Polytechnic University of Marche
Department of Agricultural, Food and Environmental Sciences

Scientific field: AGR/07 - Plant Genetics

PhD School of Agricultural, Food and Environmental Sciences

XXXIV cycle (2018-2021)

Investigation of the evolutionary history of common bean (*Phaseolus vulgaris*)

Ph.D. Supervisor Prof.
Roberto Papa

Ph.D. School director Prof.
Bruno Mezzetti

Ph.D. candidate
Giulia Frascarelli

Table of Contents

Summary

The study of wild forms represents a fundamental resource for the development of new varieties. Indeed, evolutionary studies are at the base of the development of breeding programs because they give information about the available genetic diversity that can be utilized to recover the genetic variability lost during the process of domestication. The most important grain legumes for human consumption belong to the *Phaseolus* genus. In particular, common bean (*P. vulgaris*) has a peculiar evolutionary history which makes this species a model for the study of crop evolution. Three different eco-geographical gene pools can be recognized within the wild forms of common bean: the Mesoamerican gene pool, the Andean gene pool, and the North Peru -Ecuador gene pool. Nevertheless, only the Mesoamerican and Andean gene pools undergone domestication. Numerous studies investigated the origin of common bean and different hypotheses have been proposed: (i) the North Peruvian-Ecuadorian origin which is based on the identification of an ancestral type of the Phaseolin protein present only in accessions from North Peru-Ecuador; (ii) the monophyletic Mesoamerican origin mainly based on phylogenetic analyses and on the higher genetic diversity of the Mesoamerican gene pool compared to the other two populations; (iii) two distinct origins for the North Peru-Ecuador gene pool and the group composed by the Mesoamerican and Andean gene pools; (vi) the origin from a common ancestor that has not been found yet or has been extinct.

In this thesis work, we aimed at solving the open discussion about the origin of common bean through the investigation of its phylogeny at plastid and nuclear level. Specifically, we reconstructed the phylogeny of common bean with plastid SNPs and with chloroplast genomes that were *de-novo* assembled in the current study. In addition, plastomes were used to estimate the divergence times of the gene pools showing two migration events: from Mesoamerica to North Peru -Ecuador occurred ~ 150.000 years ago and one more recent from Mesoamerica to South Andes arisen ~ 90.000 years ago. Finally, nuclear data of a set of 10 accessions of *P. vulgaris* were used to study the evolutionary history at nuclear level. To respect the assumption of absence of recombination, SNPs from the centromeric region of each chromosome were selected and used for the analyses.

Albeit the analyses of plastid SNPs and whole plastomes clearly reflect a monophyletic and Mesoamerican origin of common bean and do not identify the North Peru-Ecuador gene pool as a different species, we found that this population has discordant behaviors when analyzing whole genome markers and SNPs located in non-recombinant regions. In the first case, this

gene pool behaves as an outgroup but when recombination is excluded, the evidence of its derivation from the Mesoamerican gene pool is clear.

In addition to shed light on the origin of common bean, this work represents an interesting example of the effect of recombination in phylogenetic analyses, confirming the key role of chloroplast genomes in this kind of studies.

Chapter 1

General Introduction

Knowledge about the origin, evolution and diffusion of crop species is a crucial aspect for their appropriate use and conservation. As pointed out by Gepts (1990), the evaluation of wild germplasm and its incorporation into breeding programs represents a resource for the recovery of genetic diversity of crop species. Indeed, strong reduction of genetic diversity due to the domestication process has been reported for various species, especially autogamous plants such as common bean (Bitocchi et al. 2013), chickpea (Abbo, Berger, and Turner 2003), soybean (Lam et al. 2010), rice (Xu et al. 2012) and wheat (Reif et al. 2005). Thus, the study, exploration and maintenance of wild forms plays a key role in the development of new varieties (Plucknett et al. 1987).

1.1 Phaseolus Genus

According to Delgado-Salinas et al. (2006), the monophyletic genus *Phaseolus* consists of ~70 species, which can be grouped into two major sister clades: A and B. Clade A is characterized by wild species mostly distributed in the higher elevations of Mexico and ascribable to the well resolved *Pauciflorus*, *Pedicellatus*, *Tuerckheimii* groups and other weakly resolved species (i.e. *P. glabellus*, *P. macrolepis*, *P. microcarpus* and *P. oaxacanus*) (Delgado-Salinas et al. 2006). Conversely, species belonging to clade B have a wider geographic distribution from southeastern Canada, south through eastern USA and across southern USA to southeastern California, through-out Mexico and central and South America. Clade B comprises the groups of *Filiformis*, *Vulgaris*, *Lunatus*, *Leptostachyus* and *Polystachios* (Delgado-Salinas et al. 2006). Both *Vulgaris* and *Lunatus* include domesticated species such as *P. vulgaris*, *P. coccineus*, *P. acutifolius*, *P. domosus* and *P. lunatus* (Freytag and Debouck 2002). The diversification of genus *Phaseolus* seems to be occurred between 4 and 6 million years ago (Mya) in Mesoamerica. *Vulgaris* group has been reported to be the oldest group, indeed its formation is dated at ~4 Mya (Delgado-Salinas et al. 2006). Within this group the species *P. vulgaris*, *P. coccineus* and *P. dumosus* are closely related to the point of being partially intercrossable when *P. vulgaris* is the female parent (Mendel, 1866 ; Flow & Wall, 1970; Shii et al., 1982; Hucl, 1985).

1.2 Phaseolus vulgaris

Common bean (*Phaseolus vulgaris L.*) is a diploid (2n=2x=22) and autogamous species, which belongs to the *Fabaceae* family, *Phaseolus* genus. Its peculiar evolutionary history and the fact that is the main grain legumes for human utilization make common bean one of the most interesting species to study.

Wild forms of common bean grow across a wide geographic area of the so called Latin America, from Northern Mexico to Northwestern Argentina (Toro Chica, Tohme, and Debouck 1990). Three eco-geographical gene pools have been identified: the Mesoamerican one (distributed in Mexico, Central America, Colombia and Venezuela), the Andean one (distributed in South Peru, Bolivia and Argentina), and the population from North-Peru and Ecuador (located in the western side of the Andes) that was initially identified in 1986 and reported in the work of Debouck et al. (1993). The formers populations include both wild and domesticated forms, instead only wild forms have been found belonging to the North-Peru and Ecuador gene pool. Indeed, common bean is characterized by a unique evolutionary scenario, in which two geographically distinct and isolated evolutionary lineages predate domestication (Mesoamerican and Andean).
Even though many studies have been published, the origin of common bean is still debated, and various hypothesis have been proposed (Figure 1.1).

1.3 The origin of common bean through the study of genetic diversity

The first hypothesis, proposed by Kami et al. (1995), located the center of origin of *P. vulgaris* in an area geographically intermediate between that of the Mesoamerican and Andean gene pools, speculating that wild beans from North Peru and Ecuador constituted the ancestral genotype. Thus, they suggested that the wild bean was dispersed from the western slopes of the Andes in Northern Peru and Ecuador to north Mesoamerica and south Andes, resulting in the Mesoamerican and Andean gene pools, respectively.
This study represents one of the major achievements for the investigation of *P. vulgaris* genetic structure. Indeed, the authors analyzed the diversity of Phaseolin, the main seed storage protein of common bean, through the implementation of a highly reproducible PCR test conducted on 15-bp and 21-bp tandem direct repeats characteristic of certain gene families that encode for the phaseolin types. The most common phaseolin proteins are S and T type (Brown et al. 1982;

Gepts et al. 1986), mainly found in Mesoamerican and Andean accessions, respectively. The main outcome supporting the North Peru-Ecuador origin of common bean is the finding of a new phaseolin type named I (from Inca, PhI) found only in accession from this area and characterized by the absence of both 15-bp and 21-bp tandem direct repeats. Since duplication that generate tandem direct repeats are more likely to occur than deletions of the sites of the tandem direct repeats, the phaseolin-I has been recognized as the ancestral protein. Additional arguments in favor of this hypothesis included isozyme data that showed that wild common beans from this area are distinct and intermediate between the accessions from Mesoamerica and Andes (Koenig & Gepts, 1989; Debouck et al., 1993)

The development of amplified fragment length polymorphisms (AFFLP), which can be considered as the first class of genome-wide markers, allowed to deeper investigate the genetic diversity of common bean, calling into question the evolutionary scenario proposed by Kami et al. (1995). As emphasized in Cortinovis et al. (2020), markers with different mutation rates can highlight very different patterns of molecular diversity even in the same species or population. This evidence is due to the inverse correlation between the number of generations needed to recover the diversity loss after a bottleneck and the mutation rate (Nei et al., 1975; Nei, 2005). In the work of Rossi et al. (2009) a set of 183 accessions of wild and domesticated common bean representing the geographic distribution of *P. vulgaris*, including accessions belonging to the three main gene pools: Mesoamerican, Andean and North Peru-Ecuador (PhI), was analyzed. A large set of 418 AFLP markers was identified and used to explore the genetic diversity of the whole sample. Overall, a higher genetic diversity was found in the Mesoamerican gene pool (1.6-fold higher) compared to the Andean one and this result was still significant examining only wild genotypes. Conversely, the results obtained by Kwak et al. (2009) using SSR markers did not highlight a strong, but still presents, differentiation, in terms of genetic diversity, between Mesoamerican and Andean wild gene pools. As explained by Rossi et al. (2009), those differences are ascribable to the strong association between genetic diversity and mutation rate of the specific markers used. Based on (i) the higher genetic diversity found in the Mesoamerican gene pool and (ii) the closer proximity of the PhI samples from North Peru-Ecuador to the Mesoamerican gene pool (Rossi et al., 2009; Kwak et al., 2008), a Mesoamerican origin of common bean was suggested by Rossi et al. (2009). In addition, as described by the model of Nei et al. (1975), this result implied that an event corresponding to a strong bottleneck occurred before domestication in the Andean population.

Strong support to this hypothesis has been provided by the work of Bitocchi et al. (2012), in which the authors investigated the nucleotide diversity for a set of five genes in a wide sample (105 accessions) of wild *P. vulgaris*, representative of its entire geographical distribution and the three gene pools. The analysis of 4 legume specific gene fragments (Leg044, Leg100, Leg133, Leg223) and PvSHP1, homologous to SHATTERPROOF gene of *A. thaliana*, stressed the reduction of genetic diversity of the Andean gene pool. Indeed, the lower mutation rate, characteristic of SNP (single nucleotide polymorphism) markers compared to multilocus molecular markers, empowered the detection of genetic loss, resulting in a reduction of genetic diversity of 90% in the Andean gene pool respect to the Mesoamerican sample. Moreover, structure and phylogenetic analyses revealed both a strong population structure of the Mesoamerican gene pool with the lack of a clear distinction between the Mesoamerican and Andean gene pools, and the presence of two Mesoamerican sub-groups from Mexico closer to the North Peru-Ecuador and the Andean populations. Evidence of high diversity and strong population structure in the Mesoamerican gene pool have also been seen in Gepts et al. (1986); Singh (1989); Kwak et al. (2009); Cortés et al. (2011); Desiderio et al. (2013); Bellucci et al. (2014); Goretti et al. (2014); Schmutz et al. (2014); Ariani et al. (2018). Thus, one of the possible explanations to the results obtained by Bitocchi et al. (2012) is given by the migration of beans from Mesoamerica to the South of the country, through different events, leading the formation of the North Peru-Ecuador gene pool and Andean gene pool. Indeed, a scenario in which the origin of common bean would occur in South Latin America, would be reflected in PhI accessions being intermediates between Mesoamerican and Andean samples.

Phylogenetic analyses have always been subjected to bias due to recombination events, especially if performed with nuclear data.

To overcome this limit, an approach widely used in plant studies is to use organelle genomes such as chloroplast. Indeed, chloroplast characteristics represented by haploidy, uniparental inheritance and lack of recombination make this organelle suitable for population genetics and evolutionary and phylogenetic studies (Provan et al. 2001).

Even though in the work of Bitocchi et al. (2012) fragments of a few hundreds of base pairs were used to prevent that recombination affected the data, further investigations on the origin of common bean have been carried out by Desiderio et al. (2013) at chloroplast level. A set of 17 chloroplast microsatellites (cpSSR) was used to explore plastome diversity of a wide sample of wild *P. vulgaris* from the Americas to compare the results to those obtained from nuclear nucleotide data. Consistently with the evidence from nuclear genome, a reduction of genetic

diversity (26%) in the Andean gene pool compared to that of Mesoamerica was observed also in chloroplasts, providing additional evidence of a bottleneck occurred in the population from the Andes, before domestication. The analysis of the relationship between *P. vulgaris* samples and the genetic divergence estimated by F_{st}, D and R_{st} measures highlighted a non-significant differentiation of PhI and Mesoamerican populations (F_{st}= 0.08; R_{st}= 0.12), conversely the greater and significant differentiation was observed between the PhI and Andean gene pools (F_{st}= 0.21; R_{st}= 0.70). Despite the analysis of chloroplast genome data could revealed different evolutionary processes potentially in contrast with those observed from the study of nuclear data, the results found by Desiderio et al. (2013) strongly confirm the hypothesis of a Mesoamerican origin of common bean already proposed by Rossi et al. (2009) and Bitocchi et al. (2012). Moreover, the strong subdivision of the Mesoamerican population also at plastid level with high presence of genetic groups from Central Mexico supports this area as the cradle of *P. vulgaris* diversity. Consequently, the presumed center of domestication of Mesoamerican common bean has been pointed to be the Lerma Santiago Basin (Kwak et al. 2009) or the Oaxaca valley (Bitocchi et al., 2013; Rodriguez et al., 2016).

More recently, the analyses of whole genome nuclear, chloroplast and metabolomic data of wild and domesticated common bean (Rendón-Anaya, Montero-Vargas, et al. 2017) have raised the question about the origin of common bean one more time. The hypothesis advanced and defined as "Pseudovulgaris hypothesis" identifies the PhI gene pool from North Peru and Ecuador as a separate lineage within the *Vulgaris* group. Thus, the PhI gene pool has been formalized as distinct species called *Phaseolus debouckii* (Rendón-Anaya, Herrera-Estrella, et al. 2017). Indeed, exploration of genetic diversity of nuclear data from WGS of 29 *Phaseolus* samples comprising wild and domesticated *P. vulgaris* accessions from Mesoamerica, Andes and North Peru and Ecuador revealed that the PhI accessions showed the lowest absolute pairwise genetic divergence among all comparison (d_{xy}=0.0023) (Rendón-Anaya, Montero-Vargas, et al. 2017). Moreover, the differences between intra-species (Mesoamerican/Andean) and inter-species (PhI/Mesoamerican and PhI/Andean) distances hinted the derivation from different populations. Thus, the divergence of the PhI group from *P. vulgaris* and from the other member of the *Vulgaris* group (*P. domosus, P.coccineus, P. costaricensis*) is such that PhI subpopulation could represent a different lineage. Phylogenetic analyses based on WGS nuclear data and 55-kb chloroplast genome fragment (cpDNA) enhanced the "Pseudovulgaris hypothesis" placing the North Peruvian Ecuadorian accessions in a separate clade sister to *P. vulgaris*, contrary to previous works where PhI samples derived from the Mesoamerican gene

pool (Bitocchi et al. 2012). A similar pattern was also showed by metabolomic analysis which placed the PhI group as an intermediate between *P. vulgaris* (Mesoamerican and Andean gene pools) and *P. coccineus*.

Those results were corroborated by coalescent simulations performed with nuclear and chloroplast data on a subset of individuals (G21245, wild accession from North Peru Ecuador, BAT93, domesticated accession from Mesoamerica, and JaloEEP558, domesticated accession from Andes). Both simulations highlighted the early divergence of the PhI group (0.9 Mya from plastid data and 0.26 Mya from nuclear data) compared to the split between Mesoamerican and Andean gene pools (0.2 Mya from plastid data and 0.002 Mya form nuclear data). With regard to the Mesoamerican and Andean populations, the estimated divergence time based on 55-kb chloroplast genome fragment reported by Rendón-Anaya, Montero-Vargas, et al. (2017) was much earlier than that proposed by both Schmutz et al. (2014) (165.000 years ago) and Mamidi et al. (2013) (110.000 years ago); on the other hand, the estimation based on nuclear data ensued to be much later. Thus, even though the Mesoamerican origin of common bean was supported by the results of Rendón-Anaya, Montero-Vargas, et al. (2017), the authors introduced a new interpretation of the data, proposing that an early speciation event occurred in the area of western Andes. To explain the data, two models of migration were proposed by the authors: (i) a "two-waved" migration event likely occurred through seed dispersal favored by birds, in which *P. vulgaris* spread from Mesoamerica to North Peru-Ecuador where it was subjected to isolation and consequently underwent allopatric speciation. Subsequently, a small population of the Mesoamerican gene pool migrated to the South mainly in center and South Andes leading the formation of the Andean gene pool. (ii) Glacial periods occurred in South Andes during Pleistocene could have limited the gene flow between *Phaseolus* populations, allowing the isolation and diversification of PhI populations. Instead, the Andean gene pool could have been originated from a small founder population of *P. vulgaris*.

Moreover, the gene flow, also reported by Rendón-Anaya, Montero-Vargas, et al. (2017) between the North Peru Ecuador population (*P. debouckii*) and *P. vulgaris* could be due to the reasonably recent speciation event. Indeed, reproductive barriers could not have been completely established even though the geographic separation forced by Andes Mountains could have limited the outcrossing with the Mesoamerican gene pool.

Ariani et al. (2018) provided further evidence for the distinctness and older age of the PhI group compared to the Mesoamerican and Andean groups. Phylogenetic analyses, based on genic and

non-genic variants, revealed the intermediate behavior of PhI accessions. Indeed, the PhI group was placed between *P. coccineus* and *P. vulgaris*, resulting to be a different taxon with no nesting inside the *P. vulgaris* clade or even within the Mesoamerican group. Nevertheless, coalescent simulations and investigation of genetic diversity in the three gene pools agreed with the Mesoamerican origin of common bean. However, the reported results allowed the authors to speculate about the existence of an ancestral population, phenotypically similar to *P. vulgaris* but also carrying the type I of the phaseolin protein (Figure 1.1, panel c, d, and e). This common ancestor, located in Mesoamerica, remains to be discovered or it may be extinguished before the Mesoamerican and Andean diversification (Protovulgaris hypothesis). Hypothesis about the contemporary distribution of common bean was also advanced. Since Mesoamerica has been identified as the core area of the entire *Phaseolus* genus (Delgado-Salinas et al. 1999; Freytag and Debouck 2002), the presence of species outside of this area can be explained invoking Long Distance Dispersal events. Thus, three spatial scale of seed dispersal were proposed, each of them associated with its own temporal scale. Pod shattering would cover seed dispersal at short distances, rodents, birds, and megafauna are potential agents for the dispersion of seeds at medium range leading the formation of the structured distribution of each of the three gene pools. Instead, long distance dispersal likely shaped the current distribution of wild *P. vulgaris*.

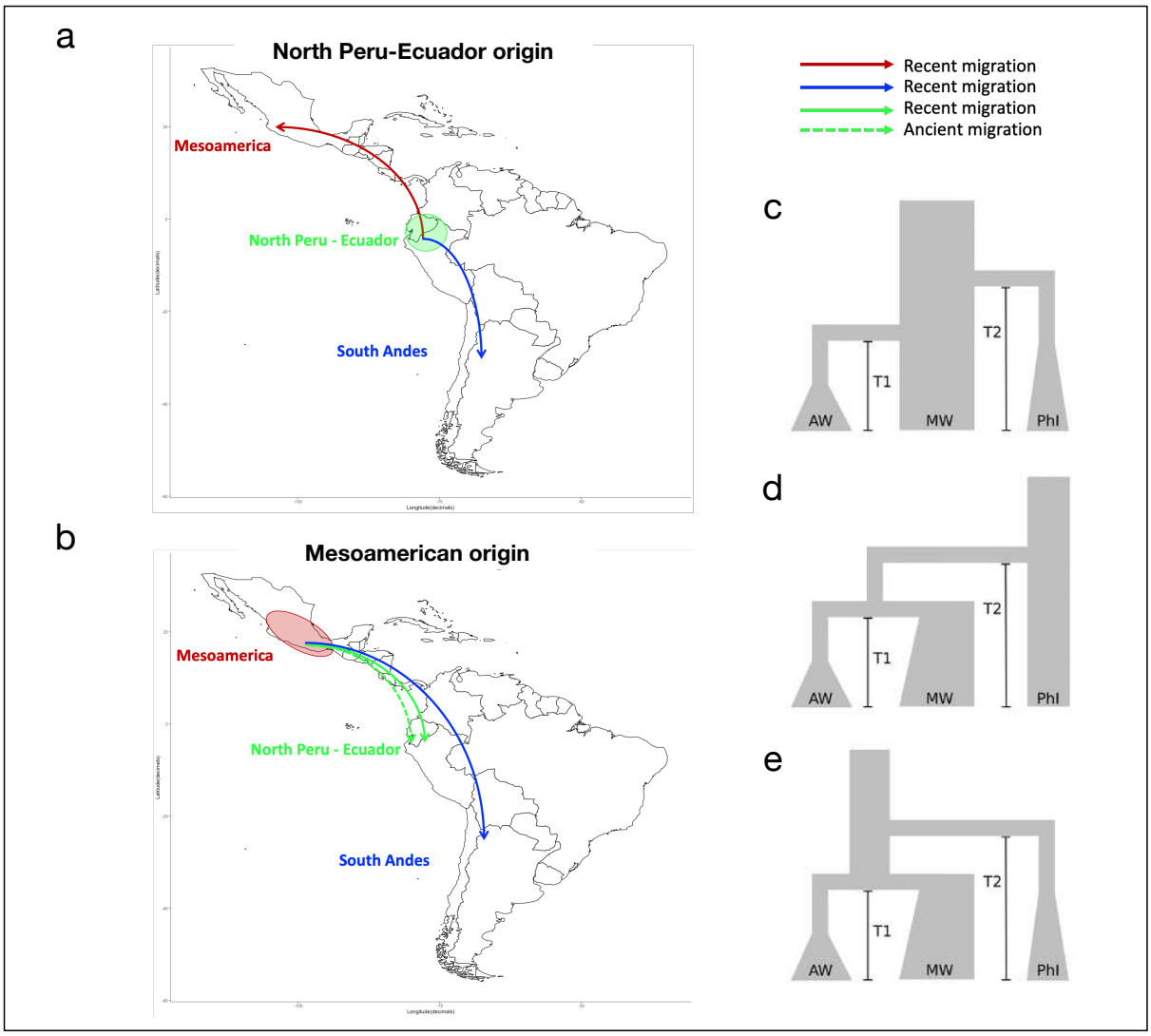


Figure 1.1 Graphical representation of the different hypotheses on the origin of common bean. Panel a: P vulgaris originated in North Peru-Ecuador and subsequently migrated in Mesoamerica and South Andes. Panel b: P. vulgaris originated in Mesoamerica and spread in South America. Panel c,d, and e are from Ariani et al., (2018) *and are the graphical representation of the different demographic models performed in* Ariani et al., (2018). *(c) Mesoamerican model where the Mesoamerican wild (MW) population did not experience any population bottleneck; (d) the Northern Peru–Ecuador model where the Northern Peru–Ecuador (PhI) gene pool did not experience any population bottleneck; and (e) the Protovulgaris model where the ancestral population went extinct after the Mesoamerican and Andean differentiation.*

1.4 The aim and the objectives of the research

Considering all the studies carried out so far and summarized in the previous paragraphs (Kami et al., 1995; Rossi et al., 2009; Bitocchi et al., 2012; Desiderio et al., 2013; Rendón-Anaya, Montero-Vargas, et al., 2017; Ariani et al., 2018), it is evident how the origin of common bean is still an open topic for discussion. The present project aims at clarifying the phylogeny of common bean and thus, the relationships among the three main gene pools. The origin of

common bean was investigated by assessing patterns of nucleotide variability using both plastid and nuclear data.

