Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional
Unidad Irapuato
Unidad de Genómica Avanzada (LANGEBIO)

Firmas genómicas del proceso de domesticación del frijol común y de los eventos de introgresión dentro del género *Phaseolus*

“Genomic landmarks of common bean domestication and introgression events within the genus *Phaseolus*”

Tesis que presenta

M. en C. Martha Rosalía Rendón Anaya

Para obtener el grado de

Doctor en Ciencias

En la especialidad de

Biotecnología de Plantas

Directores de Tesis

Dr. Alfredo Herrera Estrella

Dr. Luis Delaye Arredondo

Irapuato, Guanajuato. México. Agosto, 2016.

This work was done at the Laboratory of Gene Expression and Fungal Development at LANGEBIO-UGA, Cinvestav Unidad Irapuato, under the guidance and supervision of Dr. Alfredo Herrera Estrella and Dr. Luis Delaye Arredondo. Funding for this project was provided by the National Council of Science and Technology, CONACyT, through the project J010-214-2009 granted to AHE and the doctoral fellowship number 234807 granted to MRRA from September 1st, 2011 until August 31st, 2015.

Hice una pausa. Me levanté del escritorio porque reapareció frente a tu ventana el colibrí que tanto te gustaba. Si él regresó, es imposible que no regreses tú.

Cristina Pacheco.
Mar de Historias. La Jornada. Feb, 2014.

In memory of my mother

Acknowledgments

My appreciation and gratitude for continuous support are extended to the following people that contributed to the accomplishment of this project.

Drs. Alfredo Herrera Estrella and Luis Delaye Arredondo, my co-advisors, for their guidance and encouragement along the five years we worked together. It has been a privilege learning from your expertise and being part of your research groups.

Drs. Alfonso Delgado Salinas, Luis Herrera Estrella, Octavio Martínez de la Vega and Ruairidh Sawers, members of the tutorial committee, for their valuable comments to improve the quality and impact of the project.

The PhasIbeAm consortium, which encloses research groups from Mexico, Spain, Argentina and Brazil. Such a multidisciplinary group of bioinformaticians, statisticians, evolutionary biologists and agronomists taught me integrate different tools and concepts to answer complex biological questions.

Dr. Paul Gepts, professor at the University of California, Davis, for providing biological samples and his careful revision of the manuscript derived from this work.

Araceli Fernández Cortés, for her time and excellent technical support in the use of our computational resources.

Drs. Soledad Saburido and Miguel Angel Hernández, for contributing to the analysis and biological interpretation of the results derived in the first stage of the project.

Pedro Martínez, our laboratory technician, who was in charge of the greenhouse, plant propagation and DNA/RNA extractions.

All members of the Laboratory of Genomic Services at Langebio, for their expertise in genomic technologies and the efficiency in providing high-quality genomic material for the project.

Current and former members of the Laboratory of Gene Expression and Fungal Development, for their friendship and closeness that always made me feel at home.

Alejandro Sánchez Arreguín, whose spontaneity has made me laugh but most importantly, whose friendship, loyalty and kindness have accompanied me through the best and worst moments in the past five years.

Finally, my deepest gratitude to my parents Irasema and Uriel, my sister Xochitl and my grandmother Martha; my beautiful family and my biggest source of strength, for their love and support that prove that in spite of the distance that separated us twelve years ago, we are closer than ever.

Content

Figures

Resumen

Los cultivares modernos se han derivado fundamentalmente a partir de miles de años de selección humana que ha transformado ancestros silvestres en descendientes domesticados de alto rendimiento; el frijol común (*Phaseolus vulgaris*) representa un ejemplo de este tipo de procesos evolutivos. La importancia en términos agronómicos del frijol común así como el papel de este cultivo como fuente primordial de carbohidratos y proteína en la dieta de millones de personas alrededor del mundo, han colocado a esta leguminosa como blanco de estudios genéticos que buscan determinar los loci y polimorfismos asociados, detrás del surgimiento de rasgos de domesticación y mejoramiento. En este estudio, determinamos la secuencia completa del genoma de una variedad élite Mesoamericana de frijol, BAT93, que junto con el genoma disponible de otra variedad Andina, sienta las bases para estudios genómicos a gran escala. Así mismo, integrando señales genómicas, filogenómicas y metabolómicas obtenidas a partir de la resecuenciación de 29 genomas adicionales que representan 12 especies diferentes del género, reconstruimos un modelo evolutivo que describe la separación de los linajes de frijol en el continente Americano. Observamos además sesgos en la frecuencia de eventos de introgresión genómica inter- e intra-especie, que ponen en evidencia la movilidad de genes asociados a la respuesta a estreses de tipo biótico y abiótico. Estos loci junto con los genes de domesticación aquí identificados, conforman un grupo de elementos codificantes y no codificantes de proteínas que han dado origen a muchos de los rasgos morfoagronómicos que los actuales programas de mejoramiento buscan incorporar o incrementar en las variedades cultivadas. Finalmente, nuestros resultados evidencian un particular evento de especiación ocurrido en la zona tropical Peruana-Ecuatoriana de la cordillera de los Andes, el cual precede la divergencia de las pozas genéticas Mesoamericana y Andina del frijol común y que dio origen a una especie filogenéticamente próxima a *P. vulgaris*.

