)= Core collection accessions

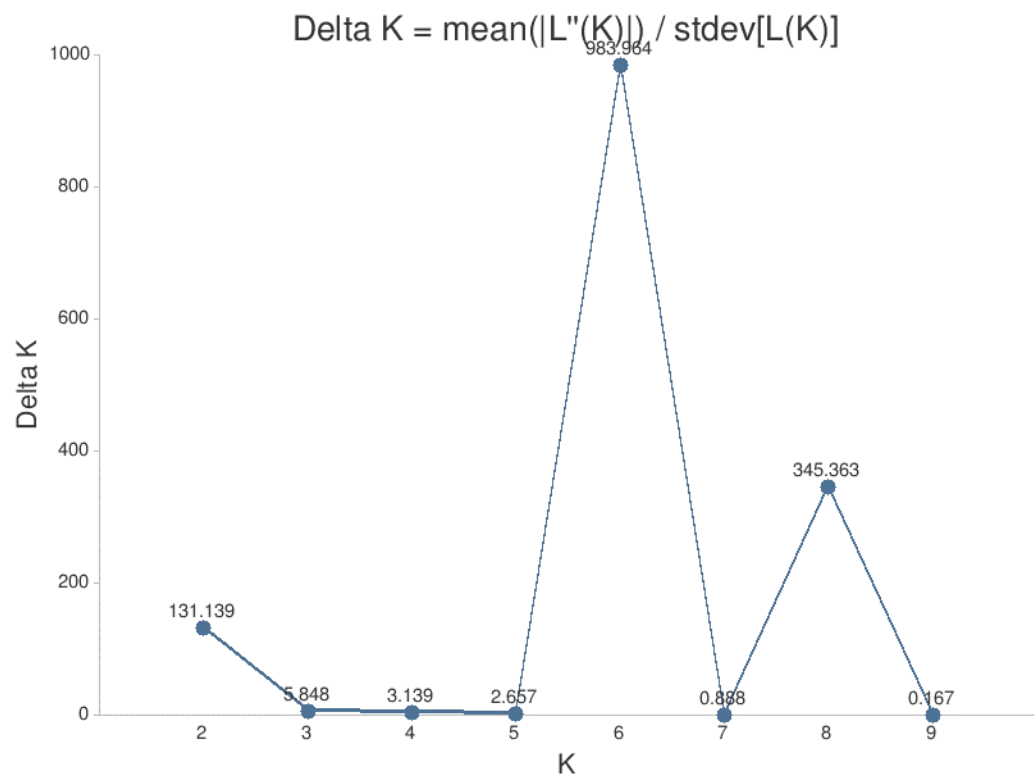
Specie	Accession	Biological Status	Germplasm Bank
<i>Phaseolus acutifolius</i> A.Gray	G40286	Wild	CIAT
<i>Phaseolus acutifolius</i> A.Gray	G40178	Wild	CIAT
<i>Phaseolus albescens</i> McVaugh ex R.Delgado. & A.Delgado	PL3592	Wild	CIAT
<i>Phaseolus coccineus</i> L.	PI449381	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	PI325603	Landrace	USDA
<i>Phaseolus coccineus</i> L.	G36353	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G35757*	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G35837A*	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G36002	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G36006	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G36206	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G36285	Wild	CIAT
<i>Phaseolus coccineus</i> L.	PI433237	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	PI451863	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	PI451868	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	PI451870	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	G36256	Cultivated	CIAT
<i>Phaseolus coccineus</i> L.	PI451872	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	PI317572	Landrace	USDA
<i>Phaseolus coccineus</i> L.	PI205360	Landrace	USDA
<i>Phaseolus coccineus</i> L.	G35650	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G35859	Wild	CIAT
<i>Phaseolus coccineus</i> L.	G36334	Wild	CIAT
<i>Phaseolus coccineus</i> L.	PI313503	Landrace	USDA
<i>Phaseolus coccineus</i> L.	PI346951	Cultivated	USDA
<i>Phaseolus coccineus</i> L.	PI201336	Landrace	USDA
<i>Phaseolus costaricensis</i> Freytag & Debouck	G40805	Wild	CIAT
<i>Phaseolus costaricensis</i> Freytag & Debouck	G40604	Wild	CIAT
<i>Phaseolus dumosus</i> Macfad. L.	PI355423	Cultivated	USDA
<i>Phaseolus dumosus</i> Macfad. L.	G35758*	Wild	CIAT
<i>Phaseolus dumosus</i> Macfad. L.	G35877*	Wild	CIAT
<i>Phaseolus dumosus</i> Macfad. L.	G35908	Wild	CIAT
<i>Phaseolus dumosus</i> Macfad. L.	G35946	Wild	CIAT

<i>Phaseolus dumosus Macfad. L.</i>	G36045	Wild	CIAT
<i>Phaseolus dumosus Macfad. L.</i>	G36286	Wild	CIAT
<i>Phaseolus dumosus Macfad. L.</i>	PI190074	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI311194	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI311165	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI317563	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI194585	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI195389	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI311183	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	PI195399	Landrace	USDA
<i>Phaseolus dumosus Macfad. L.</i>	G35641	Cultivated	CIAT
<i>Phaseolus dumosus Macfad. L.</i>	G36127	Cultivated	CIAT
<i>Phaseolus leptostachyus</i>	G40553	Wild	CIAT
<i>Phaseolus lunatus</i>	G25224	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G12853*	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G19908*	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G19909*	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G50387	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G50388B	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G23463	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G24358	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G7479	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	G21244*	Wild	CIAT
<i>Phaseolus vulgaris L.</i>	PI109863	Cultivated	USDA
<i>Phaseolus vulgaris L.</i>	PI109860	Cultivated	USDA
<i>Phaseolus vulgaris L.</i>	PI109861	Cultivated	USDA
<i>Phaseolus vulgaris x coccineus</i>	G50998C	Cultivated	CIAT
<i>Phaseolus vulgaris x Phaseolus dumosus</i>	G35980E	Cultivated	CIAT
<i>Phaseolus vulgaris x Phaseolus dumosus</i>	G35980F	Cultivated	CIAT
<i>Phaseolus tuerckheimii</i>	G40578	Wild	CIAT

C2 AMOVA derived from DArT array

Source of variation	d. f.	Sum of squares	Variance components	Percentage of variation	Fixation index	P-value
Among groups (wild <i>P. dumosus</i> , wild <i>P. coccineus</i> and wild <i>P. vulgaris</i>)	3	7682.8830	103.0170	66.3500	FCT: 0.6635	0.0029
Among populations within groups	5	1151.2530	18.5060	11.9200	FSC: 0.3542	0.0000
Within populations	91	3069.6640	33.7320	21.7300	FST: 0.7827	0.0000
Total	99	11903.8000	155.2560			

C3 Estimation of K by the median values of loge of likelihood value L (k)



C4 Estimation of K by the Δk value.