References

Abbo, S.l, J. B., and N. C. Turner. 2003. “Evolution of Cultivated Chickpea: Four Bottlenecks Limit Diversity and Constrain Adaptation.” *Functional Plant Biology* 30(10):1081–87.

Ariani, A., Jorge C. B. Mier Teran, and P. Gepts. 2018. “Spatial and Temporal Scales of Range Expansion in Wild Phaseolus Vulgaris.” *Molecular Biology and Evolution* 35(1):119–31. doi: 10.1093/molbev/msx273.

Bellucci, E., E. Bitocchi, A. Ferrarini, A. Benazzo, E. Biagetti, S. Klie, A. Minio, D. Rau, M. Rodriguez, A. Panziera, L. Venturini, G.a Attene, E. Albertini, S. A. Jackson, L.Nanni, A. R. Fernie, Z. Nikoloski, G. Bertorelle, M. Delledonne, and R. Papa. 2014. “Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean.” *Plant Cell* 26(5):1901–12. doi: 10.1105/tpc.114.124040.

Bitocchi, E., E. Bellucci, A. Giardini, D. Rau, M. Rodriguez, E. Biagetti, R. Santilocchi, P. Spagnoletti Zeuli, T. Gioia, G. Logozzo, G. Attene, L. Nanni, and R. Papa. 2013. “Molecular Analysis of the Parallel Domestication of the Common Bean (Phaseolus Vulgaris) in Mesoamerica and the Andes.” *New Phytologist* 197(1):300–313. doi: 10.1111/j.1469-8137.2012.04377.x.

Bitocchi, E., L. Nanni, E.Bellucci, M. Rossi, A. Giardini, P.Spagnoletti Zeuli, G. Logozzo, J. Stougaard, P. McClean, G. Attene, and R. Papa. 2012. “Mesoamerican Origin of the Common Bean (Phaseolus Vulgaris L.) Is Revealed by Sequence Data.” *Proceedings of the National Academy of Sciences of the United States of America* 109(14). doi: 10.1073/pnas.1108973109.

Brown, J. W. S., J. R. McFerson, F. A. Bliss, and T. C. Hall. 1982. “Genetic Divergence among Commercial Classes of Phaseolus Vulgaris in Relation to Phaseolin Pattern.” *HortScience,* 17(5):752–54.

Cortés, A. J., M. C. Chavarro, and M. W. Blair. 2011. “SNP Marker Diversity in Common Bean (Phaseolus Vulgaris L.).” *Theoretical and Applied Genetics* 123(5):827–45. doi: 10.1007/s00122-011-1630-8.

Cortinovis, G., G. Frascarelli, V. di Vittori, and R. Papa. 2020. “Current State and Perspectives in Population Genomics of the Common Bean.” *Plants* 9(3). Doi: 10.3390/plants9030330

Debouck, D. G., O. Toro, O. M. Paredes, W. C. Johnson, and P. Gepts. 1993. “Genetic Diversity and Ecological Distribution of Phaseolus Vulgaris (Fabaceae) in Northwestern South America.” *Economic Botany* 47(4):408–23. doi: https://doi.org/10.1007/BF02907356.

Delgado-Salinas, A., R. Bibler, and M. Lavin. 2006. *Phylogeny of the Genus Phaseolus (Leguminosae): A Recent Diversification in an Ancient Landscape*. Vol. 04.

Delgado-Salinas, A., T. Turley, A. Richman, and M. Lavin. 1999. “Phylogenetic Analysis of the Cultivated and Wild Species of Phaseolus (Fabaceae).” *Systematic Botany* 24(3):438–60. doi: https://doi.org/10.2307/2419699.

Desiderio, F., E.Bitocchi, E. Bellucci, D. Rau, M. Rodriguez, G. Attene, R. Papa, and L. Nanni. 2013. “Chloroplast Microsatellite Diversity in Phaseolus Vulgaris.” *Frontiers in Plant Science* 3(JAN). doi: 10.3389/fpls.2012.00312.

Wall, J.R., 1970. Experimental introgression in the genus Phaseolus. I. Effect of mating systems on interspecific gene flow. Evolution, pp.356-366.

Freytag, G. F., and D. G. Debouck. 2002. “Taxonomy, Distribution, and Ecology of the Genus Phaseolus (Leguminosae–Papilionoideae) in North America. Mexico and Central America.” *Botanical Research Institute of Texas, Forth Worth.* 23:298.

Gepts, P., T. C. Osborn, K. Rashr~a, and F. A. Bliss. 1986. “Phaseolin-Protein Variability in Wild Forms and Landraces of the Common Bean (Phaseolus Vulgaris): Evidence for Multiple Centers of Domestication I.” *Econ Bot* 40:461–68. doi: https://doi.org/10.1007/BF02859659.

Gepts, P. 1990. *Biochemical Evidence Bearing on the Domestication of Phaseolus (Fabaceae) Beans I*. Vol. 44.

Goretti, D., E. Bitocchi, E. Bellucci, M. Rodriguez, D. Rau, T. Gioia, G. Attene, P. McClean, L. Nanni, and R. Papa. 2014. “Development of Single Nucleotide Polymorphisms in Phaseolus Vulgaris and Related Phaseolus Spp.” *Molecular Breeding* 33(3):531–44. doi: 10.1007/s11032-013-9970-5.

Hucl, P. ,. and Graham J. Scoles. 1985. “Interspecific Hybridization in the Common Bean: A Review.” *HortScience 20.3* 352–57.

Kami, J., V. Becerra Velasquez, D. G. Debouck, P. Gepts, and R. W. Allard. 1995. *Identification of Presumed Ancestral DNA Sequences of Phaseolin in Phaseolus Vulgaris (Molecular Evolution/Seed Protein/Crop Evolution/Tandem Repeat/Polymerase Chain Reaction) Communicated By*. Vol. 92. doi: https://doi.org/10.1073/pnas.92.4.1101.

Koenig, R., and P. Gepts. 1989. *Allozyme Diversity in Wild Phaseolus Vulgaris: Further Evidence for Two Major Centers of Genetic Diversity*. Vol. 78. Springer-Verlag.

Kwak, M., J. A. Kami, and P. Gepts. 2009. “The Putative Mesoamerican Domestication Center of Phaseolus Vulgaris Is Located in the Lerma-Santiago Basin of Mexico.” *Crop Science* 49(2):554–63. doi: 10.2135/cropsci2008.07.0421.

Kwak, M., D.Velasco, and P. Gepts. 2008. “Mapping Homologous Sequences for Determinacy and Photoperiod Sensitivity in Common Bean (Phaseolus Vulgaris).” *Journal of Heredity* 99(3):283–91. doi: 10.1093/jhered/esn005.

Lam, H. M., X. Xu, X. Liu, W. Chen, G. Yang, F. Ling Wong, M. Wah Li, W. He, N. Qin, B. Wang, J. Li, M. Jian, J. Wang, G. Shao, J. Wang, S.Sai Ming Sun, and G. Zhang. 2010.

"Resequencing of 31 Wild and Cultivated Soybean Genomes Identifies Patterns of Genetic Diversity and Selection." *Nature Genetics* 42(12):1053–59. doi: 10.1038/ng.715.

Mamidi, S., M. Rossi, S. M. Moghaddam, D. Annam, R. Lee, R. Papa, and P. E. McClean. 2013. "Demographic Factors Shaped Diversity in the Two Gene Pools of Wild Common Bean Phaseolus Vulgaris L." *Heredity* 110(3):267–76. doi: 10.1038/hdy.2012.82.

Mendel, G. 1866. "Versuche Über Pflanzenhybriden Verh." *Des Naturf. Vereines in Brünn (Abhandlungen)* 4:3–47.

Nei, M. 2005. *Bottlenecks, Genetic Polymorphism and Speciation*. Genetics, Volume 170, Issue 1, 1–4, https://doi.org/10.1093/genetics/170.1.1

Nei, M., T. Maruyama, and R. Chakraborty. 1975. *The Bottleneck Effect and Genetic Variability in Populations*. Vol. 29.

Plucknett, D. L., N. Smith, J. Williams, and A. N. Murthi. 1987. *Gene Banks and the World's Food*. Princeton, NJ: Princeton University Press.

Provan, J., W.Powell, P. Hollingsworth, and P. M. Hollingsworth. 2001. *Review Review Chloroplast Microsatellites: New Tools for Studies in Plant Ecology and Evolution*. Vol. 16.

Reif, J. C., P. Zhang, S. Dreisigacker, M. L. Warburton, M. van Ginkel, D. Hoisington, M. Bohn, and A. E. Melchinger. 2005. "Wheat Genetic Diversity Trends during Domestication and Breeding." *Theoretical and Applied Genetics* 110(5):859–64. doi: 10.1007/s00122-004-1881-8.

Rendón-Anaya, M., A. Herrera-Estrella, P. Gepts, and A. Delgado-Salinas. 2017. "A New Species of Phaseolus (Leguminosae, Papilionoideae) Sister to Phaseolus Vulgaris, the Common Bean." *Phytotaxa* 313(3):259–66. doi: 10.11646/phytotaxa.313.3.3.

Rendón-Anaya, M., J.M. Montero-Vargas, S. Saburido-Álvarez, A. Vlasova, S.Capella-Gutierrez, J. Juan Ordaz-Ortiz, O. M. Aguilar, R. P. Vianello-Brondani, M. Santalla, L. Delaye, T. Gabaldón, P. Gepts, R. Winkler, R. Guigó, A. Delgado-Salinas, and A. Herrera-Estrella. 2017. "Genomic History of the Origin and Domestication of Common Bean Unveils Its Closest Sister Species." *Genome Biology* 18(1). doi: 10.1186/s13059-017-1190-6.

Rodriguez, M., D. Rau, E. Bitocchi, E. Bellucci, E. Biagetti, A. Carboni, P. Gepts, L. Nanni, R. Papa, and G. Attene. 2016. "Landscape Genetics, Adaptive Diversity and Population Structure in Phaseolus Vulgaris." *New Phytologist* 209(4):1781–94. doi: 10.1111/nph.13713.

Rossi, M., E. Bitocchi, E. Bellucci, L. Nanni, D. Rau, G. Attene, and R. Papa. 2009. "Linkage Disequilibrium and Population Structure in Wild and Domesticated Populations of Phaseolus Vulgaris L." *Evolutionary Applications* 2(4):504–22. doi: 10.1111/j.1752-4571.2009.00082.x.

Schmutz, J., P. E. McClean, S. Mamidi, G. A. Wu, S. B. Cannon, J. Grimwood, J. Jenkins, S. Shu, Q. Song, C. Chavarro, M. Torres-Torres, V. Geffroy, S. Mafi Moghaddam, D. Gao, B. Abernathy, K. Barry, M. Blair, M. A. Brick, M. Chovatia, P. Gepts, D. M. Goodstein, M. Gonzales, U. Hellsten, D. L. Hyten, G. Jia, J. D. Kelly, D. Kudrna, R. Lee, M. M. S. Richard, P. N. Miklas, J. M. Osorno, J. Rodrigues, V. Thareau, C. A. Urrea, M. Wang, Y. Yu, M. Zhang, R. A. Wing, P. B. Cregan, D. S. Rokhsar, and S. A. Jackson. 2014. "A Reference Genome for Common Bean and Genome-Wide Analysis of Dual Domestications." *Nature Genetics* 46(7):707–13. doi: 10.1038/ng.3008.

Shii, L. C. T., A. Rabakoarihanta, M. C. Mok, and D. W. S. Mok. 1982. "Embryo Development in Reciprocal Crosses of Phaseolus Vulgaris." *Theor. Appl. Genet* 62:59–64.

Singh, S. P. 1986. "Patterns of Variation in Cultivated Common Bean (Phaseolus Vulgaris, Fabaceae)." *Econ Bot* 43:39–57. doi: https://doi.org/10.1007/BF02859324.

Toro Chica, O., J. M. Tohme, and D. G. Debouck. 1990. *Wild Bean (Phaseolus Vulgaris L.): Description and Distribution*. Vol. 181. CIAT.

Xu, X., X. Liu, S. Ge, J. D. Jensen, F. Hu, X. Li, Y. Dong, R. N. Gutenkunst, L. Fang, L. Huang, J. Li, W. He, G. Zhang, X. Zheng, F. Zhang, Y. Li, C. Yu, K. Kristiansen, X. Zhang, J. Wang, M. Wright, S. McCouch, R. Nielsen, J.Wang, and W. Wang. 2012. "Resequencing 50 Accessions of Cultivated and Wild Rice Yields Markers for Identifying Agronomically Important Genes." *Nature Biotechnology* 30(1):105–11. doi: 10.1038/nbt.2050.

Chapter 2

The assembly of 39 plastomes of *Phaseolus spp.* and the construction of *P. vulgaris* pan-plastome as resources for future investigations

Abstract

Chloroplast genomes have a key role in the phylogenetic reconstruction of plant species. Here, we determined the complete sequence of thirty-nine plastomes belonging to four species of the genus *Phaseolus*. A particular attention was paid to common bean accessions (33 samples) which were selected to be as representative as possible of the geographic distribution of the three main gene pools from Mesoamerica, Andes and North Peru-Ecuador. Comparative analysis revealed a high level of conservation of *Phaseolus spp.* plastomes. Nevertheless, we found small deletions in common bean samples from North Peru-Ecuador and one bigger deletion characteristics of *P. acutifolius*. Finally, the chloroplast diversity of *P. vulgaris* was collected in a consensus pan-plastome. The *de-novo* assembled chloroplast genomes and the development of common bean pan-plastome represent an important resource for the study of the remarkable evolutionary history of *Phaseolus vulgaris*.

Introduction

Chloroplasts are organelles characteristic of eucaryotic algae and land plants in which photosynthesis takes place. In addition, plastomes have an important role in producing starch, lipids, essential proteins, vitamins and various flower pigments (Bausher et al. 2006).

"The endosymbiosis theory", firstly proposed by Mereschkowsky (1905), explains the origin of chloroplasts from cyanobacteria through endosymbiosis.

Free-living prokaryotes, settled within primitive eukaryotic cells as permanent intracellular elements, gave rise to eukaryotic organelles such as chloroplasts and mitochondria (Margulis 1970).

As part of the prokaryotic inheritance, plastomes of land plants usually present a circular structure of 120-160 kilobase pairs (kb), genome packaging in nucleoids, organization of genes in operons and a prokaryotic gene expression machinery (Bock 2007). A quadripartite

structure, composed by a large and a small single copy regions (LSC and SSC, respectively) divided by a pair of inverted repeats (IRs), is typical of the circular molecules of Angiosperms. Contrary to the high variability of chloroplast genomes of algae, plastomes of land plants present a similar gene content of 100-120 genes and gene order (Bendich, 2004).

Due to their characteristics such as the haploid genome, the uniparental inheritance (primarily maternal) and absence of recombination, plastomes are valuable for genetic and phylogenetic studies.

Given that comparative analysis of nuclear genomes from multiple individuals of the same species revealed that a single reference genome is inadequate to capture the genetic diversity of a species and more accurate sequencing technologies are now available, pangenomes of numerous important crops have been developed recently: pepper (Ou et al., 2018), soybean (Torkamaneh et al., 2018), rice (Zhao et al., 2018), cucumber (Gao et al., 2019), rape (Song et al., 2020), barley (Jayakodi et al., 2020), cotton (Li et al., 2021), sorghum (Tao et al., 2021), and chickpea (Varshney et al., 2021). Despite the high conservation of chloroplast genomes, pan-plastomes can also represent valuable resources to empower the assessment of genetic variation at plastid level, as reported for *Capsicum* (Magdy et al. 2019).

The common bean (*Phaseolus vulgaris*) is one of the most important grain legumes for human consumption and its unique evolutionary history makes this species a model for understanding crop evolution (Bitocchi et al. 2017). Wild forms of common bean grow across a wide geographic area of the Americas, from Mexico to Northwestern Argentina (Toro Chica, Tohme, and Debouck 1990) and they can be structured in three main gene pools: Mesoamerican, Andean and one from North-Peru and Ecuador. The Mesoamerican and Andean gene pools are geographically distinct and isolated, and they were already present before domestication of common bean. Indeed, both of them include wild and domesticated forms (for review see Cortinovis et al., 2020). The third gene pool, the one from Northern Peru-Ecuador, is represented by only wild populations characterized by an ancestral type of the phaseolin seed storage-protein: the Phaseolin I (Kami et al., 1995).

Even though the physical map of common bean chloroplast genome was published in 1983 (Mubumbila et al. 1983), only recently the complete plastome sequence of *P. vulgaris* cv. Negro Jamapa was published (Guo et al. 2007). The chloroplast described in the work of Guo et al. (2007) is characterized by a circular structure of 150,285 kb containing two identical IRs of 26,426 bp, an LSC of 79,824 bp and an SSC of 17,66 bp. Until now, the complete plastome of common bean was generated only from domesticated accessions (Guo et al. 2007; Meng and

Li 2018) while the study of wild samples may bring a precious contribution to the analysis of genetic variation and phylogenetic relationships among common bean gene pools, as showed for chickpea (Mehmetoglu et al. 2022).

In this work we assembled 33 *de-novo* chloroplast genomes of wild *P. vulgaris*, including accessions from the Mesoamerican, Andean and North Peru-Ecuador gene pools. In addition, plastomes of *P. coccineus*, *P. acutifolius* and *P. lunatus* were also reconstructed and included in the comparative analyses. Finally, the alignment of *P. vulgaris* plastomes was used to generate the pan-plastome of common bean.

Material and Methods

Plant Material and DNA extraction

A total of 33 wild accessions of *Phaseolus vulgaris* were selected to cover the area from Northern Mexico to Northwestern Argentina corresponding to the distribution range of wild common bean. Genotypes were chosen to be representative of the three different gene-pools: 19 Mesoamerican, 8 Andean and 6 from North Peru-Ecuador. According to the information given by the seed providers, the North Peru-Ecuador accessions are characterized by the ancestral phaseolin type, namely Phaseolin type I (PhI) (Debouck et al., 1993 Kami et al., 1995). In addition, wild samples of three other *Phaseolus* species were included: 4 *P. coccineus*, 1 *P. acutifolius*, 1 *P. lunatus*. Seeds were provided by the United States Department of Agriculture Western Regional Plant Introduction Station (USDA) and the International Center of Tropical Agriculture (CIAT) in Colombia.

Genomic DNA was extracted from young leaves of a single plant grown in greenhouse using the DNeasy Plant Mini kit (QIAGEN). DNA libraries were constructed and sequenced from both ends (paired-ends) by using the Illumina Nextera XT sample preparation kit.

To easily associate each accession to the own gene pool and country of origin, a code was developed and assigned to the genotypes (Table 2.1): (i) a unique numeric code for each accession, (ii) species of belonging (Pv: *Phaseolus vulgaris*; Pc: *Phaseolus coccineus*; Pa: *Phaseolus acutifolius*; Pl: *Phaseolus lunatus*), (iii) genepool (M: Mesoamerica; A: Andes; PhI: Phaseolin type I) and/or the corresponding accession status (W: wild; D: domesticated), (iv) country of origin.

Table 2.1 Panel of accessions analyzed.

Project code	Species	Accession Number	Country
010_Pv_MW_MX	*P. vulgaris*	G11050	Mexico
016_Pv_MW_MX	*P. vulgaris*	G12877	Mexico
031_Pv_AW_AR	*P. vulgaris*	G19888	Argentina
044_Pv_MW_MX	*P. vulgaris*	G20515	Mexico
056_Pv_MW_CO	*P. vulgaris*	G22304	Colombia
057_Pv_MW_MX	*P. vulgaris*	G22837	Mexico
059_Pv_MW_CR	*P. vulgaris*	G23418	Costa Rica
069_Pv_MW_CO	*P. vulgaris*	G23462	Colombia
076_Pv_MW_MX	*P. vulgaris*	G23652	Mexico
081_Pv_MW_MX	*P. vulgaris*	G24571	Mexico
205_Pv_MW_MX	*P. vulgaris*	PI417775	Mexico
501_Pv_MW_MX	*P. vulgaris*	PI417671	Mexico
505_Pv_MW_MX	*P. vulgaris*	PI535409	Mexico
506_Pv_MW_MX	*P. vulgaris*	PI535450	Mexico
787a_Pv_MW_GT	*P. vulgaris*	G23439	Guatemala
790_Pv_MW_HN	*P. vulgaris*	G50724	Honduras
835_Pv_MW_MX	*P. vulgaris*	G23551	Mexico
887_Pv_MW_MX	*P. vulgaris*	NI1433	Mexico
911_Pv_MW_MX	*P. vulgaris*	86	Mexico
id837fa7_Pv_MW_MX	*P. vulgaris*	G50899	Mexico
013_Pv_AW_PE	*P. vulgaris*	G12856	Perù
033_Pv_AW_AR	*P. vulgaris*	G19891	Argentina
039_Pv_AW_AR	*P. vulgaris*	G19898	Argentina
062_Pv_AW_PE	*P. vulgaris*	G23422	Peru
668a_Pv_AW_BO	*P. vulgaris*	G23442	Bolivia
715_Pv_AW_PE	*P. vulgaris*	G23456A	Peru
P718a_Pv_AW_PE	*P. vulgaris*	G23419	Peru
054_Pv_PhI_PE	*P. vulgaris*	G21245	Perù
073_Pv_PhI_EC	*P. vulgaris*	G23582	Ecuador
075_Pv_PhI_PE	*P. vulgaris*	G23587	Perù
078_Pv_PhI_EC	*P. vulgaris*	G23726	Ecuador
id837fa5Pv_PhI_PE	*P. vulgaris*	G23420	Peru
id837fa6_Pv_PhI_EC	*P. vulgaris*	G23724	Ecuador
Pl_W_PE	*P. lunatus*	NI1771	Peru
Pc_11_W_MX	*P. coccineus*	PI346950	Messico
Pc_14_W_MX	*P. coccineus*	NI726	Mexico
Pc_18_W_MX	*P. coccineus*	NI1120	Mexico
Pc_3_W_MX	*P. coccineus*	NI677	Mexico
Pa_W_MX	*P. acutifolius var. acutifolius*	PI319445	Mexico

Pre-processing and reference mapping

The raw reads of *Phaseolus* accessions were checked for quality using FastQC (Andrews et al., 2010), before and after the pre-processing. The command line tool Trimmomatic (Bolger et al., 2014) was used to remove Illumina technical sequences and to filter out low quality reads. Reads ≥ 75 nucleotides in length with a minimum Q-value of 20 were retained for down stream analyses. Since both nuclear and plastid DNA was extracted and sequenced, FasQscreen (Wingett and Andrews 2018) and the sequence aligner Bowtie2 (default settings) (Langmead and Salzberg 2012) were used to align the high-quality reads to the nuclear (G19833) and plastid (NC_009259) reference genomes and to extract chloroplast reads. The coverage mapping was calculated using the *-genomecov* option implemented in BEDtools (Quinlan and Hall 2010).