Abstract

Modern cultivars have been mostly derived by thousands of years of human mediated selection, which transformed wild ancestors into high-yielding domesticated descendants; common bean (*Phaseolus vulgaris*) represents an example of such evolutionary events. The agronomic importance of common bean and its pivotal role as a major dietary component in developing countries throughout the world have placed this legume as a target for genomic analyses that aim at defining the loci and associated polymorphisms behind the emergence of domestication and improvement traits. In this study, we produced the reference genome of a *P. vulgaris* Mesoamerican elite variety, BAT93, complementing the available genomic resources for the Andean gene pool. By integrating genomic, phylogenomic and metabolomic signals from 29 resequenced genomes from 12 different *Phaseolus* species that represent most of the phylogenetic clade diversity in the genus, we reconstructed an evolutionary model of common bean lineage divergence in the the New World. We uncovered intra and inter species unbalanced introgression events evidencing the mobility of stress response genes that, together with domestication protein coding and non-coding genes herein described, have given rise to important morpho-agronomic traits to be enhanced or incorporated in modern varieties. Moreover, our results evidence a particular speciation event in the Peruvian-Ecuadorian region of tropical Andes that predates the split of Mesoamerican and Andean *P. vulgaris* gene pools.

1 Introduction

The common dry bean, *Phaseolus vulgaris* L., is the most important food legume for direct human consumption; it is a major source of calories and protein in many developing countries throughout the world (FAO: http://faostat.fao.org/) providing as much as 15% of the total daily calories and more than 30% of the protein intake per day. Even though the New World origin of the genus was established by phylogenetic studies using nuclear and chloroplast markers (Delgado-Salinas *et al.* 2006), the origin of *P. vulgaris* has been strongly debated. It was initially suggested to have taken place in the Peruvian-Ecuadorian region, given that accessions therein collected appear to have a presumably ancient form of the seed storage protein phaseolin (Kami *et al.* 1995; Kwak and Gepts, 2009). Based on the analysis of five loci, Bitocchi *et al.* 2012 proposed that common bean originated in Mexico and then invaded the Southern hemisphere, giving rise to the Peruvian-Ecuadorian population and the wild Andean gene pool, both of them phylogenetically derived from Mesoamerican clades.

In spite of the uncertainty regarding the geographic origin of *P. vulgaris*, several lines of evidence from traditional (allozymes or seed proteins) to more recent molecular markers (reviewed by McClean *et al.* 2004; Cortés *et al.* 2011; Bitocchi *et al.* 2013), converge in the establishment of at least two geographically and genetically isolated gene pools, one in Mesoamerica and one in the Andes, from which, two independent domestication events took place, starting ~8000 years ago (Kaplan *et al.* 1973; Gepts 1998; Kaplan and Lynch, 1999). Genetic diversity of wild beans however has evidenced a more complex population structure, defining up to five gene pools (Blair *et al.* 2012 [1]) corresponding to segments of the geographical range of *P. vulgaris* distribution in America, including Mesoamerican (Mexican), Guatemalan, Colombian, Central Andean (Ecuador, Northern Peru) and Southern Andean (northern Argentina, Bolivia and Southern Peru).

Domestication has independently occurred several times within the genus *Phaseolus*, resulting in at least another four clearly domesticated species: *P. dumosus* (year-long bean), *P. coccineus* (runner bean), *P. acutifolius* (tepary bean) and *P. lunatus* (lima bean). Commonly, domestication has been followed by local adaptations and further population expansions. Along these processes, not only the genetic diversity of the domesticated varieties has been compromised due to the domestication bottlenecks, but also, hybridization events between wild and domesticated populations, as suggested by morphological variation and microsatellite diversity, have occurred (Beebe *et al.* 1997; Payró *et al.* 2005; Zizumbo-Villarreal *et al.* 2005; Martínez-Castillo *et al.* 2006; Worthington *et al.* 2012), displacing the original genetic diversity in these regions but at the same time, allowing domesticated varieties to acquire adaptive traits. Genetic flow along crop evolution has been shown to be crucial for the adaptation of cultivars to different environmental conditions (Hufford *et al.* 2013), as well as for the introduction of certain morpho-agronomic traits that increase the commercial value of domesticated varieties (Tomato Genome Consortium, 2012) and thus, it should be carefully examined in the context of common bean evolution.

The availability of two sequenced *P. vulgaris* genomes of Mesoamerican (Vlasova *et al.* 2016) and Andean origin (Schmutz *et al.* 2014), set the framework for deeper analyses on the genomic and population dynamics behind the emergence of different common bean lineages. Along this study, we constructed an evolutionary model of its domestication history by sequencing the reference genome of a Mesoamerican elite variety, BAT93, and re-sequencing ten additional *P. vulgaris* accessions from Mesoamerica (MA), three from the Andes (AN), five genotypes from the Peruvian-Ecuadorian area enclosed in the Amotape-Huancabamba Zone (AHZ) in the Andes, together with eleven Mesoamerican *Phaseolus* species from the Vulgaris, Filiformis, Lunatus, Leptostachyus, Polystachios and Tuerckheimii groups. Moreover, by exploring the patterns of genomic flow we identified signals of unbalanced inter- and intra-species genomic introgression. Finally, the combination of genomic, phylogenetic and metabolomics signals, allowed us to postulate a speciation event in the Amotape-Huancabamba region, giving rise to a separate lineage that should be considered a cryptic sister species of *P. vulgaris*.