Chloroplast genomes assembly and annotation

The *de-novo* assembly was completed with NOVOPasty (version 3.2; Dierckxsens et al., 2017) using the sequences of *matk*, *accd*, *psbh*, *rrn16* and *rpl32* of *P. vulgaris* as seeds to initialize the process. Genome annotation was carried out using the PGA software (Qu et al. 2019) for common bean samples and the annotation of missing genes was manually curated with BLASTn (Zhang et al. 2000). The circular map of assembled plastome of accession 016_Pv_MW_MX (*P. vulgaris*) was drawn using the online webtools OGDRAW-Draw Organelle Genome Maps (Greiner, Lehwark, and Bock 2019).

Comparative analysis

To visualize the differences among the assembled chloroplast genomes, a multiple alignment was built in mVISTA (Stanford University, Stanford CA, USA) in LAGAN (Limited Area Global Alignment of Nucleotides) mode using *P. vulgaris* published plastome and its annotation as reference (NC_009259). An additional alignment was carried out using the MAUVE alignment software (Darling et al. 2004). This alignment was performed on a subset of plastomes, chosen based on the gene content differences revealed by the annotation, using the progressive MAUVE option.

Analysis of nucleotide diversity

In order to determine nucleotide diversity in the 39 *de-novo* assembled plastomes a multi-sequence alignment (MSA) was performed using MAFFT (Katoh et al., 2018) with default parameters. The MSA was used as input for DNAsp (Rozas et al., 2017) and nucleotide diversity was calculated. A sliding window of 200 bp with 50 bp step size was used to summarize the diversity statistics for visualization.

Pan-plastome development

The de-novo assembled chloroplast genomes of *P. vulgaris* accessions were used to develop a wild common bean pan-plastome. The thirty-three sequences of *P. vulgaris* were aligned with MAFFT (Katoh et al. 2018) and collapsed using the EMBOSS package (Rice, Longden, and Bleasby 2000) with the option *-cons*. The pan-plastome was annotated with PGA software (Qu et al. 2019) and the annotation of missing genes weas manually curated with BALSTn (Zhang et al. 2000). OGDRAW (Greiner et al. 2019) was used to produce a map of the consensus pan-plastome. The sequence was analyzed with DNAsp (Rozas et al., 2017) and tandem repeats were investigated with Tandem Repeats Software (Benson, 1999)

Results

Chloroplast genomes organization

In this study, a set of 39 chloroplast genomes of *Phaseolus* species (*P. vulgaris*, *P. coccineus*, *P. acutifolius*, *P. lunatus*) were *de-novo* assembled. Raw sequences of the accessions were mapped to the reference plastome of *P. vulgaris* (NC_009259). An average of 74055 reads per sample were mapped to the reference with an average base coverage of 253. Since the chloroplast DNA was not isolated during the extraction, most of the reads were successfully aligned only to the nuclear genome, whilst only 22,25% of the sequenced reads were of plastid origin. From the *de-novo* assembly, we obtained 39 complete plastomes with a genome size average of 150,466 bp (Table 2.2). As expected, all *Phaseolus* plastomes presented a quadripartite structure including a large single copy (LSC) ranging from 79669 bp (*P. acutifolius*) to 81619 bp (*P. lunatus*), a small single copy (SSC) from 17583 bp to 18260 bp divided by two inverted repeats (IR, 26387 bp min - 26539 bp max). Overall chloroplast genomes, the GC (*Guanine or Cytosine*) content ranged from 35.38% to 35.47% (Table 2.2). Common bean plastomes were annotated using the *P. vulgaris* chloroplast genome (NC_009259) as reference. A total of 111 unique genes were found in all *P. vulgaris*

chloroplast genomes, 68 protein coding genes, 39 transfer RNA and 8 ribosomal RNA. Eighteen genes, located in IR regions, were duplicated (Figure2.1).

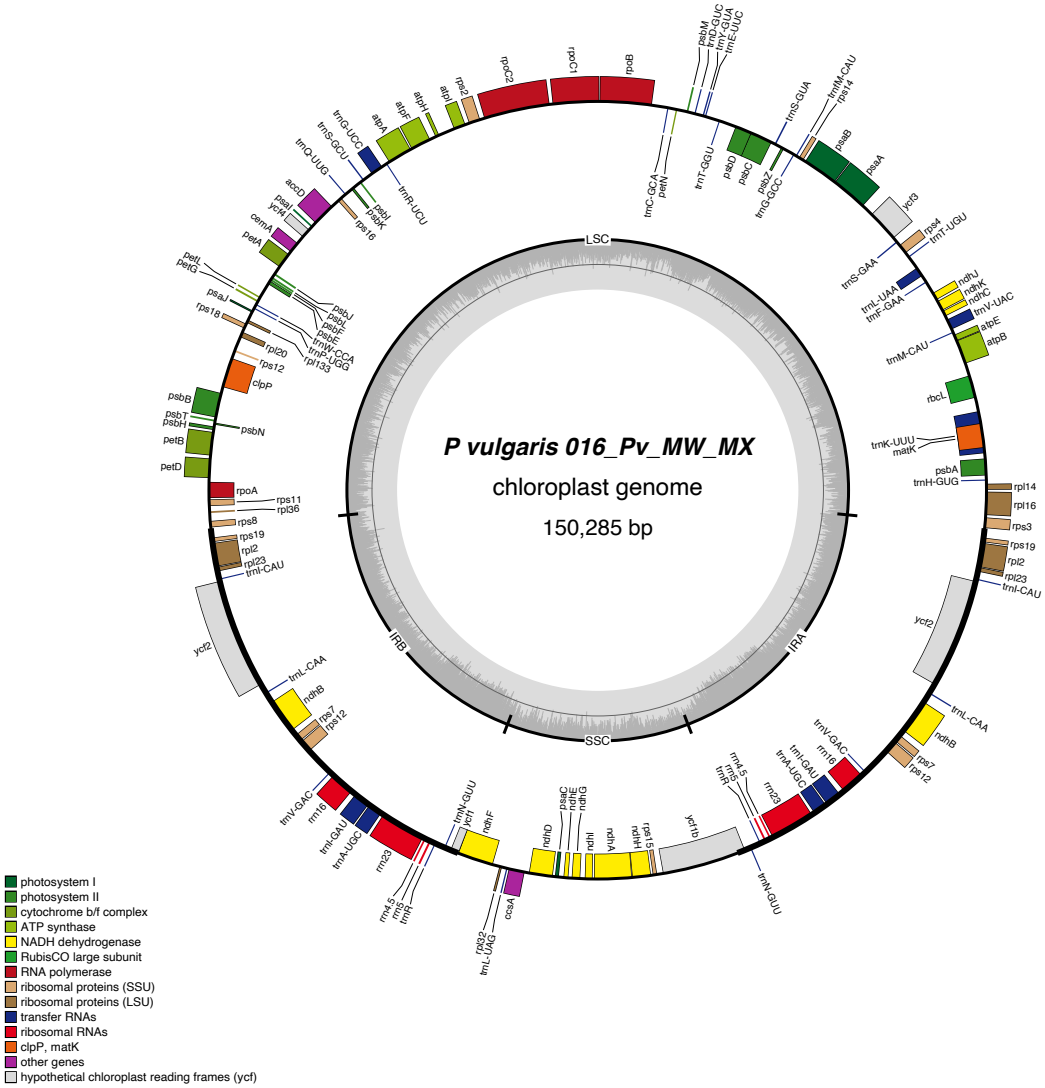


Figure 2.1 Map of the chloroplast genome of the accessions 016_Pv_MW_MX (Phaseolus vulgaris). Genes inside of the outer circle are transcribed in the clockwise direction, while those outsides are transcribed in the counterclockwise direction. Different color codes represent genes belonging to various functional groups. The circle inside represents GC content graph with the 50% threshold. The inverted repeat, large single-copy, and small single-copy regions are denoted by IR, LSC, and SSC, respectively.

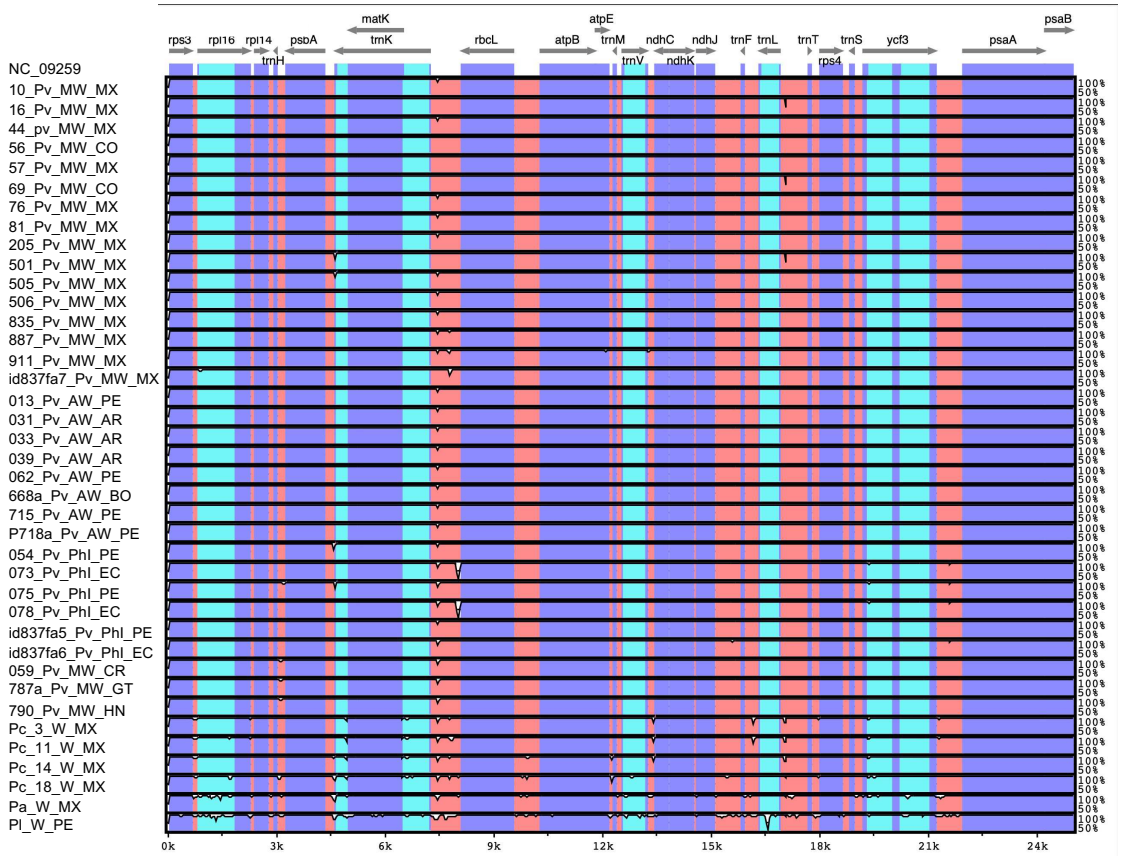
Table 2.2 Plastome features of the thirty-nine Phaseolus genotypes.

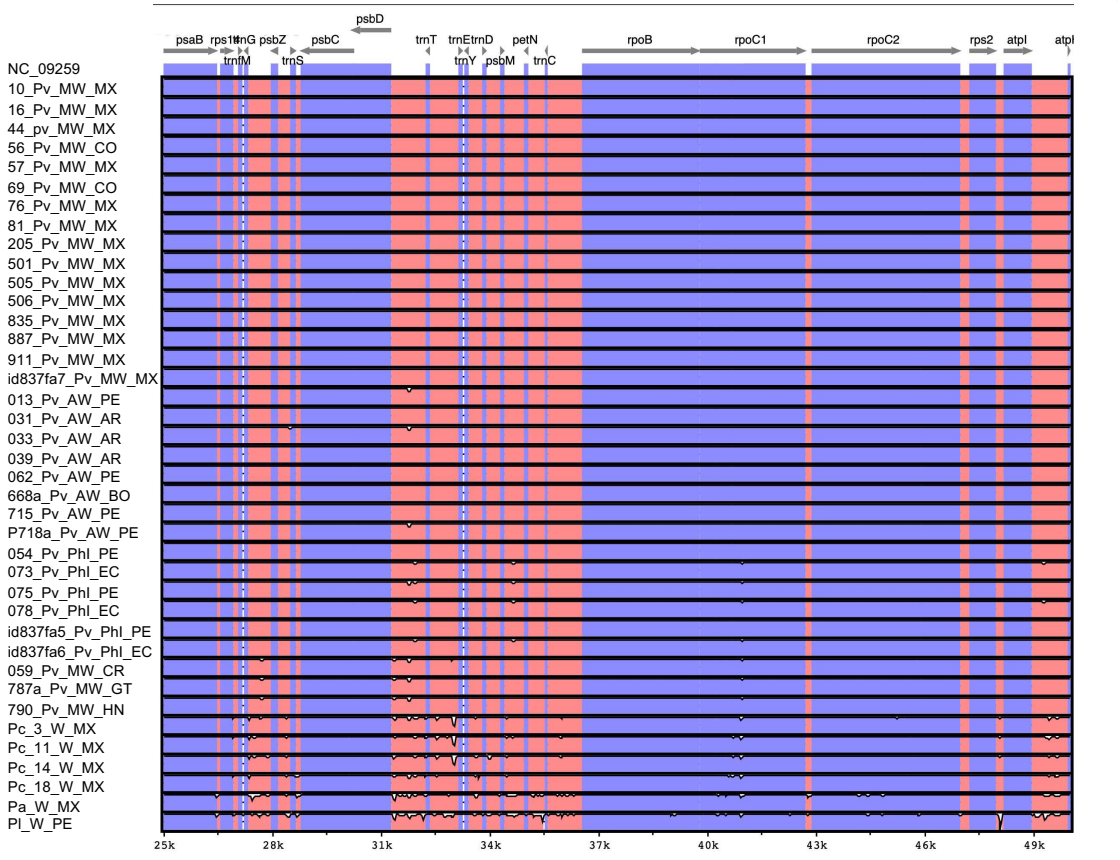
ID	Species	size	LSC	IR	SSC	LSC Start	LSC End	IRb Start	IRb End	SSC Start	SSC End	IRa Start	IRa End	GC %
NC_009259 (Reference chloroplast genome *P. vulgaris*)	*P. vulgaris*	150285	79823	26426	17610	1	79823	79824	106249	106250	123859	123860	150285	
010_Pv_MW_MX	*P. vulgaris*	150249	79782	26437	17593	1	79782	79783	106219	106220	123812	123813	150249	35,45
016_Pv_MW_MX	*P. vulgaris*	150285	79842	26425	17593	1	79842	79843	106267	106268	123860	123861	150285	35,44
031_Pv_AW_AR	*P. vulgaris*	150277	79816	26426	17609	1	79816	79817	106242	106243	123851	123852	150277	35,44
044_Pv_MW_MX	*P. vulgaris*	150281	79809	26427	17618	1	79809	79810	106236	106237	123854	123855	150281	35,43
056_Pv_MW_CO	*P. vulgaris*	150259	79814	26416	17613	1	79814	79815	106230	106231	123843	123844	150259	35,44
057_Pv_MW_MX	*P. vulgaris*	150286	79823	26435	17593	1	79823	79824	106258	106259	123851	123852	150286	35,44
059_Pv_MW_CR	*P. vulgaris*	150217	79810	26412	17583	1	79810	79811	106222	106223	123805	123806	150217	35,45
069_Pv_MW_CO	*P. vulgaris*	150382	79842	26469	17602	1	79842	79843	106311	106312	123913	123914	150382	35,46
076_Pv_MW_MX	*P. vulgaris*	150282	79815	26437	17593	1	79815	79816	106252	106253	123845	123846	150282	35,44
081_Pv_MW_MX	*P. vulgaris*	150258	79816	26394	17654	1	79816	79817	106210	106211	123864	123865	150258	35,44
205_Pv_MW_MX	*P. vulgaris*	151410	80963	26417	17613	1	80963	80964	107380	107381	124993	124994	151410	35,4
501_Pv_MW_MX	*P. vulgaris*	151491	80968	26417	17689	1	80968	80969	107385	107386	125074	125075	151491	35,38
505_Pv_MW_MX	*P. vulgaris*	150265	79799	26428	17610	1	79799	79800	106227	106228	123838	123839	150266	35,44
506_Pv_MW_MX	*P. vulgaris*	150244	79810	26425	17584	1	79810	79811	106235	106236	123819	123820	150244	35,45
787a_Pv_MW_GT	*P. vulgaris*	150199	79811	26401	17586	1	79811	79812	106212	106213	123798	123799	150199	35,45
790_Pv_MW_HN	*P. vulgaris*	150198	79810	26401	17586	1	79810	79811	106211	106212	123797	123798	150198	35,45
835_Pv_MW_MX	*P. vulgaris*	150270	79808	26433	17596	1	79808	79809	106241	106242	123837	123838	150270	35,43
887_Pv_MW_MX	*P. vulgaris*	150282	79813	26436	17597	1	79813	79814	106249	106250	123846	123847	150282	35,43
911_Pv_MW_MX	*P. vulgaris*	150313	79806	26417	17673	1	79806	79807	106223	106224	123896	123897	150313	35,44
id837fa7_Pv_MW_MX	*P. vulgaris*	150274	79830	26416	17612	1	79830	79831	106246	106247	123858	123859	150274	35,44
013_Pv_AW_PE	*P. vulgaris*	150256	79819	26417	17603	1	79819	79820	106236	106237	123839	123840	150256	35,44
033_Pv_AW_AR	*P. vulgaris*	150263	79819	26416	17612	1	79819	79820	106235	106236	123847	123848	150263	35,44
039_Pv_AW_AR	*P. vulgaris*	150259	79815	26416	17612	1	79815	79816	106231	106232	123843	123844	150259	35,44
062_Pv_AW_PE	*P. vulgaris*	150912	79816	26418	18260	1	79816	79817	106234	106235	124494	124495	150912	35,39
668a_Pv_AW_BO	*P. vulgaris*	150260	79813	26417	17613	1	79813	79814	106230	106231	123843	123844	150260	35,45
715_Pv_AW_PE	*P. vulgaris*	151333	80885	26418	17612	1	80885	80886	107303	107304	124915	124916	151333	35,42
P718a_Pv_AW_PE	*P. vulgaris*	150266	79820	26417	17612	1	79820	79821	106237	106238	123849	123850	150266	35,44
054_Pv_PhI_PE	*P. vulgaris*	150306	79836	26428	17614	1	79836	79837	106264	106265	123878	123879	150306	35,44
073_Pv_PhI_EC	*P. vulgaris*	150236	79754	26432	17618	1	79754	79755	106186	106187	123804	123805	150236	35,44
075_Pv_PhI_PE	*P. vulgaris*	150384	79907	26397	17683	1	79907	79908	106304	106305	123987	123988	150384	35,42
078_Pv_PhI_EC	*P. vulgaris*	150235	79751	26433	17618	1	79751	79752	106184	106185	123802	123803	150235	35,44
id837fa5_Pv_PhI_PE	*P. vulgaris*	151680	81178	26418	17666	1	81178	81179	107596	107597	125262	125263	151680	35,38
id837fa6_Pv_PhI_EC	*P. vulgaris*	150239	79825	26406	17602	1	79825	79826	106231	106232	123833	123834	150239	35,44
PI_W_PE	*P. lunatus*	152292	81619	26539	17595	1	81619	81620	108158	108159	125753	125754	152292	35,43
Pc_11_W_MX	*P. coccineus*	150383	79938	26409	17627	1	79938	79939	106347	106348	123974	123975	150383	35,4
Pc_14_W_MX	*P. coccineus*	150287	79836	26387	17677	1	79836	79837	106223	106224	123900	123901	150287	35,4
Pc_18_W_MX	*P. coccineus*	150305	79829	26424	17628	1	79829	79830	106253	106254	123881	123882	150305	35,44
Pc_3_W_MX	*P. coccineus*	150452	80044	26390	17628	1	80044	80045	106434	106435	124062	124063	150452	35,39
Pa_W_MX	*P. acutifolius*	150133	79669	26410	17644	1	79669	79670	106079	106080	123743	123744	150153	35,47