2 Background

2.1 *Phaseolus vulgaris*, the origin of the species and the need for a reference genome

For many years, the most accepted hypothesis regarding the origins of common bean indicated that, from a core area on the western Andes in northern Peru and Ecuador, wild beans were dispersed north -to Colombia, Central America, and Mexico- and south -to southern Peru, Bolivia, and Argentina- were indigenous people independently domesticated this crop during pre-Colombian times (Figure 1a; Kwak and Gepts, 2009). In this regard, radiocarbon dating and the evidence of starch grains in human teeth found at archaeological sites have placed common bean cultivation and consumption in South America, Northern Peru and Mexico between 4300 and 8000 B.P. (Kaplan and Lynch 1999; Piperno and Dillehay 2008; Mensack *et al.* 2010). The hypothesis of an Andean origin of the species relied on phylogenetic inferences using phaseolin, the major seed storage protein that shows an ancestral form (type I) contained in the Peruvian seeds (Kami *et al.* 1995). However, recent studies using other molecular markers contradict this theory and place Mesoamerica as a more probable centre of origin of the species. Using five different loci from 49 Mesoamerican, 47 Andean and 6 Peruvian wild *P. vulgaris*, Bitocchi *et al.* 2012 proposed that common bean originated in Mexico and then invaded the Southern hemisphere, giving rise to the Peruvian-Ecuadorian population and the wild Andean gene pool, both of them phylogenetically derived from Mesoamerican clades (Figure 1b).

These results are not surprising for several reasons. First, using AFLPs (Amplified Fragment Length Polymorphisms), several genomic loci, and chloroplast markers, it has been observed that there is a very low genetic diversity in wild and domesticated *Phaseolus vulgaris* of Andean origin~~,~~ (Chacón *et al.* 2005; Mensack *et al.* 2010; Mamidi *et al.* 2011), whereas genetic diversity is higher in the Mesoamerican varieties and is accompanied by a lower linkage disequilibrium (LD) estimation compared to the Andean values (Rossi *et al.* 2009). Additionally, although this genus extends from the US to Argentina, a large majority of species is found in Mexico, which suggests that *Phaseolus vulgaris* was originated in Mesoamerica by sympatric or allopatric speciation and latter, migrated to the south of the continent. So, the fact that the ancestral phaseolin type has not been found in Mesoamerican accessions might be due to a sampling limitation or because it might be extinct from these populations. Nevertheless, without whole genome analyses, the accurate determination of the centre of origin has remained challenging.