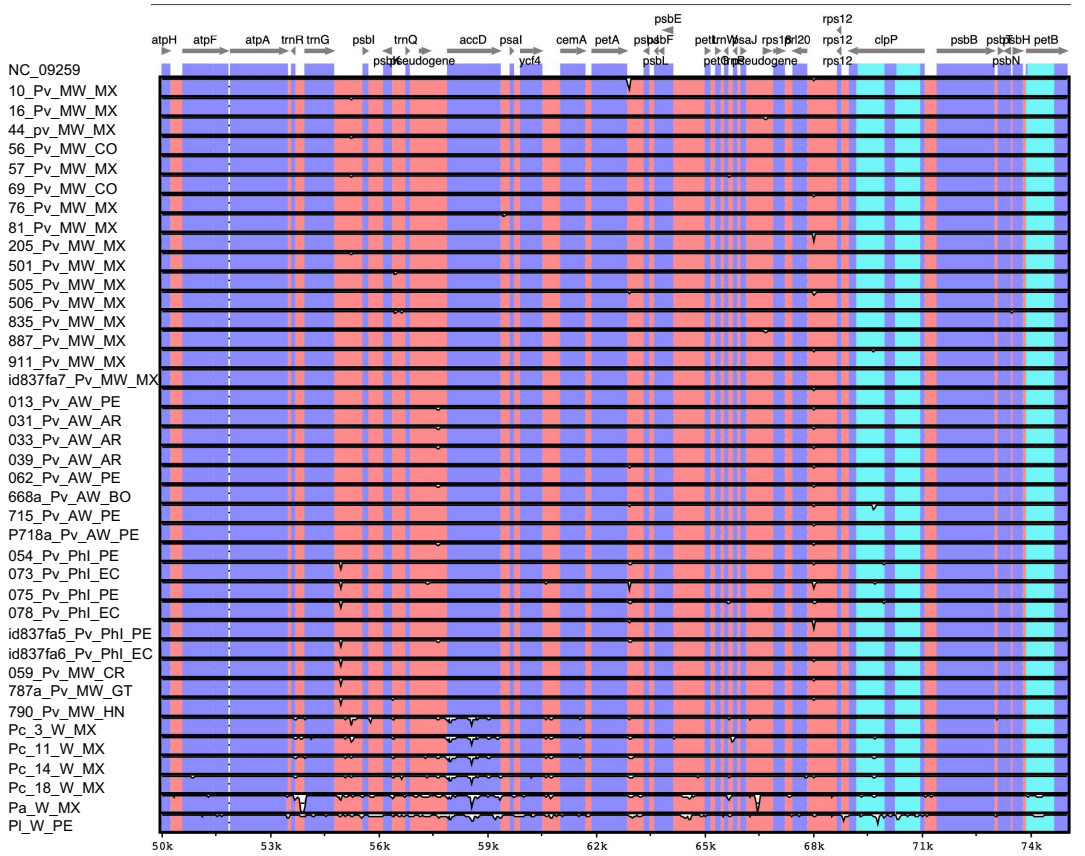
Comparative genome analysis

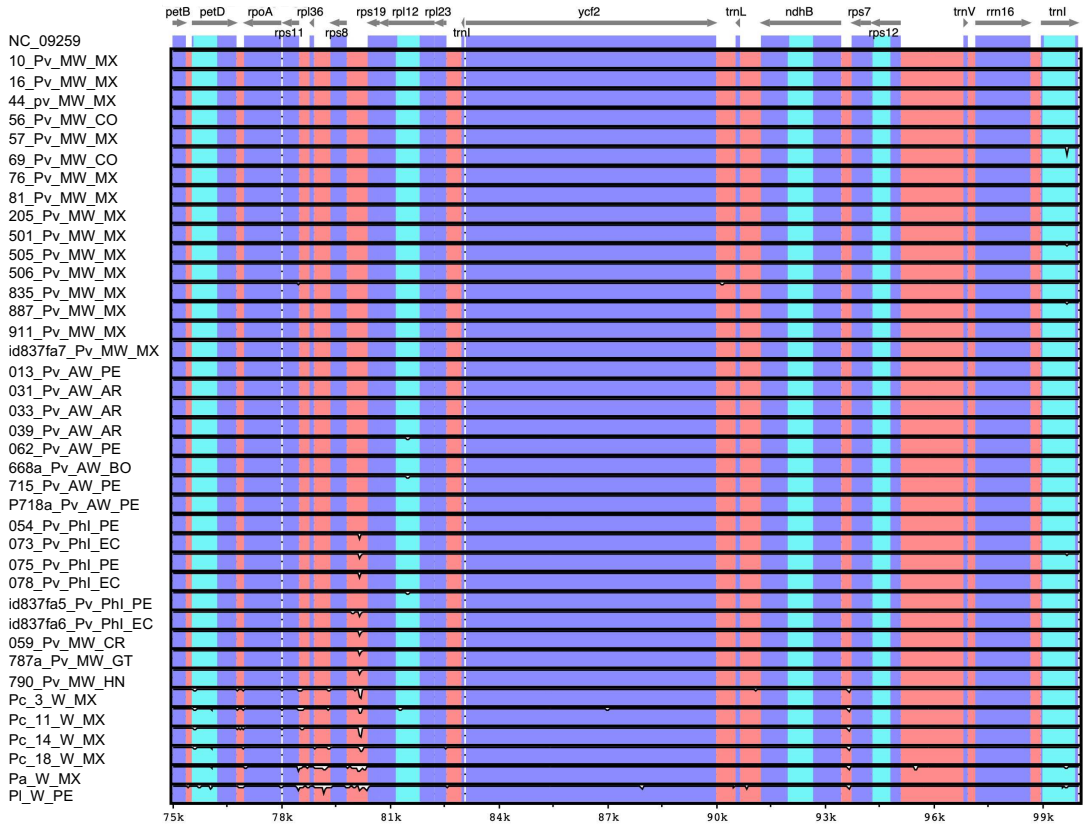
The genome variation among all *de-novo* assembled chloroplast genomes was analyzed with the online tool mVISTA unsing *P. vulgaris* (NC_09259) as reference. The mVISTA identity plot did not reveal meaningful differences between the *P. vulgaris* plastomes, with the exception of a small deletion in the intergenic region between *trnK* and *rbcL* gene, which was found in two PhI samples (i.e., 073_Pv_PhI_EC and 078_Pv_PhI_EC). In addition, a deletion (over 300 pb) was identified in the plastome of the *P. acutifolius* accession (i.e., Pa_W_MX) in the intergenic region between the *trnR* and *trnG* genes (Figure 2.2).

To identify the gene order and organization, a subset of 15 plastomes was aligned with MAUVE (Figure 2.3). Most of the regions were conserved among all plastomes and no rearrangements of gene order was detected.


matK
atpE
psaB
rps3
rpl16
rpl14
psbA
trnK
rbcL
atpB
trnM
ndhC
ndhJ
trnF
trnL
trnT
trnS
ycf3
psaA
trnH
trnV
ndhK
rps4
NC_09259
10_Pv_MW_MX
16_Pv_MW_MX
44_pv_MW_MX
56_Pv_MW_CO
57_Pv_MW_MX
69_Pv_MW_CO
76_Pv_MW_MX
81_Pv_MW_MX
205_Pv_MW_MX
501_Pv_MW_MX
505_Pv_MW_MX
506_Pv_MW_MX
835_Pv_MW_MX
887_Pv_MW_MX
911_Pv_MW_MX
id837fa7_Pv_MW_MX
013_Pv_AW_PE
031_Pv_AW_AR
033_Pv_AW_AR
039_Pv_AW_AR
062_Pv_AW_PE
668a_Pv_AW_BO
715_Pv_AW_PE
P718a_Pv_AW_PE
054_Pv_PhI_PE
073_Pv_PhI_EC
075_Pv_PhI_PE
078_Pv_PhI_EC
id837fa5_Pv_PhI_PE
id837fa6_Pv_PhI_EC
059_Pv_MW_CR
787a_Pv_MW_GT
790_Pv_MW_HN
Pc_3_W_MX
Pc_11_W_MX
Pc_14_W_MX
Pc_18_W_MX
Pa_W_MX
PI_W_PE
100%
50%
0k
3k
6k
9k
12k
15k
18k
21k
24k


psbD
psaB
rps14
trnG
psbZ
psbC
trnT
trnE
trnD
petN
rpoB
rpoC1
rpoC2
rps2
atpI
trnM
trnS
trnY
psbM
trnC
NC_09259
10_Pv_MW_MX
16_Pv_MW_MX
44_pv_MW_MX
56_Pv_MW_CO
57_Pv_MW_MX
69_Pv_MW_CO
76_Pv_MW_MX
81_Pv_MW_MX
205_Pv_MW_MX
501_Pv_MW_MX
505_Pv_MW_MX
506_Pv_MW_MX
835_Pv_MW_MX
887_Pv_MW_MX
911_Pv_MW_MX
id837fa7_Pv_MW_MX
013_Pv_AW_PE
031_Pv_AW_AR
033_Pv_AW_AR
039_Pv_AW_AR
062_Pv_AW_PE
668a_Pv_AW_BO
715_Pv_AW_PE
P718a_Pv_AW_PE
054_Pv_PhI_PE
073_Pv_PhI_EC
075_Pv_PhI_PE
078_Pv_PhI_EC
id837fa5_Pv_PhI_PE
id837fa6_Pv_PhI_EC
059_Pv_MW_CR
787a_Pv_MW_GT
790_Pv_MW_HN
Pc_3_W_MX
Pc_11_W_MX
Pc_14_W_MX
Pc_18_W_MX
Pa_W_MX
PI_W_PE
100%
50%
25k
28k
31k
34k
37k
40k
43k
46k
49k


atpH atpF atpA trnR trnG psbI trnQ accD psaI cemA petA psbE psbF psbL rps12 clpP psbB psbT psbH petB
NC_09259
10_Pv_MW_MX
16_Pv_MW_MX
44_pv_MW_MX
56_Pv_MW_CO
57_Pv_MW_MX
69_Pv_MW_CO
76_Pv_MW_MX
81_Pv_MW_MX
205_Pv_MW_MX
501_Pv_MW_MX
505_Pv_MW_MX
506_Pv_MW_MX
835_Pv_MW_MX
887_Pv_MW_MX
911_Pv_MW_MX
id837fa7_Pv_MW_MX
013_Pv_AW_PE
031_Pv_AW_AR
033_Pv_AW_AR
039_Pv_AW_AR
062_Pv_AW_PE
668a_Pv_AW_BO
715_Pv_AW_PE
P718a_Pv_AW_PE
054_Pv_PhI_PE
073_Pv_PhI_EC
075_Pv_PhI_PE
078_Pv_PhI_EC
id837fa5_Pv_PhI_PE
id837fa6_Pv_PhI_EC
059_Pv_MW_CR
787a_Pv_MW_GT
790_Pv_MW_HN
Pc_3_W_MX
Pc_11_W_MX
Pc_14_W_MX
Pc_18_W_MX
Pa_W_MX
PI_W_PE
50k
53k
56k
59k
62k
65k
68k
71k
74k


petB petD rpoA rpl36 rps19 rpl12 rpl23 ycf2 trnL ndhB rps7 trnV rrn16 trnI
rps11 rps8 trnI rps12
NC_09259
10_Pv_MW_MX
16_Pv_MW_MX
44_pv_MW_MX
56_Pv_MW_CO
57_Pv_MW_MX
69_Pv_MW_CO
76_Pv_MW_MX
81_Pv_MW_MX
205_Pv_MW_MX
501_Pv_MW_MX
505_Pv_MW_MX
506_Pv_MW_MX
835_Pv_MW_MX
887_Pv_MW_MX
911_Pv_MW_MX
id837fa7_Pv_MW_MX
013_Pv_AW_PE
031_Pv_AW_AR
033_Pv_AW_AR
039_Pv_AW_AR
062_Pv_AW_PE
668a_Pv_AW_BO
715_Pv_AW_PE
P718a_Pv_AW_PE
054_Pv_PhI_PE
073_Pv_PhI_EC
075_Pv_PhI_PE
078_Pv_PhI_EC
id837fa5_Pv_PhI_PE
id837fa6_Pv_PhI_EC
059_Pv_MW_CR
787a_Pv_MW_GT
790_Pv_MW_HN
Pc_3_W_MX
Pc_11_W_MX
Pc_14_W_MX
Pc_18_W_MX
Pa_W_MX
PI_W_PE
75k
78k
81k
84k
87k
90k
93k
96k
99k

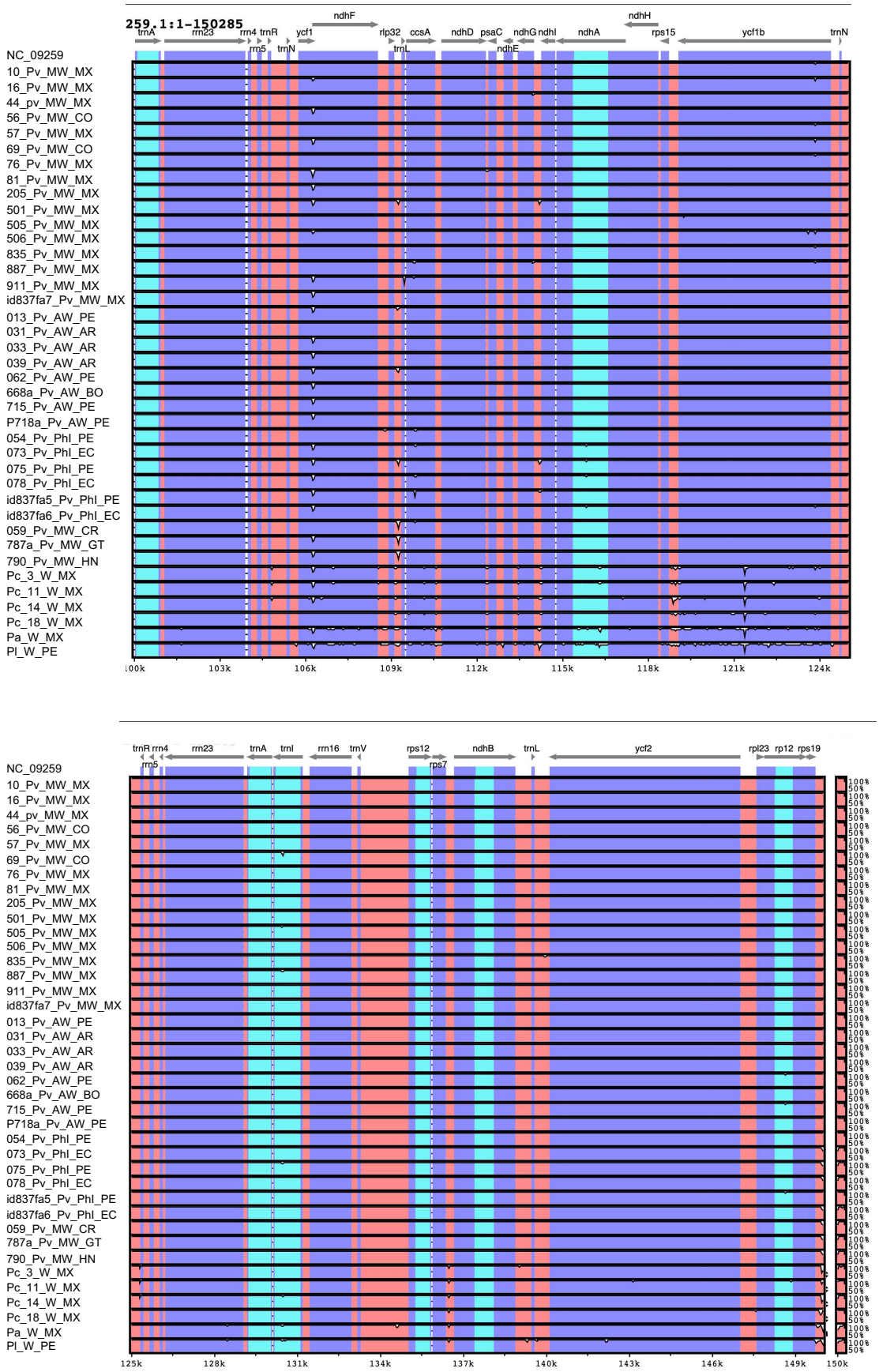


Figure 2.2 Sequence similarity plot by using mVISTA, among 39 de-novo assembled chloroplast genomes and NC_09259 as reference. In the y-axis percentage of sequence identity was shown between 50% and

100%. Transcriptional orientations of genes were assigned by grey arrows. Red bars represented non-coding sequences (NCS), purples bars represented exons and light blue bars represented untranslated regions (UTRs). Genomic differences were shown as white peaks.

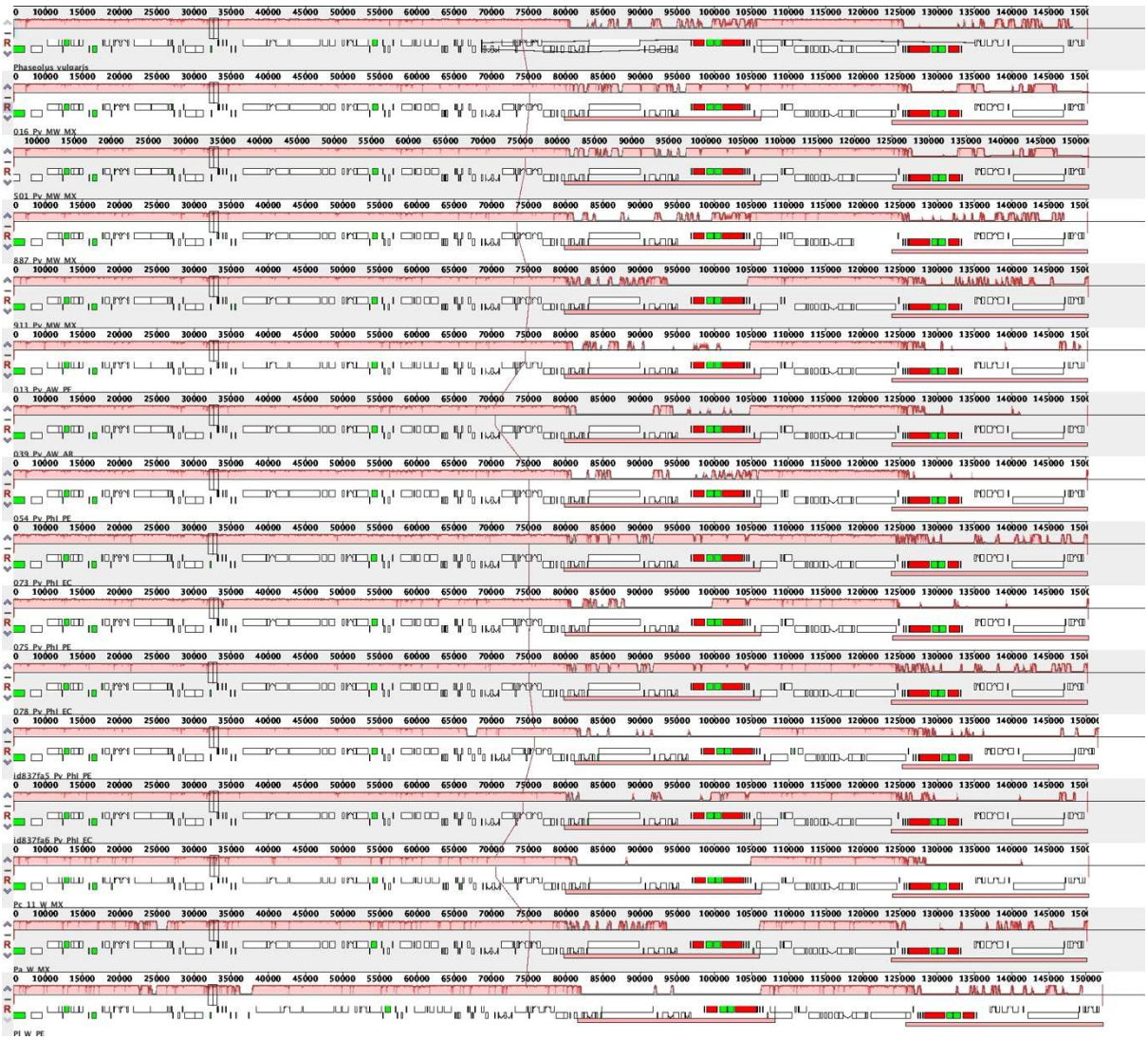


Figure 2.3 MAUVE alignment showing the gene order and homology between fifteen accessions P. vulgaris (16_Pv_MW_MX, 501_Pv_MW_MX, 887_Pv_MW_MX, 911_Pv_M_MX, 013_Pv_AW_PE0, 39_Pv_AW_AR,54_Pv_PhI_PE, 073_Pv_PhI_EC, 075_Pv_PhI_PE, 078_Pv_PhI_EC), P. coccineus (Pc_11_W_MX), P. acutifolius (Pa_W_MX), P. lunatus (Pl_W_PE) , and NC_09259 as reference. Locally Collinear Blocks (LCBs) include the histograms that show sequence identity with peaks. Protein coding genes, rRNA genes, tRNA genes and intron containing tRNA genes are marked with block in white, red, black and green colors, respectively.

Analysis of nucleotide diversity

The nucleotide diversity was investigated with DNAsp software. The multi sequence alignment revealed 145133 monomorphic sites 3586 polymorphic sites of which 1254 were defined as informative. A sliding window analysis was performed to calculate the nucleotide variability (Pi) across the 39 plastomes. Even though, the results showed a high sequence similarity, three divergent hot spots were detected (Pi>0.02): trnY-GUA, petA, rrn23 (Figure 2.4)

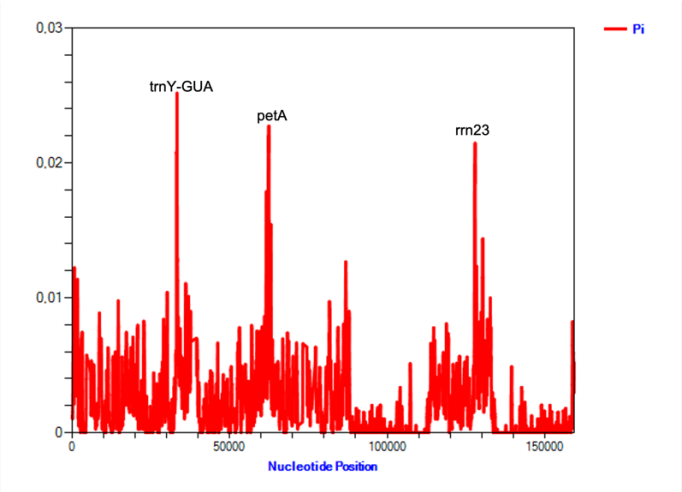


Figure 2.4 Sliding window analysis among the whole chloroplast genome of 39 de-novo assembled plastomes, including P. vulgaris, P. coccineus, P. acutifolius, P. lunatus. Regions with higher nucleotide variability are indicated.

Pan-plastome of *P. vulgaris*

The thirty-three chloroplast genomes, reconstructed from the sequences of wild *P. vulgaris* accessions, were used to build a consensus pan-plastome of 156,269 bp (Figure 2.5). The alignment of the full length plastomes revealed 6473 sites identified as missing data or gaps, of the remaining loci, 146207 are monomorphic, 236 singletons and 353 can be considered parsimony informative sites. The nucleotide diversity described by the Pi value is in a range between 0 and 0.012. A total of 4880 bp were recorded to be InDels forming 181 InDel events. The InDels were neutral, as Tajma D was not significant.

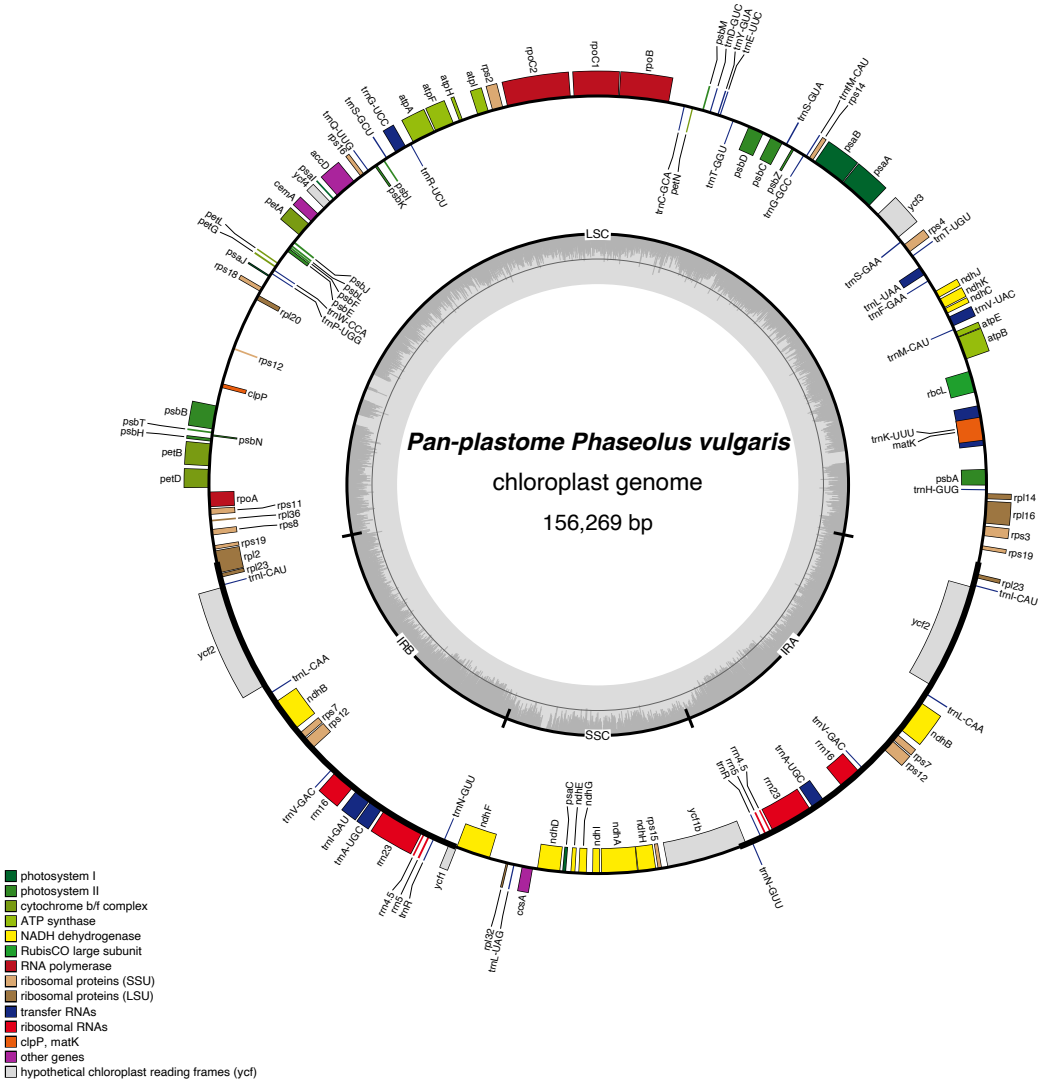


Figure 2.5 Map of P. vulgaris pan-plastome. Genes inside of the outer circle are transcribed in the clockwise direction, while those outsides are transcribed in the counterclockwise direction. Different color codes represent genes belonging to various functional groups. The circle inside represents the GC content graph with the 50% threshold. The inverted repeats, large single-copy, and small single-copy regions are denoted by IR, LSC, and SSC, respectively.

Tandem repeats were identified for the pan-plastome. A total of 17 tandem repeats were detected, 9 of which were polymorphic. The TRs were characterized by a period size ranging between 9 bp and 21 bp. All the tandem repeats placed on the LSC were intergenic (7), the remaining 10 TRs were in exon of ycf1, ycf2, ndhF, and rps19 (Table 2.3).

Table 2.3 Summary of the TRs identified in the pan-plastome.

Gene	Region	Period Sequence	Period size	Copy number	Consensus size	Percent matches
rpl14-trnH-GUG	intergenic	AGTATTTCTTACTTTTA	16	2.4	17	82
psbA-trnK-UUU	intergenic	TAGTATGTTCTTATTCAT	18	2.1	18	100
trnK-UU-rbcL	intergenic	ATTGAATATAGAAT	14	2.0	14	100
atpA-trnG-UCC	intergenic	AACCTTATTAACTA	14	2.1	14	93
trnG-UCC-trnSGCU	intergenic	ATATACAATTTAC	13	1.9	13	100
trnP-UGG-psaJ	intergenic	TCATAGAATACGGAATA	17	2.3	17	100
rps8-rps19	intergenic	CTTCTGTATAGAGGTT	16	3.3	16	94
ycf2	exon	TTTTTGTCCAAGTTACTTCTC	21	2.4	21	86
ycf2	exon	TATTGATGATAGTGACGA	18	2.1	18	100
ycf2	exon	GATAGTGAC	9	4.4	9	87
ycf1	exon	AAAATTAATAT	10	3.1	11	87
ndhF	exon	AAAATTTTTTTCAAAT	16	2.2	16	89
ycf1b	exon	TTTTCAGTA	9	3.1	9	100
ycf2	exon	TATCGTCAC	9	4.4	9	87
ycf2	exon	TATCGTCACTATCATCAA	18	2.1	18	100
ycf2	exon	GACAAAAAAAGAAGAAACTTG	21	2.4	21	86
rps19	exon	AAGAACCTCTATACAG	16	3.2	16	97

Discussion

The peculiar characteristics of plastome such as the haploid genome, the uniparental inheritance and the lack of recombination make the genome of this circular organelle particularly suitable for genetic investigation and especially phylogenetic analysis. Indeed, cpDNA of various crops was used to investigate and solve relationship among species such as *Phaseolus spp*, *Cucurbitaceae spp*, *Brassica spp*, *Capsicum spp* and *Macadamia* spp (Salinas, 1993; Kocyan et al., 2007; Arias & Chris Pires, 2012; D'Agostino et al., 2018; Nock et al., 2019 and She et al., 2022). Up to date, only two full length chloroplast genomes of *P. vulgaris* and one cpDNA of *P. lunatus* were published (Guo et al. 2007; Meng and Li 2018). With the present study, we contributed to improve the cpDNA sequence space available for *Phaseolus spp.* by the assembly of 39 chloroplast genomes.

Comparative analyses of the thirty-nine *de-novo* assembled chloroplasts of *Phaseolus spp.* revealed a high conserved structure of the plastome and gene order. Conversely to the wild *Cicer echinospermum* in which two inversions were found when compared to the *Cicer arietinum* (Mehmetoglu et al. 2022), no structural rearrangements were detected among the four species (*P. vulgaris*, *P. coccineus*, *P. acutifolius*, *P. lunatus*) when compared to each other and to the domesticated reference chloroplast genome (*P. vulgaris* cv Negro Jamapa). A small deletion, unique of two common bean accessions belonging to the North Peru-Ecuador gene pool (i.e., 073_Pv_PhI_EC and 078_Pv_PhI_EC) was identified in the intergenic region between *trnK* and *rbcL* genes. In addition, a deletion of over 300 bp was found to be peculiar of *P. acutifolius* plastome.

Compared to the previously published *P. vulgaris* (Guo et al. 2007) and *P. lunatus* (Tian et al. 2021) chloroplast genomes, the *de-novo* assembled plastomes showed similar dimension, gene content and GC percentage. The first automatic annotation was not able to detect one gene (trnV) and 2 pseudogenes (rps16 and rpl133). Since the pseudo gene rps16 was reported to be a distinctive characteristic of the *P. vulgaris* chloroplast genome in the work of Guo et al. (2007), an additional BLAST search was performed for the one gene and two pseudo genes confirming their presence.

The nucleotide diversity of the plastomes was investigated to understand the level of polymorphism between the *Phaseolus* accessions. As expected, due to the high number of samples from the same species (*P. vulgaris*) and the high conservation of chloroplast genomes, the nucleotide diversity was low with a range between 0 and 0.025 but comparable with other studies conducted on samples of the same species (Song et al., 2017). However, three hot spot

regions showed a Pi value grater that 0.02. Two are located in the LSC region (trnY-GUA and PetA) and one was found in the IR region (rrn23). The intragenic region of PetA was already reported to be one of the most variable loci in the work of Dong et al. (2012) where chloroplast genome belonging to 12 different genera were scanned. Regions such as ycf1, ycf2, matk, trnL, rbcl, trnK, rpl32-trnL, and trnH-psbA found to be highly variable in previous studies (Dong et al., 2012; Guo et al., 2007; Song et al., 2017), did not showed great diversity in the analyzed set of samples confirming the high plastome conservation among *Phaseolus* species.

Recently, pan-plastomes have been proposed for many species such as *Capsicum* spp (Magdy et al. 2019), the Asian lotus (Wang et al. 2022) and the sugar beet (Sielemann et al. 2022). With the aim to collect the chloroplast diversity of wild common bean, a consensus pan-plastome was built with 33 full length plastomes of *P. vulgaris,* following the work of Magdy et al. (2019). Despite the low nucleotide diversity found in the consensus sequence, seventeen tandem repeats were detected. The LSC region was characterized by intergenic TRs, while ycf1, ycf2, nadhF and rps19 showed TRs in the exons. Two tandem repeats of the gene ycf2 were already reported in common bean plastome (Guo et al. 2007). Repetitive regions in the chloroplast genome play a key role in the analysis of evolutionary scenarios, moreover they can be informative in phylogenetic studies (Cavalier-Smith, 2002). Due to the high degree of polymorphisms that characterizes TRs, they can be used as molecular markers such as the case of the variable tandem repeat found in the gene rpl23-trnI that can discriminate different *Capsicum* species (Magdy et al. 2019).

Conclusion

In this work we provide 39 plastome assemblies for four different *Phaseolus* species (*P. vulgaris*, *P. coccineus*, *P. acutifolius*, *P. lunatus*), with particular attention for common bean. Indeed, trying to cover most of the *P. vulgaris* species' diversity, we *de-novo* assembled 33 chloroplast genomes of wild accessions belonging to the three main gene-pools: Mesoamerican, Andean and the one from North Peru-Ecuador.

Comparative analysis revealed a high level of conservation of *Phaseolus spp.* plastomes. Nevertheless, we found small deletions in common bean samples from North Peru-Ecuador and one bigger deletion characteristics of *P. acutifolius*. Three hot-spots of diversity were detected close to trnY-GUA, PetA and rrn23 genes.

The chloroplast diversity of *P. vulgaris* was collected in a consensus pan-plastome for which nucleotide diversity, tandem repeats and InDels were investigated. Since the use of chloroplast genomes for reconstructing the species phylogeny is preferable compared to the analysis of single genes, plastomes analyzed in the present study represent a resource for future investigations of *Phaseolus* phylogeny. Furthermore, the great number of chloroplast sequences of wild common bean accessions allows a wide exploration of the evolutionary history and origin of common bean, as will be reported in the next chapter.

References

Andrews, S., J. Gilley, and M. P. Coleman. 2010. “Difference Tracker: ImageJ Plugins for Fully Automated Analysis of Multiple Axonal Transport Parameters.” *Journal of Neuroscience Methods* 193(2):281–87. doi: 10.1016/j.jneumeth.2010.09.007.

Arias, T., and J. C. Pires. 2012. “A Fully Resolved Chloroplast Phylogeny of the Brassica Crops and Wild Relatives (Brassicaceae: Brassiceae): Novel Clades and Potential Taxonomic Implications.” *Taxon* 61(5):980–88. doi: 10.1002/tax.615005.

Bausher, M. G., N.D. Singh, S. Bum Lee, R. K. Jansen, and H. Daniell. 2006. “The Complete Chloroplast Genome Sequence of Citrus Sinensis (L.) Osbeck Var ‘Ridge Pineapple’: Organization and Phylogenetic Relationships to Other Angiosperms.” *BMC Plant Biology* 6. doi: 10.1186/1471-2229-6-21.

Bendich, A. J. 2004. “Circular Chloroplast Chromosomes: The Grand Illusion.” *The Plant Cell* 16(7):1661–66. doi: https://doi.org/10.1007/BF02907356.

Benson, G. 1999. “Tandem repeats finder: a program to analyze DNA sequences”. Nucleotide acids research. Vol 27(2), 573-580

Bitocchi, E., D. Rau, E. Bellucci, M. Rodriguez, M. L. Murgia, T. Gioia, D. Santo, Laura Nanni, G. Attene, and R. Papa. 2017. “Beans (Phaseolus Ssp.) as a Model for Understanding Crop Evolution.” *Frontiers in Plant Science* 8.

Bock, R. 2007. *Structure, Function, and Inheritance of Plastid Genomes. Bock R. (eds) Cell and Molecular Biology of Plastids. Topics in Current Genetics, vol 19. Springer, Berlin, Heidelberg. https://doi.org/10.1007/4735_2007_0223*

Bolger, A. M., M. Lohse, and B. Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30(15):2114–20. doi: 10.1093/bioinformatics/btu170.

Cavalier-Smith T. 2002. “Chloroplast Evolution: Secondary Symbiogenesis and Multiple Losses Dispatch.” *Current Biology* 12(2):R62–64. doi: https://doi.org/10.1007/BF02907356.

Cortinovis, G., G. Frascarelli, V. di Vittori, and R. Papa. 2020. “Current State and Perspectives in Population Genomics of the Common Bean.” *Plants* 9(3). Doi: 10.3390/plants9030330

D’Agostino, N., R. Tamburino, C. Cantarella, V. de Carluccio, L. Sannino, S. Cozzolino, T. Cardi, and N. Scotti. 2018. “The Complete Plastome Sequences of Eleven Capsicum Genotypes: Insights into DNA Variation and Molecular Evolution.” *Genes* 9(10). doi: 10.3390/genes9100503.

Darling, A. C. E., B. Mau, F. R. Blattner, and N. T. Perna. 2004. “Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements.” *Genome Research* 14(7):1394–1403. doi: 10.1101/gr.2289704.

Debouck, D. G., O. Toro, O. M. Paredes, W. C. Johnson, and P. Gepts. 1993. "Genetic Diversity and Ecological Distribution of Phaseolus Vulgaris (Fabaceae) in Northwestern South America." *Economic Botany* 47(4):408–23. doi: https://doi.org/10.1007/BF02907356.

Dierckxsens, N., P. Mardulyn, and G. Smits. 2017. "NOVOPlasty: De Novo Assembly of Organelle Genomes from Whole Genome Data." *Nucleic Acids Research* 45(4). doi: 10.1093/nar/gkw955.

Dong, W., J. Liu, Jing Yu, L. Wang, and S. Zhou. 2012. "Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding." *PloS One* 7(4). doi: 10.1371/journal.pone.0035071.

Gao, L., I. Gonda, H. Sun, Q. Ma, K. Bao, D. M. Tieman, E. A. Burzynski-Chang, T. L. Fish, K. A. Stromberg, G. L. Sacks, T. W. Thannhauser, M. R. Foolad, M. Jose Diez, J. Blanca, J. Canizares, Y. Xu, E.Knaap, S. Huang, H. J. Klee, J. J. Giovannoni, and Z. Fei. 2018. "The Tomato Pan-Genome Uncovers New Genes and a Rare Allele Regulating Fruit Flavor." doi: 10.1038/s41588-019-0410-2.

Gao, L., I. Gonda, H. Sun, Q. Ma, K. Bao, D. M. Tieman, E. A. Burzynski-Chang, T. L. Fish, K. A. Stromberg, G. L. Sacks, T. W. Thannhauser, M. R. Foolad, M. Jose Diez, J. Blanca, J. Canizares, Y. Xu, E. Knaap, S. Huang, H. J. Klee, J. J. Giovannoni, and Z.Fei. 2022. "Graph-Based Pan-Genome Reveals Structural and Sequence Variations Related to Agronomic Traits and Domestication in Cucumber." doi: 10.1038/s41467-022-28362-0.

Greiner, S., P. Lehwark, and R. Bock. 2019. "OrganellarGenomeDRAW (OGDRAW) Version 1.3.1: Expanded Toolkit for the Graphical Visualization of Organellar Genomes." *Nucleic Acids Research* 47(W1):W59–64. doi: 10.1093/nar/gkz238.

Guo, X. , S. Castillo-Ramírez, V. González, P. Bustos, J. L.Fernández-Vázquez, R. Santamaría, J. Arellano, M. A. Cevallos, and G. Dávila. 2007. "Rapid Evolutionary Change of Common Bean (Phaseolus Vulgaris L) Plastome, and the Genomic Diversification of Legume Chloroplasts." *BMC Genomics* 8. doi: 10.1186/1471-2164-8-228.

Jayakodi, M., S. Padmarasu, G. Haberer, V. Suresh Bonthala, H. Gundlach, C. Monat, T. Lux, N. Kamal, D. Lang, A. Himmelbach, J. Ens, X. Q. Zhang, T. T. Angessa, G. Zhou, C. Tan, C. Hill, P. Wang, M. Schreiber, L. B. Boston, C. Plott, J. Jenkins, Y. Guo, A. Fiebig, H. Budak, D. Xu, J. Zhang, C. Wang, J. Grimwood, J. Schmutz, G. Guo, G. Zhang, K. Mochida, T. Hirayama, K. Sato, K. J. Chalmers, P. Langridge, R. Waugh, C. J. Pozniak, U. Scholz, K. F. X Mayer, M. Spannagl, C. Li, M. Mascher, and N. Stein. 2020. "The Barley Pan-Genome Reveals the Hidden Legacy of Mutation Breeding." *284 | Nature |* 588. doi: 10.1038/s41586-020-2947-8.

Kami, J., V. Becerra Velasquez, D. G. Debouck, P. Gepts, and R. W. Allard. 1995. *Identification of Presumed Ancestral DNA Sequences of Phaseolin in Phaseolus Vulgaris (Molecular Evolution/Seed Protein/Crop Evolution/Tandem Repeat/Polymerase Chain Reaction) Communicated By*. Vol. 92. doi: https://doi.org/10.1073/pnas.92.4.1101.

Katoh, K., J. Rozewicki, and K. D. Yamada. 2018. “MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization.” *Briefings in Bioinformatics* 20(4):1160–66. doi: 10.1093/bib/bbx108.

Kocyan, A., L.B. Zhang, H. Schaefer, and S. S. Renner. 2007. “A Multi-Locus Chloroplast Phylogeny for the Cucurbitaceae and Its Implications for Character Evolution and Classification.” *Molecular Phylogenetics and Evolution* 44(2):553–77. doi: 10.1016/j.ympev.2006.12.022.

Langmead, B., and S. L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9(4):357–59. doi: 10.1038/nmeth.1923.

Li, J., D. Yuan, P. Wang, Q. Wang, M. Sun, Z. Liu, H. Si, Z. Xu, Y. Ma, B. Zhang, L. Pei, L. Tu, L. Zhu, L. L. Chen, K. Lindsey, X. Zhang, S. Jin, and M. Wang. 2021. “Cotton Pan-Genome Retrieves the Lost Sequences and Genes during Domestication and Selection.” *Genome Biology* 22(1). doi: 10.1186/s13059-021-02351-w.

Magdy, M., L. Ou, H. Yu, R. Chen, Y. Zhou, H. Hassan, B. Feng, N. Taitano, E. van der Knaap, X. Zou, F. Li, and B.Ouyang. 2019. “Pan-Plastome Approach Empowers the Assessment of Genetic Variation in Cultivated Capsicum Species.” *Horticulture Research* 6(1). doi: 10.1038/s41438-019-0191-x.

Margulis, L. 1970. *Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant and Animal Cells on the Precambrian Earth*. Yale University Press.

Mehmetoglu, E., Y. Kaymaz, D. Ates, A. Kahraman, and M. B. Tanyolac. 2022. “The Complete Chloroplast Genome Sequence of Cicer Echinospermum, Genome Organization and Comparison with Related Species.” *Scientia Horticulturae* 296. doi: 10.1016/j.scienta.2022.110912.

Meng, X., and M. Li. 2018. “Characterisation of the Complete Chloroplast Genome of the Common Bean, Phaseolus Vulgaris L.” *Mitochondrial DNA Part B: Resources* 3(2):920–22. doi: 10.1080/23802359.2018.1502634.

Mereschkowsky, C.. 1905. “Uber Natur Und Ursprung Der Chromatophoren Im Pflanzenreiche.” *Biologisches Centralblatt* 203–604.

Mubumbila, M., K. H. J. Gordon, E.n J. Crouse, G. Burkard, and J. H. Weil. 1983. “Construction of the Physical Map of the Chloroplast DNA of Phaseolus Vulgaris and Localization of Ribosomal and Transfer RNA Genes.” *Gene* 21(3):257–66. doi: 10.1016/0378-1119(83)90009-4.

Nock, C. J., C. M. Hardner, J. D. Montenegro, A. A. A. Termizi, S. Hayashi, J. Playford, D. Edwards, and J. Batley. 2019. “Wild Origins of Macadamia Domestication Identified through Intraspecific Chloroplast Genome Sequencing.” *Frontiers in Plant Science* 10. doi: 10.3389/fpls.2019.00334.

Ou, L., D. Li, J. Lv, W. Chen, Z. Zhang, X. Li, B. Yang, S. Zhou, S. Yang, W. Li, H. Gao, Q. Zeng, H. Yu, B. Ouyang, F. Li, F. Liu, J. Zheng, Y. Liu, J. Wang, B. Wang, X. Dai, Y. Ma, and X. Zou. 2018. "Pan-Genome of Cultivated Pepper (Capsicum) and Its Use in Gene Presence-Absence Variation Analyses." *Source: The New Phytologist* 220(2):360–63. doi: 10.2307/90025389.

Qu, X. J., M. J. Moore, D. Z. Li, and T. Shuang Yi. 2019. "PGA: A Software Package for Rapid, Accurate, and Flexible Batch Annotation of Plastomes." *Plant Methods* 15(1). doi: 10.1186/s13007-019-0435-7.

Quinlan, A. R., and I. M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26(6):841–42. doi: 10.1093/bioinformatics/btq033.

Rice, P., I. Longden, and A. Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics* 16(6):276–77. doi: 10.1016/S0168-9525(00)02024-2.

Rozas, J., A. Ferrer-Mata, J. C. Sanchez-DelBarrio, S. Guirao-Rico, P. Librado, S. E. Ramos-Onsins, and A. Sanchez-Gracia. 2017. "DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets." *Molecular Biology and Evolution* 34(12):3299–3302. doi: 10.1093/molbev/msx248.

Salinas, A. Delgado. 1993. "Chloroplast DNA as an Evolutionary Marker in the Phaseolus Vulgaris Complex." *Theoret. Appl. Genetics* 88:646–52. doi: https://doi.org/10.1007/BF01253966.

She, H., Z. Liu, Z. Xu, H. Zhang, F. Cheng, J. Wu, X. Wang, and W. Qian. 2022. "Comparative Chloroplast Genome Analyses of Cultivated Spinach and Two Wild Progenitors Shed Light on the Phylogenetic Relationships and Variation." *Scientific Reports* 12(1):2045–2322. doi: 10.1038/s41598-022-04918-4.

Sielemann, K., B. Pucker, N. Schmidt, P. Viehöver, T. Heitkam, and Da. Holtgräwe. 2022. "Complete Pan-Plastome Sequences Enable High Resolution Phylogenetic Classification of Sugar 2 Beet and Closely Related Crop Wild Relatives." *BioRxiv*. doi: 10.1101/2021.10.08.463637.

Song, J. M., Z. Guan, J. Hu, C. Guo, Z. Yang, S. Wang, D. Liu, B. Wang, S. Lu, R. Zhou, W. Z. Xie, Y. Cheng, Y. Zhang, K. Liu, Q. Y. Yang, L. L. Chen, and L. Guo. 2020. "Eight High-Quality Genomes Reveal Pan-Genome Architecture and Ecotype Differentiation of Brassica Napus." *Nature Plants* 6:34–45. doi: 10.1038/s41477-019-0577-7.

Song, Y., S. Wang, Y. Ding, J. Xu, M. Fu Li, S. Zhu, and N. Chen. 2017. "Chloroplast Genomic Resource of Paris for Species Discrimination." *Scientific Reports* 7(1). doi: 10.1038/s41598-017-02083-7.

Tao, Y., H. Luo, J. Xu, A. Cruickshank, X. Zhao, F. Teng, A. Hathorn, X. Wu, Y. Liu, T. Shatte, D. Jordan, H. Jing, and E. Mace. 2021. "Extensive Variation within the Pan-Genome of Cultivated and Wild Sorghum." *Nature Plants* 7:766–73. doi: 10.1038/s41477-021-00925-x.

Tian, S., P. Lu, Z. Zhang, J. Qiang Wu, H. Zhang, and H. Shen. 2021. “Chloroplast Genome Sequence of Chongming Lima Bean (Phaseolus Lunatus L.) and Comparative Analyses with Other Legume Chloroplast Genomes.” *BMC Genomics* 22(1). doi: 10.1186/s12864-021-07467-8.