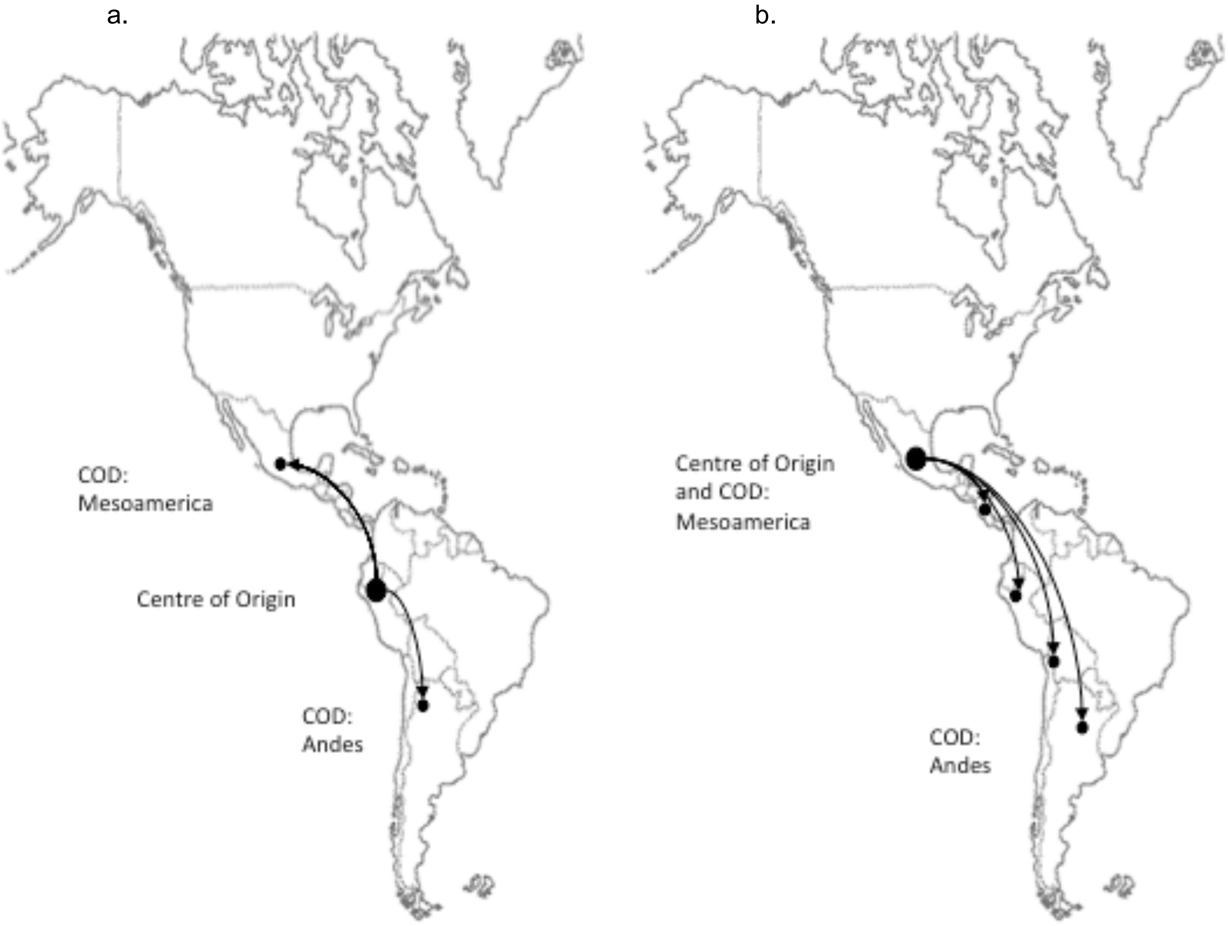


Figure 1. Centres of origin and domestication of common bean. a. Traditional view of Peruvian origin of *P. vulgaris*. b. Modern theory about Mesoamerican origin of *P. vulgaris*. Large circles represent the centres of origin; small circles the wild common bean gene pools: the arrows show dispersal routes. COD: centre of domestication.

For a long time common bean genetic resources were limited to linkage maps using reference populations that combined Mesoamerican and Andean genotypes, such as DOR364 × G19833 (Blair *et al.* 2003; Córdoba *et al.* 2010) or BAT93 × Jalo EEP558 (Grisi *et al.* 2007), that were continuously enriched with new microsatellites and SNPs. These genetic maps were useful for the identification of several QTLs associated to resistance traits (Kelly *et al.* 2003; Garzon and Blair 2014) or even the popping ability of ñuña beans (Yuste-Lisbona *et al.* 2012). However, it was until 2014 that a first reference genome was released (Schmutz *et al.* 2014), opening the possibility of deeper genomic analyses. The chosen accession corresponds to an inbred landrace line of *P. vulgaris* (G19833) derived from the Andean pool (Race Peru). The assembled genome comprised 472.5 Mb of the estimated ~587-Mb; it included 27,197 protein-coding loci (91% retained in synteny blocks with *G. max)*, and was anchored onto 11 chromosome-scale pseudomolecules (Figure 2).

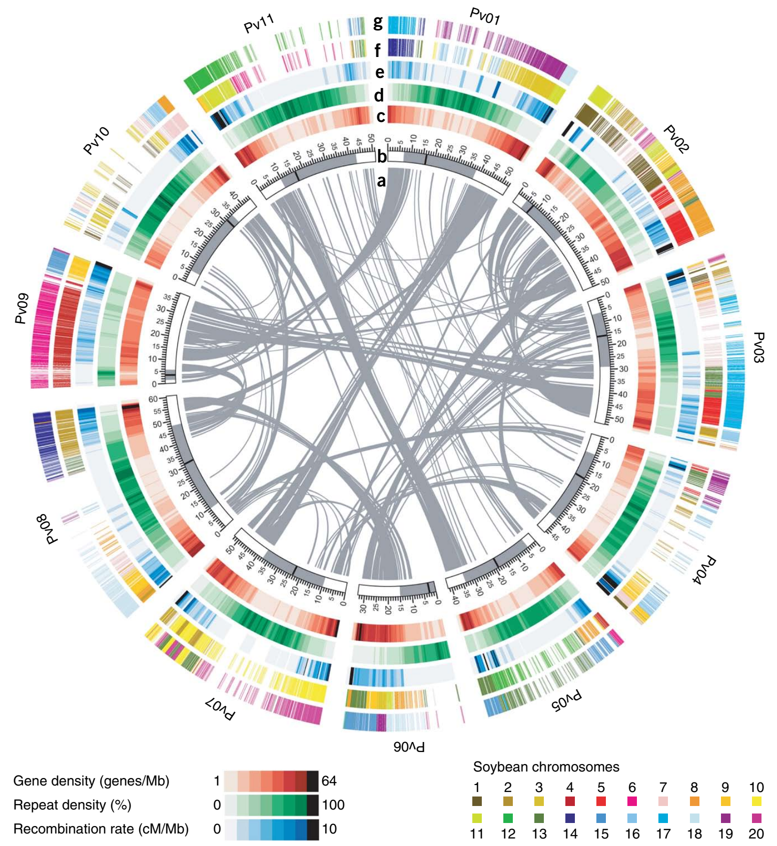


Figure 2. *P. vulgaris* genome structure and synteny with the *G. max*. a. Grey lines connect duplicated genes. b. Chromosome structure with centromeric and peri-centromeric regions in black and grey, respectively (scale is in Mb). c. Gene density in sliding windows of 1 Mb at 200 kb intervals. d. Repeat density in sliding windows of 1 Mb at 200 kb intervals. e. Recombination rate based on the genetic and physical mapping of 6,945 SNPs and SSRs. f,g. First syntenic region (f) and second *G. max* syntenic region (g) due to a lineage-specific duplication resulting in two chromosome segments for every segment in *P. vulgaris* (modified from Schmutz *et al.* 2014).