Torkamaneh, D., M. A. Lemay, and F. ßois Belzile. 2021. “The Pan-Genome of the Cultivated Soybean (PanSoy) Reveals an Extraordinarily Conserved Gene Content.” *Plant Biotechnology Journal* 19:1852–62. doi: 10.1111/pbi.13600.

Toro Chica, O., J. M. T., and D. G. Debouck. 1990. *Wild Bean (Phaseolus Vulgaris L.): Description and Distribution*. Vol. 181. CIAT.

Varshney, . K., M. Roorkiwal, S. Sun, P. Bajaj, A. Chitikineni, M. Thudi, N. P. Singh, X. Du, H. D. Upadhyaya, A. W. Khan, Y. Wang, V. Garg, G. Fan, W. A. Cowling, J. Crossa, L. Gentzbittel, K. P. Voss-Fels, V. Kumar Valluri, P. Sinha, V. K. Singh, C. Ben, A. Rathore, R. Punna, M. K. Singh, B. Tar’an 19, C. Bharadwaj, M. Yasin, M. S. Pithia, S. Singh, K. Ram Soren, H. Kudapa, D. Jarquín, P. Cubry, L. T. Hickey, G. Prasad Dixit, A. C. Thuillet, A. Hamwieh, S. Kumar, A. A. Deokar, S. K. Chaturvedi, A. Francis, R. Howard, D. Chattopadhyay, D. Edwards, E. Lyons, Y. Vigouroux, B. J. Hayes, E. von Wettberg, S. K. Datta, H. Yang, H. T. Nguyen, J. Wang, K. H. M. Siddique, T. Mohapatra, and J. L. Bennetzen. 2021. “A Chickpea Genetic Variation Map Based on the Sequencing of 3,366 Genomes.” *622 | Nature |* 599. doi: 10.1038/s41586-021-04066-1.

Wang, J., X. Liao, C. Gu, K. Xiang, J. Wang, S. Li, L. R. Tembrock, Z. Wu, and W. He. 2022. “The Asian Lotus (*Nelumbo Nucifera*) Pan-Plastome: Diversity and Divergence in a Living Fossil Grown for Seed, Rhizome, and Aesthetics.” *Ornamental Plant Research* 2(1):1–10. doi: 10.48130/OPR-2022-0002.

Wingett, S. W., and S. Andrews. 2018. “Fastq Screen: A Tool for Multi-Genome Mapping and Quality Control.” *F1000Research* 7. doi: 10.12688/f1000research.15931.1.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. “A Greedy Algorithm for Aligning DNA Sequences.” *JOURNAL OF COMPUTATIONAL BIOLOGY* 7(2):203–14. doi: https://doi.org/10.1089/10665270050081478.

Zhao, Q., Q. Feng, H. Lu, Y. Li, A. Wang, Q. Tian, Q. Zhan, Y. Lu, L. Zhang, T. Huang, Y.n Wang, D. Fan, Y. Zhao, Z. Wang, C. Zhou, J. Chen, C. Zhu, W. Li, Q. Weng, Q. Xu, Z. X. Wang, X. Wei, B. Han, and X. Huang. 2018. “Pan-Genome Analysis Highlights the Extent of Genomic Variation in Cultivated and Wild Rice.” doi: 10.1038/s41588-018-0041-z.

Chapter 3

The Evolutionary History of *Phaseolus vulgaris* As Revealed By Chloroplast and Nuclear Genomes

Abstract

Knowledge about the origin, evolution and expansion of crop species is crucial for their conservation and exploitation. *Phaseolus vulgaris* has a unique evolutionary history, with the wild form originated in Mesoamerica and subsequently introduced into South America, leading to the formation of two South American wild gene pools in North Peru and Ecuador and in South Andes. However, the debate on common bean origin is still open. Indeed, recent study proposed the so-called "*Pseudovulgaris*" hypothesis on the origin of common bean, that indicates the formation of the North Peru and Ecuador gene pool as occurred much earlier than that of *P. vulgaris* species and, thus of the diversification of Mesoamerican and Andean gene pools. In this case, the North Peru-Ecuador population represents a different species, named *P. pseudovulgaris* (*P. debouckii*) and it shared a common ancestor with the Mesoamerican and Andean groups, that remains to be discovered or has become extinct. Here, by analyzing the phylogeny of *P. vulgaris* we aim to better investigate the *P. vulgaris* origin and verify the different hypotheses. A wide sample that represents the entire geographical distribution of the wild forms of the species was genetically characterized for chloroplast genome diversity. A concatenated sequence of 3,231 chloroplast informative sites was used to build a phylogenetic tree. Moreover, 37 *de-novo* chloroplast genomes were assembled and used to provide a temporal frame of the divergence for the analyzed genotypes, suggesting that the separation between the Mesoamerican and the North Peru-Ecuador gene pools occurred 0,15 Mya. Nuclear data, from the resequencing of a sample of ten accessions, were used to corroborate the results. Overall, analyses of nuclear and plastid data support monophyletic and Mesoamerican origin of common bean.

Introduction

Phylogenetics is an essential tool for inferring evolutionary relationships between individuals, species, genes, and genomes. The enormous progresses of technology lead to the development of molecular phylogenetics. Indeed, NGS approaches allow the reconstruction of phylogenetic tree from DNA or protein sequences. If at the beginnings of phylogenomics the primary purpose was to estimate the relationships among the species represented by the sequence analyzed, today the purposes have expanded to include the understanding of the relationships among the sequences themselves, inferring the functions of genes that have not been studied experimentally (Hall et al. 2009). Since the use of morphological data, the accuracy of phylogenetic inference has always been discussed and molecular data are not free from problems. As reported in Nabhan and Sarkar (2012), (i) conflicting signals, (ii) inadequate rate of sequence evolution and (iii) violation of method assumptions are some of the issues that can be observed in molecular data. Considering both gene trees and species trees, the analysis of multiple concatenated loci may result in lack of resolution and incongruent phylogenies. The second problem depends on the fact that very different phylogenetic reconstruction can be obtained from the use of genetic markers characterized by different evolutionary rate. Finally, despite of the availability of various model for sequence evolution, real data may violate one or more assumptions. Indeed, the great majority of sequence evolution models assumes that the evolutionary processes follow stationary, reversible, and homogeneous condition (Naser-Khdour et al. 2019). Those assumptions imply that the marginal frequencies of the nucleotides or amino acids are constant over time, substitution rates between nucleotides or amino acids are equal in both directions and constant along the tree.

Phylogenetic analyses have always been subjected to bias due to recombination events, especially if performed with nuclear data. Indeed, recombination implies that different parts of the sequence potentially have highly different phylogenetic histories. Even though, the analysis of recombining sequences could hypothetically represent an advantage because it would allow the investigation of various evolutionary processes, recombination is a true limit of phylogenetic analysis (Schierup and Hein 2000). The possibility that analyzed sequences are related by more than one tree, due to recombination events, makes the representation of the evolutionary relationships through a unique phylogenetic tree somewhat incomplete. Performing phylogenetic analysis ignoring recombination may lead to artifacts (i) in the estimation of the time to most recent common ancestor or (ii) in the estimation of the amount of recent divergence, (iii) overestimation of the number of mutations (iv) apparent signs of

exponential growth, (v) apparent substitution rate heterogeneity among sites (vi) apparent parallel substitutions, (vii) loss of a molecular clock (viii) more apparent ancient polymorphism (Schierup and Hein 2000).

Even if phylogenetics is usually adopted for inferring evolutionary relations among different species, it can be useful to study also relationships among populations within a species, especially if characterized by a peculiar population structure as in the case of common bean. The aim of the present study is to clarify the evolutionary history of *P. vulgaris* investigating the relationships of the three main wild gene pools: Mesoamerican, Andean and North Peru Ecuador. Both plastid and nuclear data were examined to reconstruct the phylogeny of this species and to infer times of divergence among the wild populations.

Material and Methods

Plant Material and DNA extraction

The investigation of common bean phylogenetic history was conducted with both chloroplast and nuclear DNA. Patterns of nucleotide variability of chloroplast DNA was assessed across 97 samples of *Phaseolus spp*. Seventy wild accessions of *Phaseolus vulgaris* were selected to represent the geographical distribution of wild common bean, from Northern Mexico to Northwestern Argentina. The accessions are representative of the three different gene pools of the species: 48 Mesoamerican, 18 Andean and 4 from North Peru and Ecuador that are characterized by the ancestral type I of the Phaseolin protein (Debouck et al. 1993; Kami et al. 1995) (Figure 3.1). Wild samples of *Phaseolus coccineus* (22), *Phaseolus lunatus* (3) and one wild and one domesticated accession of *Phaseolus acutifolius* were included in the panel.

Seeds were provided by the United States Department of Agriculture Western Regional Plant Introduction Station (USDA) and the International Center of Tropical Agriculture (CIAT) in Colombia. A complete list of the accessions studied is available in Table 3.1. A code was assigned to each sample: (i) a unique numeric code for each accession, (ii) species of belonging (Pv: *Phaseolus vulgaris*; Pc: *Phaseolus coccineus*; Pa: *Phaseolus acutifolius*; Pl: *Phaseolus lunatus*), (iii) gene pool (M: Mesoamerica; A: Andes; PhI: Phaseolin I type) and/or the corresponding accession status (W: wild; D: domesticated), (iv) country of origin. The DNeasy Plant Mini Kit from QIAGEN was utilized to extract genomic DNA from young leaves of single plants grown in greenhouse. DNA libraries were constructed and sequenced from both-ends (paired-ends) with Illumina Nextera XT sample preparation kit for the plastid dataset and Illumina DNA PCR free kit for the nuclear dataset.

Table 3.1 Panel of accessions selected to study the cpDNA.

Project code	Species	Accession Number	Country
P043_Pv_MW_GT	*P. vulgaris*	G19909	Guatemala
044_Pv_MW_MX	*P. vulgaris*	G20515	Mexico
P047_Pv_MW_CO	*P. vulgaris*	G21117	Colombia
069_Pv_MW_CO	*P. vulgaris*	G23462	Colombia
501_Pv_MW_MX	*P. vulgaris*	PI417671	Mexico
059_Pv_MW_CR	*P. vulgaris*	G23418	Costa Rica
746a_Pv_MW_MX	*P. vulgaris*	G12879	Mexico
787a_Pv_MW_GT	*P. vulgaris*	G23439	Guatemala
790_Pv_MW_HN	*P. vulgaris*	G50724	Honduras
887_Pv_MW_MX	*P. vulgaris*	NI1433	Mexico
911_Pv_MW_MX	*P. vulgaris*	86	Mexico
951_Pv_MW_MX	*P. vulgaris*	M31	Mexico
716_Pv_AW_PE	*P. vulgaris*	G23458	Peru
P832_Pv_MW_MX	*P. vulgaris*	G23470	Mexico
835_Pv_MW_MX	*P. vulgaris*	G23551	Mexico
id837fa1_Pv_MW_MX	*P. vulgaris*	G2771	Mexico
id837fa2_Pv_MW_MX	*P. vulgaris*	G11056	Mexico
id837fa3_Pv_MW_MX	*P. vulgaris*	G12957	Mexico
id837fa4_Pv_MW_MX	*P. vulgaris*	G13021	Mexico
id837fa7_Pv_MW_MX	*P. vulgaris*	G50899	Mexico
id837fa8_Pv_MW_MX	*P. vulgaris*	G12873	Mexico
007_Pv_MW_MX	*P. vulgaris*	G9989	Mexico
010_Pv_MW_MX	*P. vulgaris*	G11050	Mexico
011_Pv_MW_MX	*P. vulgaris*	G11051	Mexico
015_Pv_MW_MX	*P. vulgaris*	G12872	Mexico
016_Pv_MW_MX	*P. vulgaris*	G12877	Mexico
017_Pv_MW_MX	*P. vulgaris*	G12896	Mexico
019_Pv_MW_MX	*P. vulgaris*	G12922	Mexico
020_Pv_MW_MX	*P. vulgaris*	G12924	Mexico
021_Pv_MW_MX	*P. vulgaris*	G12927	Mexico
022_Pv_MW_MX	*P. vulgaris*	G12930	Mexico
028_Pv_MW_MX	*P. vulgaris*	G13505	Mexico
056_Pv_MW_CO	*P. vulgaris*	G22304	Colombia
057_Pv_MW_MX	*P. vulgaris*	G22837	Mexico
065_Pv_MW_MX	*P. vulgaris*	G23429	Mexico
076_Pv_MW_MX	*P. vulgaris*	G23652	Mexico
080_Pv_MW_MX	*P. vulgaris*	G24378	Mexico
081_Pv_MW_MX	*P. vulgaris*	G24571	Mexico

179_Pv_MW_MX	*P. vulgaris*	PI318696	Mexico
187_Pv_MW_MX	*P. vulgaris*	PI325677	Mexico
205_Pv_MW_MX	*P. vulgaris*	PI417775	Mexico
504_Pv_MW_MX	*P. vulgaris*	PI535430	Mexico
505_Pv_MW_MX	*P. vulgaris*	PI535409	Mexico
506_Pv_MW_MX	*P. vulgaris*	PI535450	Mexico
006_Pv_AW_AR	*P. vulgaris*	G7469	Argentina
031_Pv_AW_AR	*P. vulgaris*	G19888	Argentina
033_Pv_AW_AR	*P. vulgaris*	G19891	Argentina
038_Pv_AW_AR	*P. vulgaris*	G19897	Argentina
039_Pv_AW_AR	*P. vulgaris*	G19898	Argentina
052_Pv_AW_AR	*P. vulgaris*	G21199	Argentina
062_Pv_AW_PE	*P. vulgaris*	G23422	Peru
P064_Pv_AW_PE	*P. vulgaris*	G23426	Peru
066_Pv_AW_BO	*P. vulgaris*	G23444	Bolivia
067_Pv_AW_BO	*P. vulgaris*	G23445	Bolivia
068_Pv_AW_PE	*P. vulgaris*	G23455	Peru
232_Pv_AW_AR	*P. vulgaris*	W617499	Argentina
243_Pv_AW_BO	*P. vulgaris*	W618826	Bolivia
656_Pv_AW_AR	*P. vulgaris*	G19902	Argentina
665_Pv_AW_AR	*P. vulgaris*	NI1423	Argentina
668a_Pv_AW_BO	*P. vulgaris*	G23442	Bolivia
717_Pv_AW_PE	*P. vulgaris*	G23459	Peru
P718a_Pv_AW_PE	*P. vulgaris*	G23419	Peru
034_Pv_AW_AR	*P. vulgaris*	G19892	Argentina
040_Pv_AW_AR	*P. vulgaris*	G19901	Argentina
715_Pv_AW_PE	*P. vulgaris*	G23456A	Peru
013_Pv_AW_PE	*P. vulgaris*	G12856	Perù
id837fa6_Pv_PhI_EC	*P. vulgaris*	G23724	Ecuador
073_Pv_PhI_EC	*P. vulgaris*	G23582	Ecuador
075_Pv_PhI_PE	*P. vulgaris*	G23587	Perù
078_Pv_PhI_EC	*P. vulgaris*	G23726	Ecuador
Pc_1_W_MX	*P. coccineus*	PI430189	Messico
Pc_3_W_MX	*P. coccineus*	NI677	Mexico
Pc_4_W_MX	*P. coccineus*	NI819	Mexico
Pc_18_W_MX	*P. coccineus*	NI1120	Mexico
Pc_20_W_MX	*P. coccineus*	PI325598	Mexico
Pc_8_W_MX	*P. coccineus*	NI1092	Mexico
Pc_21_W_MX	*P. coccineus*	NI1117	Mexico
Pc_2_W_MX	*P. coccineus*	PI430183	Messico
Pc_5_W_MX	*P. coccineus*	NI1265	Mexico

Pc_6_W_MX	*P. coccineus*	PI417607	Mexico
Pc_7_W_MX	*P. coccineus*	PI417608	Mexico
Pc_9_W_MX	*P. coccineus*	PI417593	Messico
Pc_10_W_MX	*P. coccineus*	PI430178	Messico
Pc_11_W_MX	*P. coccineus*	PI346950	Messico
Pc_12_W_MX	*P. coccineus*	NI818	Mexico
Pc_13_W_MX	*P. coccineus*	NI1213	Mexico
Pc_14_W_MX	*P. coccineus*	NI726	Mexico
Pc_15_W_MX	*P. coccineus*	NI813	Mexico
Pc_16_W_MX	*P. coccineus*	NI1122	Mexico
Pc_17_W_MX	*P. coccineus*	NI1028	Mexico
Pc_19_W_MX	*P. coccineus*	NI1125	Mexico
Pc_22_W_MX	*P. coccineus*	NI1325	Mexico
Pa_D_SV	*P. acutifolius*	PI200902	El Salvador
Pa_W_MX	*P. acutifolius var. acutifolius*	PI319445	Mexico
Pl_D_MX	*P. lunatus*	PI313212	Mexico
Pl_W_GT	*P. lunatus*	NI1689	Guatemala
Pl_W_PE	*P. lunatus*	NI1771	Peru

For the analysis of nuclear DNA, a restricted sample of 10 accessions of *P. vulgaris* was chosen from the previous panel to be sequenced with higher coverage (Table 3.2). Particularly, the selection was based on both geographic criteria, guarantying the representation of the Mesoamerica, Andes and North Peru-Ecuador and the haplogroups highlighted by the investigation of plastomes (Figure 3.1).

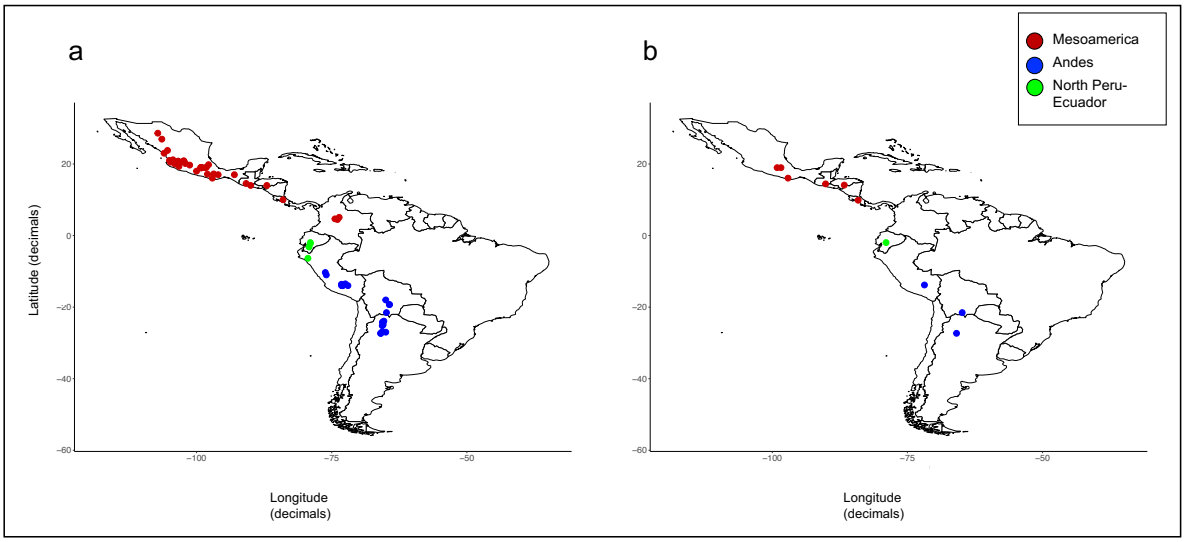


Figure 3.1 (a) Representation of the geographic distribution of the P. vulgaris samples used for the analysis of the cpDNA and (b) nuclear DNA. Mesoamerican accessions are showed in red, Andean accessions in blue and North Peru-Ecuador (PhI) accessions in green.

Table 3.2 Panel of accessions selected to study the nuclear DNA.

Project code	Species	Gene-pool	Country
187_Pv_MW_MX	*P. vulgaris*	Mesoamerican Wild	Mexico (Morelos)
065_Pv_MW_MX	*P. vulgaris*	Mesoamerican Wild	Mexico (Puebla)
081_Pv_MW_MX	*P. vulgaris*	Mesoamerican Wild	Mexico (Oaxaca)
787A_Pv_MW_GT	*P. vulgaris*	Mesoamerican Wild	Guatemala (Santa Rosa)
790_Pv_MW_HN	*P. vulgaris*	Mesoamerican Wild	Honduras (El Paraiso)
59_Pv_MW_CR	*P. vulgaris*	Mesoamerican Wild	Costa Rica (San Jose)
038_Pv_AW_AR	*P. vulgaris*	Andean Wild	Argentina (Tucuman)
067_Pv_AW_BO	*P. vulgaris*	Andean Wild	Bolivia (Tarija)
716_Pv_AW_PE	*P. vulgaris*	Andean Wild	Peru (Cuzco)
078_Pv_PhI_EC	*P. vulgaris*	North Peru-Ecuador (PhI)	Ecuador (Chimborazo)

Plastid data

Reference Mapping and SNPs calling

Quality control of raw reads from the 97 *Phaseolus* accessions was performed using FastQC (Andrews et al. 2010), before and after read pre-processing. The command line tool Trimmomatic (Bolger et al. 2014) was used to remove Illumina technical sequences and to filter out low quality reads. Reads ≥ 75 nucleotides in length with a minimum Q-value of 20 were retained for downstream analyses. FastQscreen (Wingett and Andrews 2018) and the sequence aligner Bowtie2 (default settings) (Langmead and Salzberg 2012) were used to align the high-quality reads to the nuclear (G19833) and plastid (NC_009259) reference genomes of common bean and to extract chloroplast reads. Following the pipeline in Nock et al. (2019), the SNP calling was made with the "mpileup" utility of the SAMtools/BCFtools software (Li et al. 2009) using the complete chloroplast genome sequence of *P. vulgaris* (NC_009259) as a reference. Single VCF files were merged with VCFtools (Danecek et al. 2011) to obtain the final file and SnpEff (Cingolani et al. 2012) was used to annotate and predict the effects of the SNPs.

Singletons were filtered out because not informative. SNP distribution was computed through a window analysis (window size =100 nucleotides) performed using VCFtools.

Assembly of plastomes

As described in the previous chapter, the reference guided assembly of 39 chloroplast genome was carried out. Thirty-seven of the previously *de-novo* assembled plastomes were used in the current chapter to investigate *Phaseolus* phylogeny and to clarify the phylogeny of common bean.

Genetic Structure and MDS analyses

To investigate the structure of the analyzed populations, a cluster analysis was run on the SNP dataset with the BAPS v6.0 software (Cheng et al., 2013). We chose a mixture model based analysis since all markers are likely linked.

Pair-wise identity by state (IBS) distances estimates among all 97 samples and among the *P. vulgaris* accessions were calculated using PLINK v1.90b52 and graphically represented by a multidimensional scaling (MDS) plot.

Phylogenetic analyses and haplotype network

The VCF file was converted into a FastA file using a custom *perl* script. *P. lunatus* was selected as outgroup (Delgado-Salinas et al. 1999). A Maximum Likelihood tree was computed by RAxML 8.1.2 (Stamatakis 2014) with the GTR substitution model and bootstrap value of 10,000. The same analysis was carried out with the 37 assembled plastomes. Trees were visualized with FigTree v1.4.4.