2.2 Understanding common bean domestication through genome sequencing

The study of domestication as an evolutionary model is an extremely valuable tool to identify events associated with the origin of new plant/animal species and to describe the selective pressures experienced by domesticated taxa. Agriculturalists, from prehistoric times until present, have improved their crops and livestock by choosing the best individuals as parents for the next generations. This domestication strategy is likely the most important development in the past 13,000 years of human history since it was a prerequisite to the rise of civilization (Purugganan *et al.* 2009). Different geographic areas can be distinguished as origin centres of domestication, including the Fertile Crescent, China, Mesoamerica, Andes/Amazonia, eastern US, Sahel, tropical West Africa, Ethiopia and New Guinea. Expansions of crops, livestock, people and technologies tended to occur more rapidly along east–west axes than along north–south axes since locations at the same latitude share similar climates, habitats and hence require less evolutionary change or adaptation of domesticates, technologies and cultures than do locations at different latitudes (Gepts, 2004). Some New World crops are represented by distinct but related species in North/South America and Mesoamerica, suggesting that related species were domesticated independently in these areas; this is the case of common bean *P. vulgaris*, lima bean *Phaseolus lunatus*, chilli peppers *Capsicum annuum*, squashes *Cucurbita pepo*, among other crops (Diamond, 2002).

Many morphological and physiological changes in growth habit, the lack of seed dispersal (Figure 3), dormancy or toxicity, are repeated traits in different domesticated crops and thus, have been used to define the concept of the ‘domestication syndrome’ (Gepts, 2004). The conservation and inheritance of such traits was originally based on a Mendelian strategy, and more recently, on the identification of quantitative trait loci (QTL) that represent blocks of genes that have dramatic effects on adaptation (Hancock, 2005). Therefore, domestication, considered as the outcome of a selection process that leads to increased adaptation of plant and animals to cultivation and utilization by humans, can be evaluated under a population genetics perspective (reviewed by Morell *et al.* 2011). One of the main consequences of domestication is the loss of genetic variability, compared to that observed in the wild ancestors (Gepts and Papa, 2002). This is partially explained by the reduced size of founding populations and successive bottlenecks, after which, only few allelic combinations are passed on to future generations; there is an important loss of heterozygosis and effective recombination, and thus, substantial LD can be generated. Generally, LD decays more rapidly in outcrossing species as compared to selfing ones because recombination is less effective in auto-pollinated individuals, which are more likely to be homozygous (Morell *et al.* 2011).

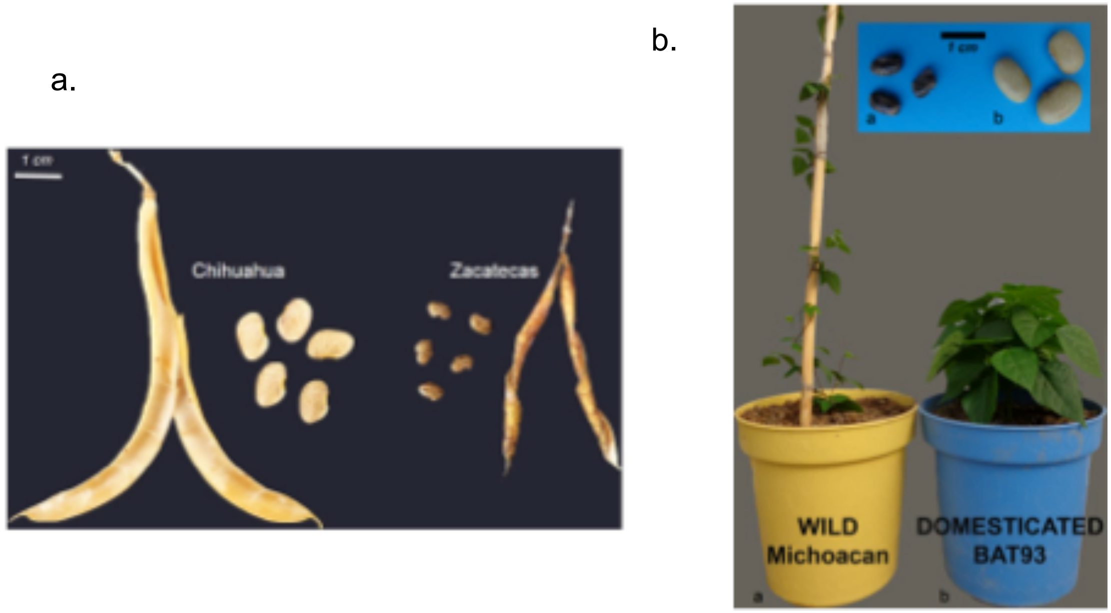


Figure 3. Domestication syndrome in common bean. a. Pod dehiscence: loss of seed dispersal is a characteristic trait in domesticated varieties (landrace from Chihuahua) compared to wild individuals (accession from Zacatecas). b. Plant architecture and seed size: climbing habit in wild plants is selected against in domesticated varieties that often grow as bushes. Domesticated cultivars produce larger seeds, with higher starch content.

A second effect of plant domestication is the modification of breeding systems: outcrossing plants are often forced to follow a self-pollinating system (Hancock, 2005). This change in the mating system produces a decrease in population sizes since lethal alleles are expressed as homozygous. Once these lethal alleles are eliminated from the population, the individual fitness increases and thus, the size of the population is balanced. At the same time, homozygosis becomes more frequent in the population and genetic diversity is greatly affected.

The first step towards the understanding of crop domestication consists in understanding parallels and contrasts between natural and artificial selection, how they have shaped genetic diversity and altered expression profiles in wild vs. domesticated populations. A useful strategy so far employed in several crops is transcriptome sequencing of wild and domesticated relatives to describe how selection on quantitative traits has affected gene expression networks. In the case of maize, the expression profiling of 18,242 genes (using an expression array) for 38 diverse maize genotypes and 24 teosinte genotypes, revealed more than 600 genes having significantly different expression levels in maize compared with teosinte. Moreover, more than 1,100 genes showed significantly altered co-expression profiles, reflective of substantial rewiring of the transcriptome since domestication (Swanson-Wagner *et al.* 2012). Although limited information on the functional consequences of the expression changes can be drawn, differentially expressed genes show a significant enrichment for genes previously identified through population genetic analyses as likely targets of selection during maize domestication and improvement. Another example is the comparison of transcriptomes from wild and domesticated cotton accessions during fiber formation, which revealed that wild cottons allocate greater resources to stress response pathways, while domestication led to reprogrammed resource allocation toward increased fiber growth, possibly through modulating stress-response networks (Yoo *et al.* 2014).

In spite of the debate regarding the origin of *P. vulgaris* as a species, several lines of evidence from traditional (allozymes or seed proteins) to more recent molecular markers [restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), AFLPs (reviewed by McClean *et al.* 2004) and single nucleotide polymorphisms (SNPs), Cortés *et al.* 2011], converge in the establishment of two geographically and genetically isolated gene pools, one in Mesoamerica and one in the Andes, from which, two independent domestication events took place starting ~8000 years ago, followed by local adaptations and further expansions. This scenario is not atypical in crops, as other plants have been domesticated more than once, offering the possibility of studying parallel evolution of independent lineages. Such an example is given by rice, *Oryza sativa*, with its two cultivated subspecies, *indica* and *japonica*, whose genomes clearly display independent origins from their wild relatives but share genomic segments bearing important agronomic traits that arose only once in one population and spread across cultivars through introgression and artificial selection (He *et al.* 2011).

Transcriptomic tools have been also used to answer intriguing points regarding the emergence of domestication traits in common bean. RNA-seq data obtained from 10 domesticated and 8 wild Mesoamerican *P. vulgaris* accessions at the first true-leaf stage, revealed that domestication not only affected the level of nucleotide diversity in about 9% of the genes, but also changed expression patterns of certain loci (Bellucci *et al.* 2014). Differentially expressed transcripts in wild accessions compared to the domesticated ones were enriched in putatively selected genes and the loss of expression diversity appeared significantly higher in selected genes compared to neutral loci. These observations could be linked to domestication but could be also explained by hitchhiking of regulatory elements.

The recent publication of a *P. vulgaris* genome of Andean origin (Schmutz *et al.* 2014), allowed for the first time to have a large scale screening of the effects of artificial selection on both gene pools. The estimation of genetic diversity losses and high differentiation values (Fst) of four resequenced pooled populations representing Mesoamerican and Andean landraces with respect to 60 wild genotypes, suggested that different sets of genes, 1835 in Mesoamerica and 748 in the Andean region, were selected for during the two independent domestication events, with only 57 of them shared by both processes (Figure 4). Even within gene pools, domestication candidates were not shared by subpopulations, suggesting that similar phenotypes in cultivated accessions were achieved following independent evolutionary trajectories. At the genomic level, 74 Mb and 60 Mb, respectively, were shown to be affected by artificial selection. Although relevant, certain aspects of this approach have to be carefully considered:

a) The fact that pooled populations were sequenced means that some biases could have been introduced in terms of over/under-representation of polymorphisms particular to certain individuals within each subpopulation.
b) Not only were the landraces sequenced in pools of DNA but they also correspond to isolated sites of collection, which means that Mesoamerican and Andean subpopulations contain a combination of genotypes that can produce artificial population structures.
c) The estimators that authors propose to identify signals of domestication (Tajima's D, Fst, π) are sensitive to population structures and are not necessarily direct indicators of the effects of artificial selection if gene flow is not considered as part of the genomic dynamics of the landraces.

The conclusions drawn by Schmutz and coworkers will be later contrasted with other strategies we propose to be more efficient in defining domestication signals. Nevertheless, an important contribution given the amount of sequenced genotypes in the above described report, allowed to make demographic inferences showing that the wild Andean gene pool diverged from the wild Mesoamerican gene pool ~165,000 years ago, with a small founding population and a strong bottleneck predating domestication that lasted ~76,000 years followed by an exponential growth phase extending to the present day (Figure 5).