The haplotype network analysis was performed using PopART (Leigh and Bryant 2015) with the TCS type of network (Clement et al. 2002.). Input samples of *P. vulgaris* were previously divided in 20 groups based on geographic distribution and gene pools.

Nuclear data

Reference Mapping and SNPs calling

For the analysis of nuclear data of 10 accessions of *P. vulgaris* a slightly different approach was carried out. SNP calling was performed using the sequence_handling pipeline (Hoffman et al., 2018) available at the Minnesota Supercomputing Institute. This pipeline is composed by a series of scripts, called *handlers* that automate and speed up DNA sequence alignment and quality control. FastQC (Wingett and Andrews 2018) was used to perform read quality check before and after the trimming of the adapters. Adapter contamination was trimmed using Scythe (v. 1.2.8, https://github.com/vsbuffalo/scythe) using a prior contamination rate of 0.05, as suggested by the Scythe documentation. Sequences were aligned to the *P. vulgaris* reference genome (accession G19833, v. 2.1) using BWA-MEM (v. 0.7.17, Li 2013) with default parameters. The resulting SAM files were sorted, de-duplicated and read groups were added with Picard v. 2.4.1 (http://broadinstitute.github.io/picard/).

Haplotype calling was performed with GATK 4.1.2 (Poplin et al. 2017) using a nucleotide sequence diversity estimate of θW =0.001. The Watterson Theta value (θW) was estimated with ANGSD (Sand Korneliussen, Albrechtsen, and Nielsen 2014) on the base of the analyzed samples and an ancestral sequence obtained from mapping *P. lunatus* reads to the reference genome of *P. vulgaris*.

The resulting gVCF files were used to jointly call SNPs on all samples. To reduce the computational time, this process was parallelized across chromosomes. VCF files for all genomic regions were then concatenated into one file. A hard-filtering approach was used to increase the quality of the call-set (Danecek et al. 2021). Indels, non-biallelic sites, low quality sites (missingness ≥ 50%) and sites with *maf* (minor allele frequencies) ≤ 0.01were filtered.

Finally, singletons were removed from the final set of SNPs to avoid noisy signals due to long-branch attraction effects. SNPs were annotated with SnpEff (Cingolani et al. 2012).

Intraspecies Phylogenetic analyses

The relationships among the re-sequenced *P. vulgaris* individuals were investigated using SNPs from whole genome filtered every 250 kb and SNP located in the centromeric region of each chromosome. The Neighbor Joining (NJ) method implemented in MEGA X (Kumar et al. 2018) was applied to reconstruct the phylogeny of common bean with the set of SNPs placed across the genome, a bootstrap value of 10.000 was used.

For a dipper investigation of the evolutionary history of the species, phylogenetic reconstruction was performed using a Maximum Likelihood approach implemented in RAxML (Stamatakis 2014). In this case, to avoid issues due to recombination, only those SNPs included in the centromeric regions of chromosomes were concatenated. A ML tree was performed for each centromere of the eleven chromosomes of common bean (bootstrap value 100,000).

Linkage disequilibrium in centromeres

To characterize the centromeres, the r^2 representing the correlation between allele frequencies at pairs of SNPs along centromeric regions was computed with PopLDdecay software (Zhang et al. 2019) setting as maximum distance between SNPs the dimension of each centromere. The mean estimated r^2 values were plotted against the physical distance between SNPs to trace the linkage disequilibrium decay.

Species divergence dating

With the aim to better understand the time of divergence among the species under investigation, a molecular clock analysis was performed with plastid data.

The 37 plastomes assembled (see Chapter 2) were aligned with three plastomes belonging to *Vigna spp.* available in Genbank (NC_013843, NC_018051, NC_021091 corresponding to *V. radiata*, *V. unguiculata* and *V. angularis*, respectively) and analyzed using the Bayesian approach implemented in BEAST v2.6.2 (Bouckaert et al. 2014). The *BEAST method was applied to produce the XML file and the coalescent simulation was run applying a relaxed lognormal molecular clock with a General Time Reversable (GTR) model and calibrating the tree with the divergence between *P. coccineus* and *Vigna spp* reported in Lavin et al. (2005) (μ= 1,23*10^{-3} sub/site/year). The MCM chain was set to 100,000,000 and two independent runs were performed and finally combined.

Results

Plastid data

Sequencing and mapping

Raw sequence reads of 97 *Phaseolus* accessions were obtained and mapped to the reference plastome of *P. vulgaris*, (NC_009259). An average of 248,794 reads per sample were mapped to the reference with an average base coverage of 209. As expected, most of the reads were successfully aligned only to the nuclear genome, whilst only the 20,45% of the sequenced reads were of plastid origin.

Identification and Analysis of SNP variation

A set of 4008 SNPs (777, singletons included) was identified across the 97 *Phaseolus spp* samples. All over, 1999 SNPs were in genes and the 66% of the genes harbor at least one SNP. Fifty-six of 128 genes were affected by more than 3 variants and the most variable genes were *ndhF*, *accD* and the pseudo gene *ycf1b* with 100, 115 and 391 SNPs, respectively. 1,526 SNPs fall within exons and 473 within introns; we found the *psbH* gene presented variants only in the exon. The majority of the SNPs were distributed across the single copy regions. A higher SNP density was found in the SSC region (5,22 SNPs per 100 bp window on average) compared to the LSC region (3,76 SNPs per 100 bp window on average). The 42.3% of the variants were synonymous and the 57,7% were non-synonymous, the 56.97% of the variants belonging to the latter category were missense SNPs.

Genetic structure

The SNP dataset was used to investigate the genetic structure within the *P. vulgaris* group. The BAPS structure analysis identified five subgroups (C1-C5) that best define the population (Figure 3.2, panel a and b) (Logmarginal likelihood of optimal partition: -6689.6434). Due to the high conservation of the plastomes, all the samples showed the highest percentage of membership to the corresponding cluster. All four PhI samples were clustered together in C1. The Mesoamerican accessions were split into three groups C2, C3, C5 (Figure 3.2). Cluster C2 was characterized by three samples from Guatemala, Costa Rica and Honduras, while cluster C3 included only Mexican accessions and cluster C5 was composed by 25 samples from Mexico, two from Colombia and one from Guatemala. All the Andean accessions clustered together in C4.

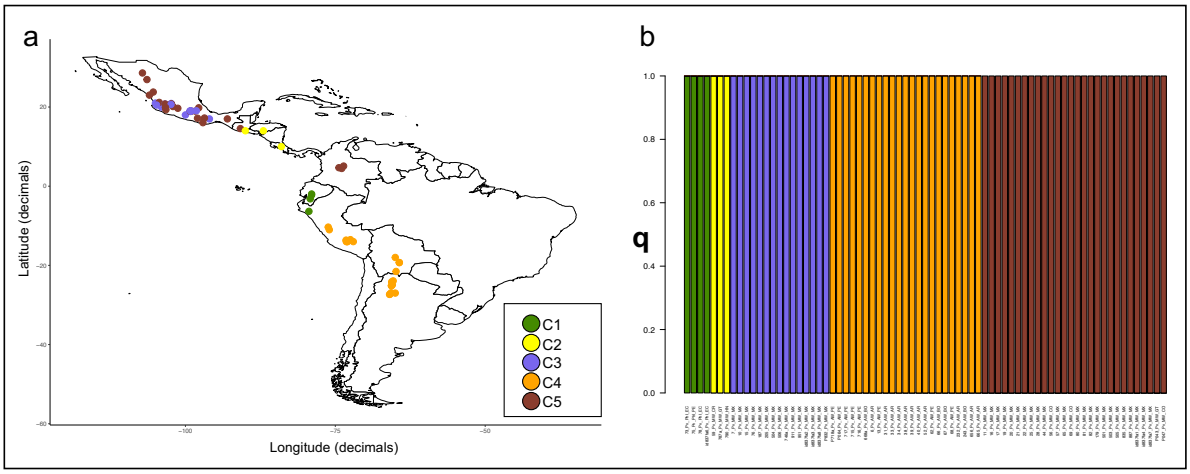


Figure 3.2 Results of the BAPS structure analysis. (a) Geographical distribution of the P. vulgaris accessions based of BAPS cluster membership. (b) Structure analysis performed with BAPS.

Relationships among the accessions and the species were also inspected by a Multidimensional scaling (MDS) analysis based on pair-wise identity by state (IBS) distances. The MDS plot of the 97 samples (Figure 3.3, panel a) separated the accessions by species. Specifically, the first component (C1) parted *P. acutifolius* and *P. lunatus*; conversely, the second component (C2) divided *P. vulgaris* from *P. coccineus*. Within the *P. vulgaris* group (Figure 3.3, panel b), the C1 component differentiated the Andean and Mesoamerican gene pools, in addition it split the accessions belonging to the latter one into three groups, one of them closer to the Andean samples. The second component (C2) separated the PhI gene pool from the Andean and Mesoamerican one as well as the 59_Pv_MW_CR, 787a_Pv_MW_GT and 790_MW_HN accessions from the rest of the Mesoamerican samples.

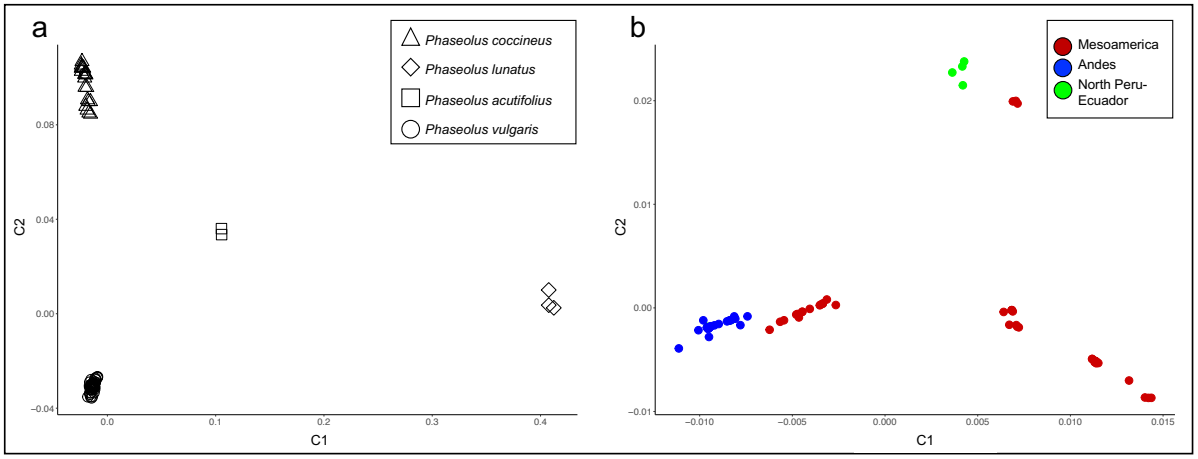


Figure 3.3 Results of the Multidimensional Scaling Analysis (MDS). Component C1 and C2 are reported in x and y axes, respectively. (a) MDS plot of all the accessions included in the study; (b) MDS plot of P. vulgaris samples.

Phylogenetics analysis and haplotype network

A concatenated sequence of 3231 informative sites of the chloroplast SNP dataset was used to investigate the phylogeny of common bean (Figure 3.4, panel a). The ML tree (10.000 bootstrap value) clearly highlighted the presence of a genetic structure in Mesoamerica with four statistically well supported groups (bootstrap value > 75%), one including accessions from Guatemala, Honduras and Costa Rica which is more closely related to North Peru-Ecuador accessions, two including only Mexican accessions with one of them closer to the Andean group, and the remaining one including samples from Mexico, Guatemala and Colombia. The accessions from North Peru-Ecuador were placed in a clade which clearly derived from the Mesoamerican gene pool. In order to clarify the phylogeny of *Phaseolus vulgaris*, a ML tree (10.000 bootstrap values) was also inferred from the alignment of the 37 chloroplast genomes obtained from the *de-novo* assembly (Figure 3.4, panel b). The phylogenetic tree had strong bootstrap supports for all the branches and showed a topology consistent with the tree based on SNPs' concatemer.

PopArt software detected forty-five haplotypes within the vulgaris group (Figure 3.4, panel c). No haplotypes were shared between the Mesoamerican and Andean gene pools. Within the Mesoamerican gene pool, four Mexican and three Columbian accessions shared the same haplotype, as well as two samples from Mexico and one from Guatemala. Moreover, in the Andean gene pool two samples from Peru shared the haplotype with an accession from Bolivia and seven samples from Argentina showed the same haplotype. A validation of the relationship between the PhI gene pool and the Mesoamerican and Andean accessions observed in the previous phylogenetic trees was provided by this haplotype network. The type I Phaseolin

accessions showed haplotypes that were closer to the Mesoamerican gene pool, specifically to the three samples from Guatemala, Honduras and Costa Rica (i.e., 787a_Pv_MW_GT, 790_Pv_MW_HN and 059_Pv_MW_CR, respectively) and mostly separated from the Andean samples.

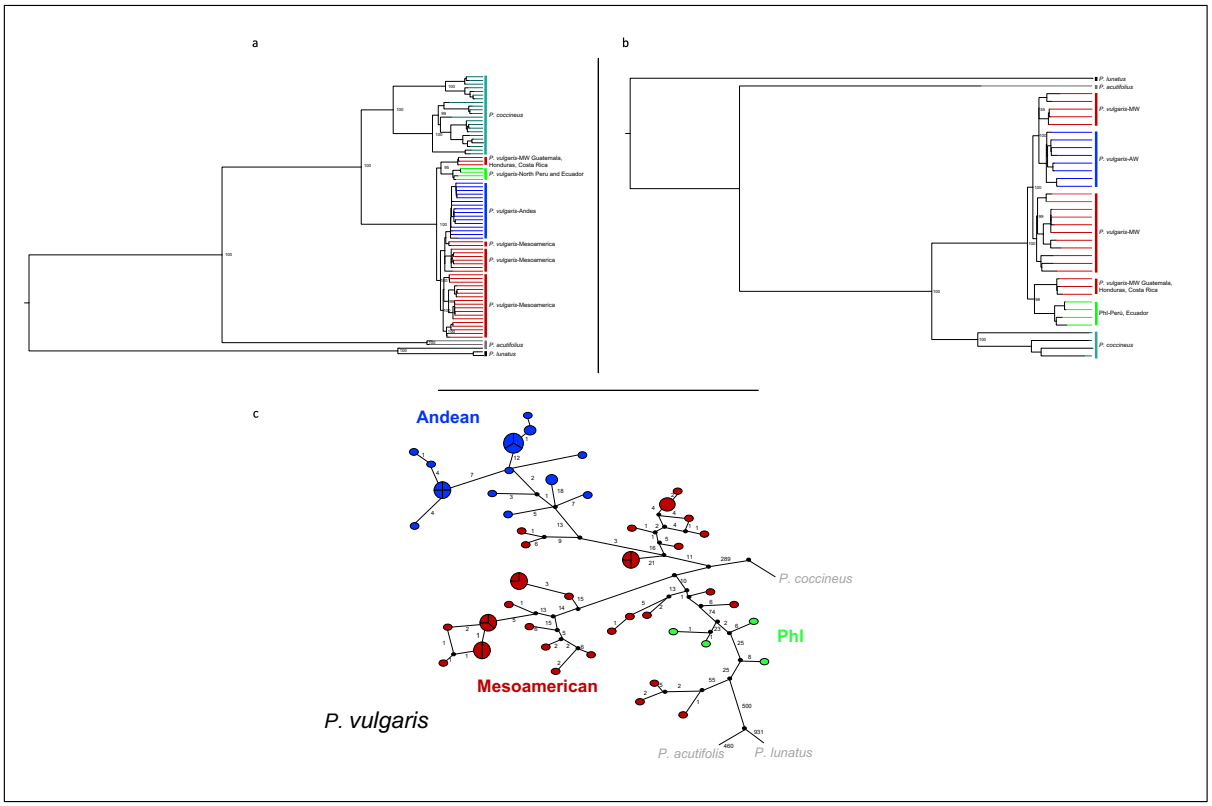


Figure 3.4 Maximum likelihood trees and haplotype network: (a) ML tree obtained from 3231 SNPs collected from 97 samples of Phaseolus spp, bootstrap value=10,000; (b) ML tree obtained from the alignment of 37 de-novo chloroplast genomes, bootstrap value=10,000. (c) Haplotype network resulted from the analysis of all Phaseolus accessions with a focus on P. vulgaris. MW: Mesoamerican wild; AW: Andean Wild; PhI: Phaseolin type I. Each circle represents a single different haplotype and circle sizes are proportional to the number of individuals that carry the same haplotype. Black dots indicate missing intermediate haplotypes and numbers correspond to mutational steps.

Nuclear data

Sequencing and mapping

Raw reads of 10 accessions of *Phaseolus vulgaris*, obtained from whole genome sequencing, were mapped against the species' reference genome (accession G19833, v. 2.1).

An average of 54768987 reads per sample were mapped to the reference, with a final mapping coverage ~10X.

Identification and Analysis of SNP variation

After filtering, a set of 11,160,422 SNPs was identified across the 10 *P. vulgaris* samples. According to the SnpEff report, the majority of SNP effects were found in intergenic (45.297 %), upstream (19.885 %) and downstream (20.00 %) regions, respectively. The SNP effects found in intronic regions was 7.452 % of the total effects, while, the percentage of SNP effects located in exonic regions was 4.862 %.

Intraspecies Phylogenetic analyses

SNPs were used to investigate the relationships among the *P. vulgaris* samples. Firstly, a NJ tree was reconstructed from the concatenation of nuclear SNPs placed across the genome and selected every 250 kb (Figure 3.5). All nodes were well supported with a percentage higher than 90%, except for the node splitting Mesoamerican and Andean samples. All Andean samples cluster together and the Mesoamerican accession from Honduras (i.e., 790_Pv_MW_HN) was included in the Andean sub-tree and placed close to the 038_Pv_AW_AR. Mesoamerican samples were split into three groups, two of them were composed by accessions from Mexico and from Guatemala and Costa Rica and included in the same bifurcation of Andean accessions. Conversely, sample 187_Pv_MW_MX clustered together with the North Peruvian-Ecuadorian genotype which behaved as an outgroup.

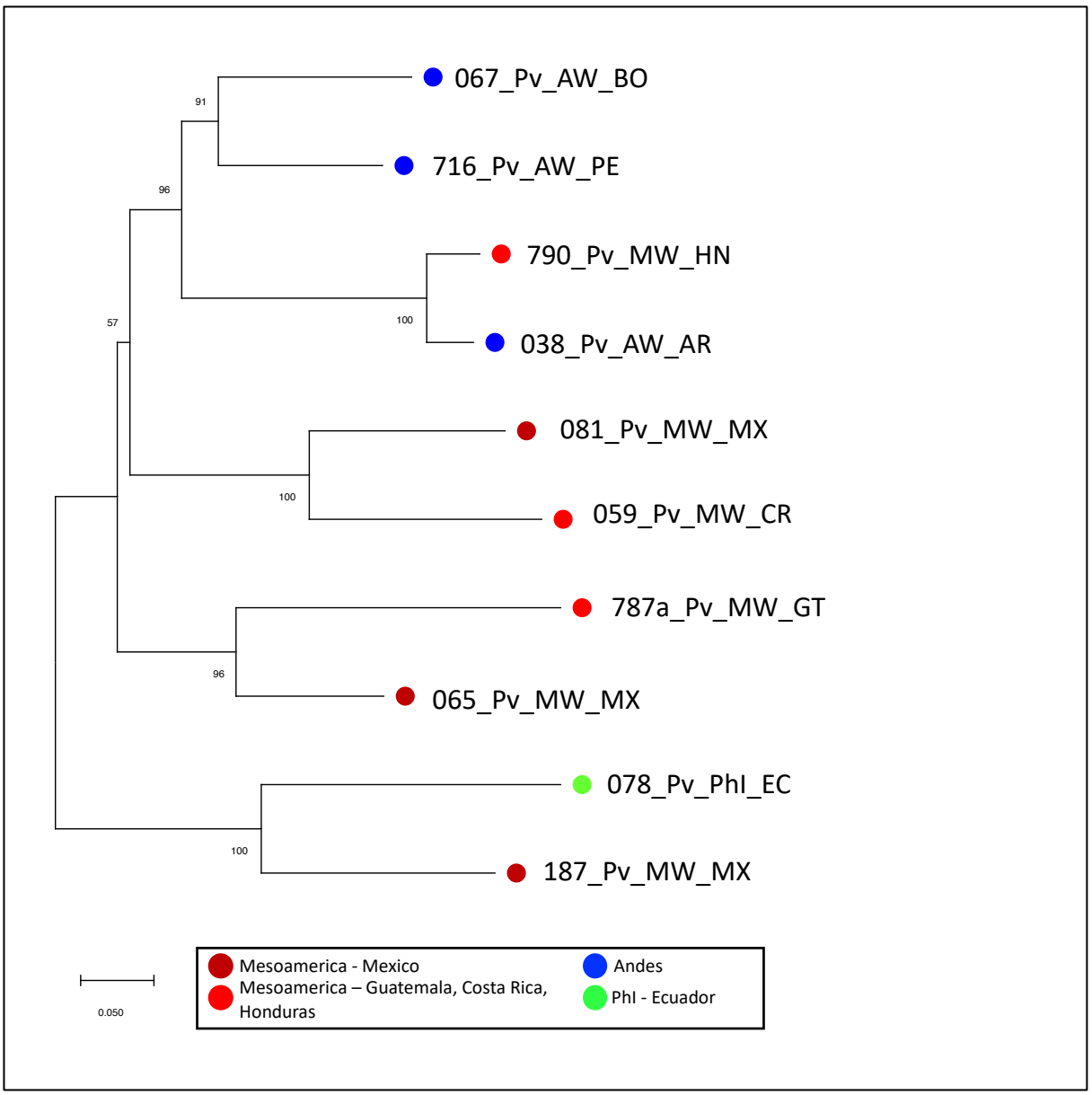


Figure 3.5 Neighbor Joining tree performed with nuclear SNPs selected every 250 kb with a bootstrap value of 10,000. In dark red: Mesoamerican samples from Mexico, light red: Mesoamerican samples from Guatemala, Costa Rica or Honduras, blue: Andean samples and in green: PhI sample from Ecuador

To better investigate common bean phylogeny, Maximum likelihood trees were constructed with SNPs located in centromeric regions of each chromosome (Table 3.3), thus a total of eleven trees were analyzed (Figure 3.6).

Table 3.3 Centromeric regions and corresponding number of analyzed SNPs.

Chr	Centromeres Dimensions (Mb)	N. SNPs
1	7.7	170779
2	4.6	113647
3	2.1	46248
4	6.5	163594
5	7.5	173932
6	0.1	2803
7	13.6	334210
8	13.9	294893
9	4.3	109616
10	0.7	16317
11	1.0	24544

Different topologies were obtained from the eleven ML tress. In all cases the great majority of nodes resulted well supported. A not clear subdivision of the Mesoamerican and Andean samples was reported. Interesting, the PhI and the Mesoamerican samples from Costa Rica, Guatemala and Mexico (Oaxaca) (i.e., 078_Pv_PhI_EC, 059_Pv_MW_CR, 787a_Pv_MW_GT, 081_Pv_MW_MX) resulted in the same clustering group.