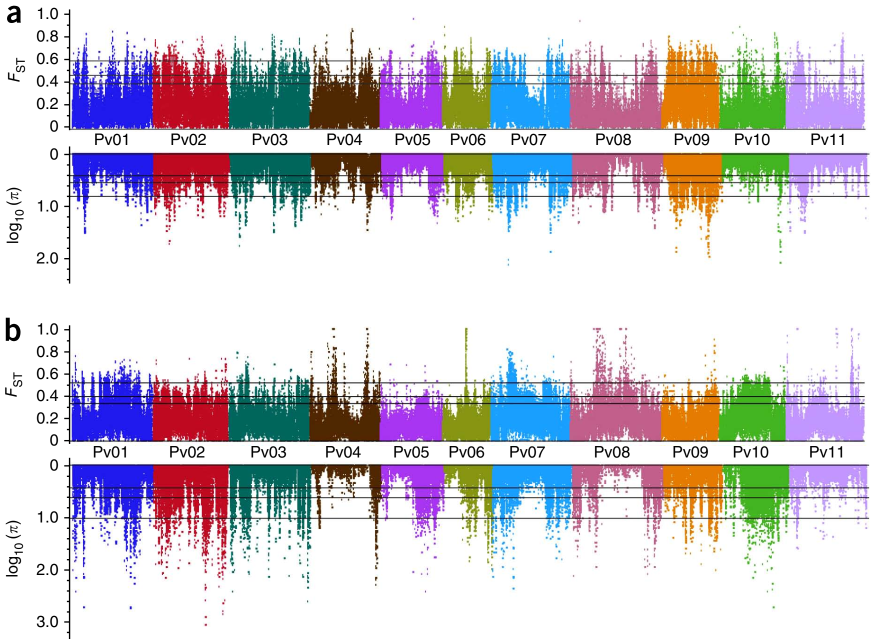


Figure 4. Differentiation and reduction in diversity during the domestication of common bean. Genome-wide view in 10-kb/2-kb sliding windows of differentiation (*F*ST) and reduction in diversity (π ratio) statistics associated with domestication within the common bean Mesoamerican (**a**) and Andean (**b**) gene pools. Log10 π ratios less than zero are not shown. Lines represent the 90%, 95% and 99% tails for the empirical distribution of each statistic (taken from Schmutz *et al.* 2014).

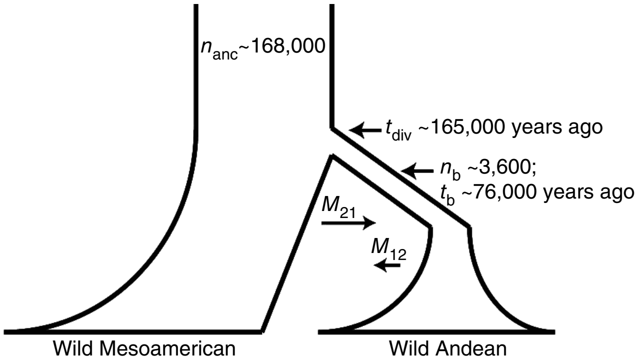


Figure 5. Divergence of the wild Mesoamerican and Andean common bean pools. *n*anc, size of ancestral population; *t*div, start of bottleneck; *n*b, size of bottleneck population; *t*b, length of bottleneck (taken from Schmutz *et al.* 2014).

Domestication processes affecting other *Phaseolus* species have attracted the attention of different research groups. It has been suggested that a single domestication event of *P. acutifolius* bean occurred in the Sonoran Desert Region of Sinaloa, since wild tepary accessions from this area were grouped with cultivated lines in distance-based trees using microsatellite sequences (Blair *et al.* 2012 [2]). Two major gene pools have been defined for lima bean, Andean and Mesoamerican, the latter subdivided in at least two groups (MI and

MII) (Andueza-Noh *et al.* 2013, Martínez-Castillo *et al.* 2014). Wild populations of the large-seeded Andean gene pool have a narrow distribution on the western slope of the Andes in Ecuador and northern Peru, while wild populations of the small-seeded Mesoamerican pool have a much broader distribution that included not only Mexico and Central America, but also the eastern slope of the Andes from Colombia to Argentina. Given the high outcrossing rate of lima beans, introgression has played a very important role in determining the level of genetic diversity of wild and domesticated populations. Just as in common bean (Papa and Gepts, 2003), gene flow is bidirectional and higher from domesticated to wild populations but highly variable when different regions are considered for sampling. This gives rise to different levels of genetic diversity, maintaining higher values in those regions where introgressions are more frequent (Martínez-Castillo *et al.* 2007; Félix *et al.* 2014). Using chloroplast markers from 262 wild and domesticated accessions of lima bean (Andueza-Noh *et al.* 2013), it was recently proposed that MI was domesticated in western central Mexico~~,~~ (Nayarit, Jalisco, Colima, Michoacán and Guerrero), while MII in Guatemala, Honduras and Costa Rica, the Mesoamerican Mayan region. On the other hand, population structure analyses suggest that domestication of runner bean could have occurred independently in two areas, Mexico and Guatemala-Honduras, followed by extensive hybridizations (Spataro *et al.* 2011).

As in common bean, until the generation of reference genomes, many questions remain open in terms of the effects of artificial selection on different genomic features and rewiring of transcriptional networks of other domesticated *Phaseolus* species. Given that domestication syndrome traits are common to most cultivated *Phaseolus* species, it is possible to imagine some degree of convergence of domestication processes into similar loci, metabolic pathways, regulatory elements and expression tuning. However, important differences in ecological niches, the degrees of availability/proximity to wild populations, reproduction habits and even human groups preferences, open the possibility of identifying alternative outcomes of domestication compared to common bean.

2.3 Hybridization, introgression and species boundaries.

A major goal of evolutionary biology is to identify evolutionary factors responsible for present-day phenotypes and species differences. In flowering plants, speciation often involves a shift in pollinator or mating system with concomitant divergence in key floral traits causing reproductive isolation between lineages. The mating system is an important determinant of the genetic variation that is maintained: outcrossing species usually show higher genetic diversity, compared to selfing species, in which heterozygosis is rapidly lost. At the same time, hybridization, the crossbreeding between individuals of different species or groups of populations (Dowling and Secor, 1997), and introgression, the transfer of genes between species mediated primarily by backcrossing, are important events that allow genetic novelties to accumulate faster than through mutation alone. The fraction of species that hybridize is variable, but on average around 10% of animal and 25% of plant species are known to hybridize with at least one other species (Mallet, 2007), even if they are distantly related (Weissmann *et al.* 2005). Hybridization can operate in different directions: reducing taxon diversity by eliminating the boundaries between species particularly if gene flow occurs into one or both parental taxa (which might facilitate adaptive evolution; Figure 6a), generating new taxa by homoploid or allopolyploid hybrid speciation (Figure 6b); and merging the two hybridizing taxa (Pastorini *et al.* 2009; Schneider *et al.* 2011). The geographic pattern and spatial scale of introgression will depend on many factors, including the environmental context in which hybridization occurs, how far individuals disperse, and the nature of selection.