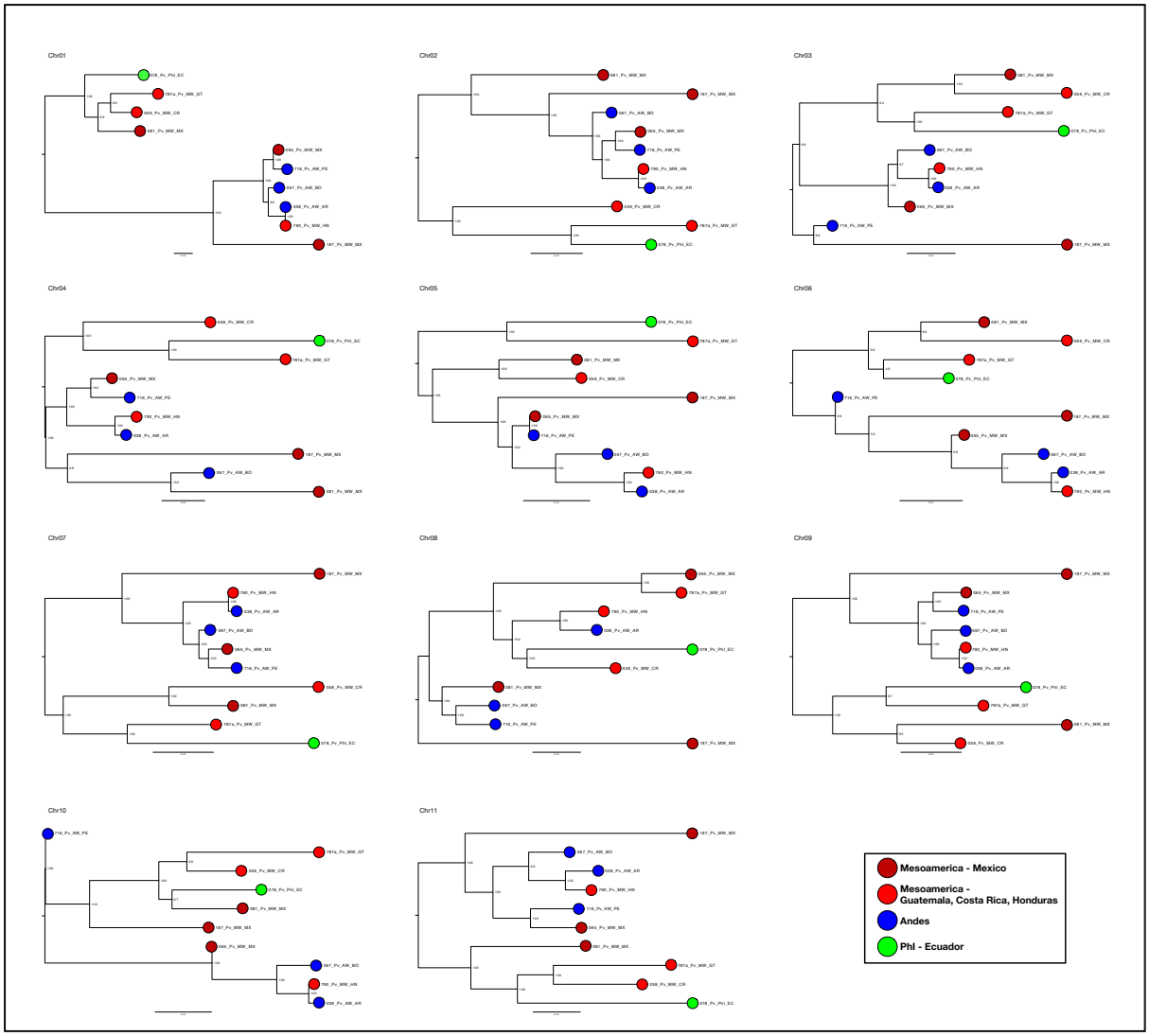


Figure 3.6 Maximum Likelihood trees of the eleven centromeric regions with bootstrap value of 100,000. In dark red: Mesoamerican samples from Mexico, light red: Mesoamerican samples from Guatemala, Costa Rica or Honduras, blue: Andean samples and in green: PhI sample from Ecuador.

To investigate the centromeric regions, LD (r^2) was computed for each centromere. On average the value of r^2 was 0.12 and the LD decay occurred very shortly (100 kb in the case of the centromeric region of chromosome 3, Figure 3.7)

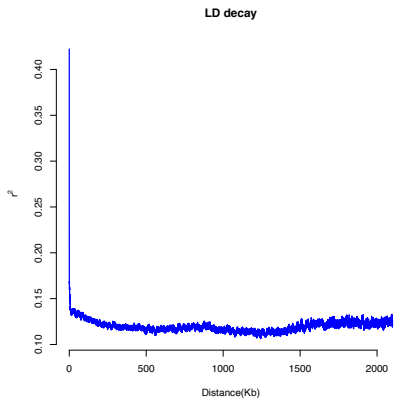


Figure 3.7 LD decay of the centromeric region of chromosome 3.

Molecular Clock analysis

To provide a divergence estimate of the vulgaris group, the 37 assembled plastomes were used to carry out a molecular clock analysis (Figure 3.8). The coalescent simulation showed a divergence time among *P. vulgaris* wild genetic groups of ~0.19 My (Million years) (0.0847-0.3082 95% Highest Posterior Probability (HPD)). The separation between the North Peru-Ecuador gene pool and the group composed of Mesoamerican and Andean gene pools was estimated ~0.15 MY (0.0607-0.2419 95% HPD). A much recent split occurred between the Mesoamerican and Andean populations (~0.09 My) (0.0422-0.1515 95% HPD).

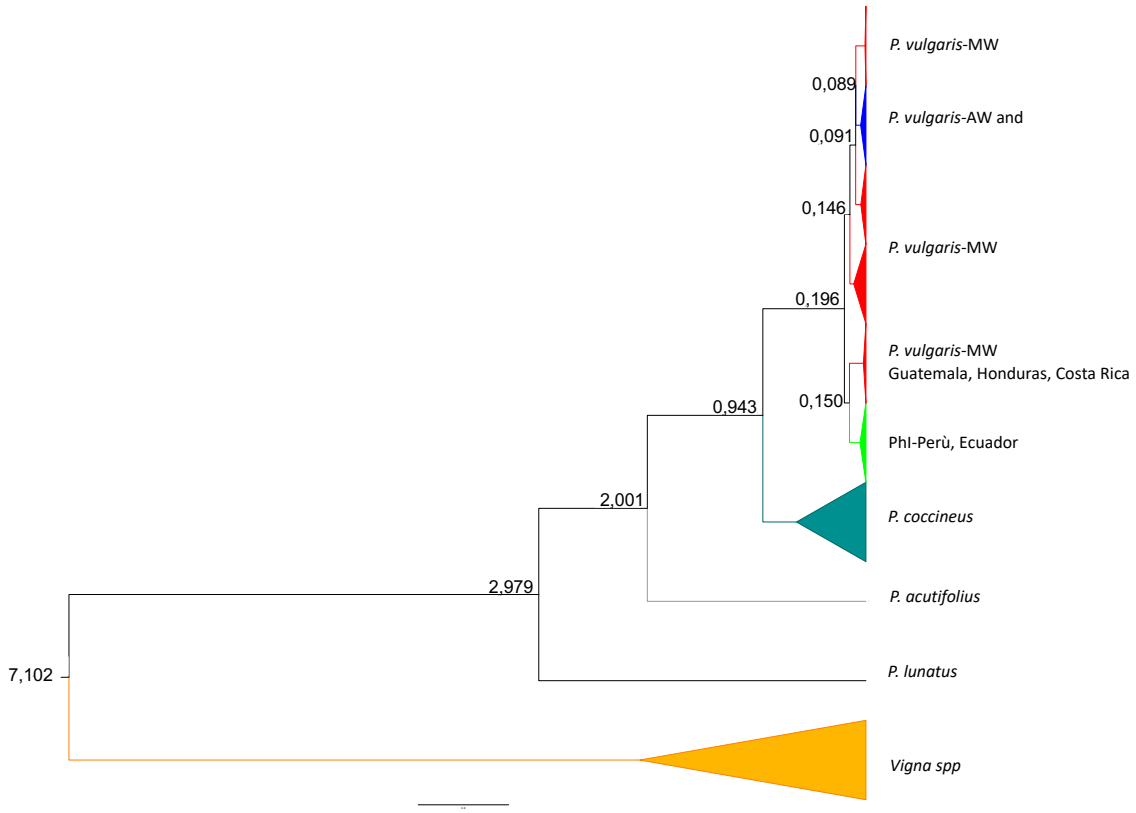


Figure 3.8 Molecular Clock analysis of the 37 de-novo assembled plastomes. Divergence times are reported on the nodes. MW: Mesoamerican Wild; AW: Andean Wild; PhI: Phaseolin type I (North Peru-Ecuador).

Discussion

The aim of the present study is to clarify the phylogeny among *Phaseolus vulgaris* wild gene pools. We analyzed both plastid and nuclear data of wild accessions of common bean, sampled to be representative of the geographic distribution of this species. To date, numerous theories have been advanced on the origin of common bean. The North Peru-Ecuador hypothesis proposed by Kami et al. (1995) was based on the analysis of the Phaseolin protein sequence, identifying an ancestral type distinctive of a core area in North Peru-Ecuador (Phaseolin I). Evidence of a Mesoamerican origin was provided by Rossi et al. (2009), Bitocchi et al. (2012) and Desiderio et al. (2013). More recently, the work of Rendón-Anaya et al. (2017) identified the North Peru-Ecuador population as a new species of *Phaseolus* (*Phaseolus debouckii*) that differentiated before the speciation event of *P. vulgaris* and Ariani et al. (2018) proposed the "Protovulgaris hypothesis" according to which an ancestral population common among the three gene pools went extinct when the Mesoamerican and Andean gene pools differentiated. The results obtained here at plastid and nuclear level clearly supported a Mesoamerican origin of common bean and did not recognize PhI accessions as different species from *P. vulgaris*.

Indeed, as previously reported by Bitocchi et al. (2012) and Desiderio et al. (2013), BAPS and MDS analysis revealed a strong subdivision of the Mesoamerican population. Four Mesoamerican clusters were identified, one of which was placed closer to the Andean cluster and another one was related to PhI accessions.

To overcome the issues due to recombination, phylogenetic analyses were performed with SNPs located in plastomes, the 37 *de-novo* assembled chloroplast genomes and SNPs located in the centromeric regions of chromosomes (nuclear data). Due to its peculiar characteristics such as the uniparental inheritance, haploidy and lack of recombination, chloroplast is suitable for the reconstruction of evolutionary history through phylogenetic analyses. Both the Maximum Likelihood (ML) trees obtained with the plastid SNPs and plastomes corroborated the results obtained from the MDS and population structure analyses. According to previous studies (Bitocchi et al, 2012) the wild accessions from North Peru-Ecuador were assigned to a clade which clearly derived from the Mesoamerican gene pool. As mentioned before, nuclear data from the centromeric regions of a sample of 10 *P. vulgaris* accessions were investigated. Indeed, centromeres are regions of the genome known to be not subjected to recombination (cross over) and in which gene presence is limited. Due to low gene density, centromeric regions are not exposed to the selection constraint, conversely a very high mutational rate characterized those areas of the genome (Bensasson 2011). As reported in other works (Rendón-Anaya et al. 2017; Ariani et al. 2018) when phylogenetic analyses are carried out with markers situated along the genome, PhI samples behave as outgroup. The same behave was observed here, when SNPs from across the whole genome were used to reconstruct a NJ tree. Instead, a deeper investigation showed that even though centromeres presented different topologies, the PhI sample clearly showed a relation with the Mesoamerican gene pool and specifically with accessions from Guatemala, Costa Rica and Oaxaca valley, the latter identified to be the presumed center of domestication (Bitocchi et al. 2013; Rodriguez et al. 2016).

With the aim to corroborate our analyses and to provide a divergence estimate of the *P. vulgaris* group, the 37 assembled plastomes were also used to carry out a molecular clock analysis. The coalescent simulation showed a divergence time among *P. vulgaris* wild genetic groups of 0.19 My (Million years). This estimate is similar to that found by Rendón-Anaya et al. (2017) for the split between an Andean (Jalo EEP558) and a Mesoamerican (BAT93) domesticated genotype based on the analysis of plastid data, but strongly different from that obtained by the same author between the North Peru and Ecuador genotype G21245 and the previously mentioned domesticated Andean and Mesoamerica genotypes (0.9 My). According to our

analysis, the divergence between the Mesoamerican and Andean gene pools was estimated ~0.09 My ago, a time comparable with the one proposed by Mamidi et al. (2013) and Ariani et al. (2018), ~110.000 years ago and ~87.000 years ago respectively.

Even though, our analysis set the divergence between the Mesoamerican gene pool and the North Peru-Ecuador accessions earlier compared to the divergence between the Mesoamerican and Andean gene pools, the time difference between the two events is strongly dissimilar from the one proposed by Rendón-Anaya et al. (2017).

Finally, our data strongly support a monophyletic and Mesoamerican origin of common bean. Moreover, no evidence was found for the identification of the PhI population as a different species from *P. vulgaris*. We propose that two migration events occurred, the first one from Mesoamerica to North Peru and Ecuador about 150.000 years ago and the second one, more recent, from Mesoamerica to South Andes about 90.000 years ago.

Conclusion

The primary definition of phylogenetics is the science that studies genealogical relationships between species over time. It is based on the acceptance of the theory of evolution as a way to explain similarities and differences between species, due to the accumulation of mutations from yesterday to today time (Cavalli-Sforza and Edwards, 1966).

The advent of DNA sequencing technologies revolutionized phylogenetic studies. Indeed, before the development of this tools, the use of phylogenetic trees was mainly restricted to taxonomy and systematics. Nowadays, phylogenies are used in the great majority of life sciences, from the analysis of paralogues in gene family to the studies of populations' histories (Yang & Rannala, 2012).

This work represents an example of phylogenetics applied to the study of the evolutionary history of populations belonging to the same species: *P. vulgaris*.

Our findings corroborate the previous hypothesis about the monophyletic and Mesoamerican origin of common bean, but they also give a deeper understanding of relationships between the three main wild gene-pools and their divergence events.

In addition, we give clear evidence of the bias due to recombination events when using nuclear data to reconstruct phylogenetic trees.

We believe that this study will help shedding light on the *P. vulgaris* intraspecific phylogeny and its origin.

Data availability

Scripts for conducting the analyses are available at GitHub at the following link: https://github.com/giuliafrascarelli/Common_bean.

References

Andrews, S., J. Gilley, and M. P. Coleman. 2010. “Difference Tracker: ImageJ Plugins for Fully Automated Analysis of Multiple Axonal Transport Parameters.” *Journal of Neuroscience Methods* 193(2):281–87. doi: 10.1016/j.jneumeth.2010.09.007.

Ariani, A., J. C. Berny M. Teran, and P. Gepts. 2018. “Spatial and Temporal Scales of Range Expansion in Wild Phaseolus Vulgaris.” *Molecular Biology and Evolution* 35(1):119–31. doi: 10.1093/molbev/msx273.

Bensasson, D. 2011. “Evidence for a High Mutation Rate at Rapidly Evolving Yeast Centromeres.” *BMC Evolutionary Biology* 11(1). doi: 10.1186/1471-2148-11-211.

Bitocchi, E., E. Bellucci, A. Giardini, D. Rau, M. Rodriguez, E. Biagetti, R. Santilocchi, P. Spagnoletti Zeuli, T. Gioia, G. Logozzo, G. Attene, L. Nanni, and R. Papa. 2013. “Molecular Analysis of the Parallel Domestication of the Common Bean (Phaseolus Vulgaris) in Mesoamerica and the Andes.” *New Phytologist* 197(1):300–313. doi: 10.1111/j.1469-8137.2012.04377.x.

Bitocchi, E., L. Nanni, E. Bellucci, M. Rossi, A. Giardini, P. Spagnoletti Zeuli, G. Logozzo, J. Stougaard, P. McClean, G. Attene, and R. Papa. 2012. “Mesoamerican Origin of the Common Bean (Phaseolus Vulgaris L.) Is Revealed by Sequence Data.” *Proceedings of the National Academy of Sciences of the United States of America* 109(14). doi: 10.1073/pnas.1108973109.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30(15):2114–20. doi: 10.1093/bioinformatics/btu170.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C. Hsi Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. “BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.” *PLoS Computational Biology* 10(4). doi: 10.1371/journal.pcbi.1003537.

Cavalli-Sforza, L. L., and A. W. F. Edwards’. 1966. *PHYLOGENETIC ANALYSIS: MODELS AND ESTIMATION PROCEDURES*. doi: 10.1111/j.1558-5646.1967.tb03411.x.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3.” *Fly* 6(2):80–92. doi: 10.4161/fly.19695.

Clement, M., Q. Snell, P. Walker, D. Posada, and K. Crandall. 2002. *TCS: Estimating Gene Genealogies*. In Parallel and Distributed Processing Symposium, International (Vol. 3, pp. 0184-0184). IEEE Computer Society.

Danecek, P., A. Auton, G. Abecasis, C. A. A., E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27(15):2156–58. doi: 10.1093/bioinformatics/btr330.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10(2). doi: 10.1093/gigascience/giab008.

Debouck, D. G., O. Toro, O. M. Paredes, W. C. Johnson, and P. Gepts. 1993. "Genetic Diversity and Ecological Distribution of Phaseolus Vulgaris (Fabaceae) in Northwestern South America." *Economic Botany* 47(4):408–23. doi: https://doi.org/10.1007/BF02907356.

Delgado-Salinas, A., T. Turley, A. Richman, and M. Lavin. 1999. "Phylogenetic Analysis of the Cultivated and Wild Species of Phaseolus (Fabaceae)." *Systematic Botany* 24(3):438–60. doi: https://doi.org/10.2307/2419699.

Desiderio, F., E. Bitocchi, E. Bellucci, D. Rau, M. Rodriguez, G. Attene, R. Papa, and L. Nanni. 2013. "Chloroplast Microsatellite Diversity in Phaseolus Vulgaris." *Frontiers in Plant Science* 3(JAN). doi: 10.3389/fpls.2012.00312.

Kami, J., V. Becerra Velasquez, D. G. Debouck, P. Gepts, and R. W. Allard. 1995. *Identification of Presumed Ancestral DNA Sequences of Phaseolin in Phaseolus Vulgaris (Molecular Evolution/Seed Protein/Crop Evolution/Tandem Repeat/Polymerase Chain Reaction) Communicated By*. Vol. 92. doi: https://doi.org/10.1073/pnas.92.4.1101.

Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura. 2018. "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms." *Molecular Biology and Evolution* 35(6):1547–49. doi: 10.1093/MOLBEV/MSY096.

Langmead, B., and S. L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9(4):357–59. doi: 10.1038/nmeth.1923.

Lavin, M., P. S. Herendeen, and M. F. Wojciechowski. 2005. "Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary." *Systematic Biology* 54(4):575–94. doi: 10.1080/10635150590947131.

Leigh, J. W., and D. Bryant. 2015. "POPART: Full-Feature Software for Haplotype Network Construction." *Methods in Ecology and Evolution* 6(9):1110–16. doi: 10.1111/2041-210X.12410.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25(16):2078–79. doi: 10.1093/bioinformatics/btp352.

Mamidi, S., M. Rossi, S. M. Moghaddam, D. Annam, R. Lee, R. Papa, and P. E. McClean. 2013. "Demographic Factors Shaped Diversity in the Two Gene Pools of Wild Common Bean Phaseolus Vulgaris L." *Heredity* 110(3):267–76. doi: 10.1038/hdy.2012.82.

Nabhan, A. R., and I. N. Sarkar. 2012. “The Impact of Taxon Sampling on Phylogenetic Inference: A Review of Two Decades of Controversy.” *Briefings in Bioinformatics* 13(1):122–34.

Naser-Khdour, S., B. Q. Minh, W. Zhang, E. A. Stone, R. Lanfear, and D. Bryant. 2019. “The Prevalence and Impact of Model Violations in Phylogenetic Analysis.” *Genome Biology and Evolution* 11(12):3341–52. doi: 10.1093/gbe/evz193.

Nock, C. J., C. M. Hardner, J. D. Montenegro, A. A. Ahmad Termizi, S. Hayashi, J. Playford, D. Edwards, and J. Batley. 2019. “Wild Origins of Macadamia Domestication Identified through Intraspecific Chloroplast Genome Sequencing.” *Frontiers in Plant Science* 10. doi: 10.3389/fpls.2019.00334.

Poplin, R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks. 2017. “Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples.” *BioRxiv*. doi: 10.1101/201178.

Rendón-Anaya, M., J. M. Montero-Vargas, S. Saburido-Álvarez, A. Vlasova, S. Capella-Gutierrez, J. J. Ordaz-Ortiz, O. M. Aguilar, R. P. Vianello-Brondani, M. Santalla, L. Delaye, T. Gabaldón, P. Gepts, R. Winkler, R. Guigó, A. Delgado-Salinas, and A. Herrera-Estrella. 2017. “Genomic History of the Origin and Domestication of Common Bean Unveils Its Closest Sister Species.” *Genome Biology* 18(1). doi: 10.1186/s13059-017-1190-6.

Rodriguez, M., D. Rau, E. Bitocchi, E. Bellucci, E. Biagetti, A. Carboni, P. Gepts, L. Nanni, R. Papa, and G. Attene. 2016. “Landscape Genetics, Adaptive Diversity and Population Structure in Phaseolus Vulgaris.” *New Phytologist* 209(4):1781–94. doi: 10.1111/nph.13713.

Rossi, M., E. Bitocchi, E. Bellucci, L. Nanni, D. Rau, G. Attene, and R. Papa. 2009. “Linkage Disequilibrium and Population Structure in Wild and Domesticated Populations of Phaseolus Vulgaris L.” *Evolutionary Applications* 2(4):504–22. doi: 10.1111/j.1752-4571.2009.00082.x.

Sand, K., Thorfinn, A. Albrechtsen, and R. Nielsen. 2014. *ANGSD: Analysis of Next Generation Sequencing Data*. BMC Bioinformatics 15, 356. Doi: https://doi.org/10.1186/s12859-014-0356-4

Schierup, M. H., and J. Hein. 2000. *Consequences of Recombination on Traditional Phylogenetic Analysis*. *Genetics*, Volume 156(2, 1): 879–891/. doi: https://doi.org/10.1093/genetics/156.2.879

Stamatakis, A. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30(9):1312–13. doi: 10.1093/bioinformatics/btu033.

Wingett, S. W., and S. Andrews. 2018. “Fastq Screen: A Tool for Multi-Genome Mapping and Quality Control.” *F1000Research* 7. doi: 10.12688/f1000research.15931.1.

Yang, Z., and B. Rannala. 2012. “Before the Advent of DNA Sequencing Technologies.” *Nature Publishing Group*. doi: 10.1038/nrg3186.

Zhang, C., S. Shan Dong, J. Yang Xu, W. Ming He, and T. Lin Yang. 2019. “PopLDdecay: A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files.” *Bioinformatics* 35(10):1786–88. doi: 10.1093/bioinformatics/bty875.