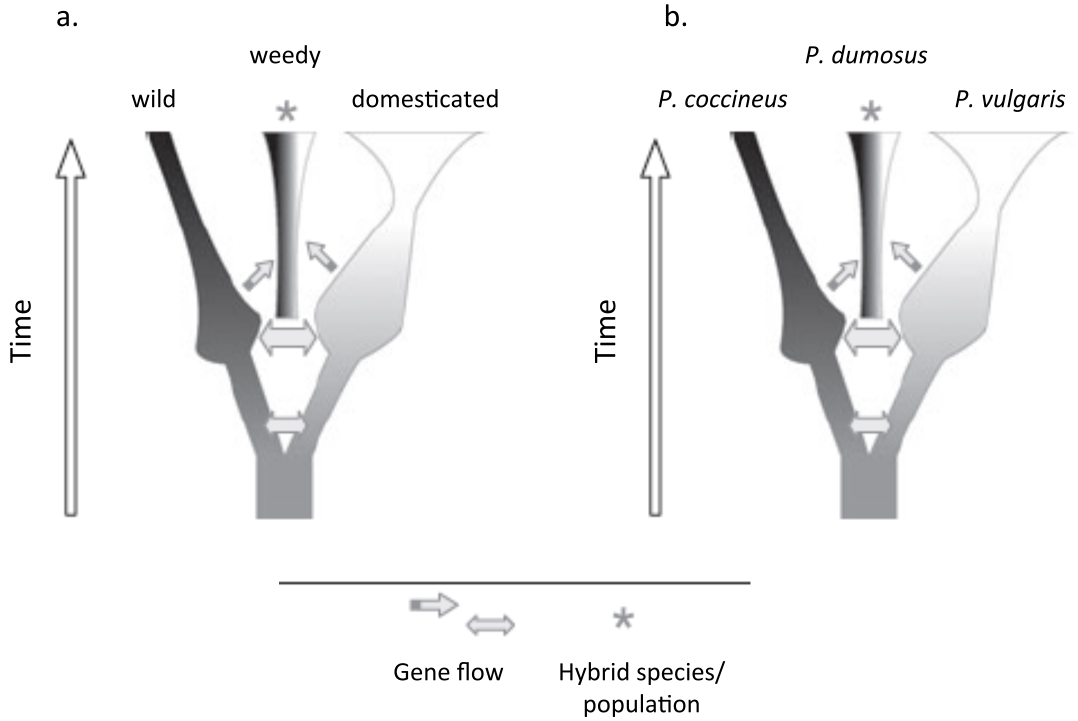


Figure 6. Hybridization and gene flow reintroduce genetic diversity. a. Hybridization inside gene pool one originates weedy populations with high levels of genetic diversity. b. Hybridization between two *Phaseolus* species produced a stable hybrid species, *P. dumosus* (modified from Abbott *et al.* 2013).

The importance of gene flow along crop evolution has been controversial. A challenging idea pointed out by several studies and revised by Harrison and Larson (2014), suggests that boundaries between species are “semi-permeable” depending on the genetic marker and that genetic isolation must be considered as a property of individual genes (or chromosome segments), not as a characteristic of entire genome. Following this theory, differential introgression documented in many hybrid zones, refers to the observation that alleles at some loci introgress more than others (Figure 7). Theoretically, globally advantageous alleles will tend to introgress easily; neutral alleles will introgress to varying extents, but linkage to genes that contribute to local adaptation or reproductive isolation will inhibit their movement. Alleles will introgress little or not at all when they represent variants at loci subject to divergent directional selection (i.e. domestication loci; Hufford *et al.* 2013; Papa *et al.* 2005) and/or loci that determine speciation phenotypes (phenotypes that are responsible for reproductive isolation).

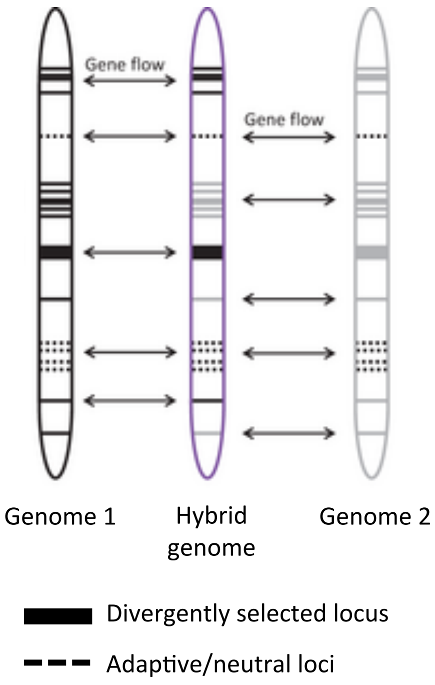


Figure 7. Genome permeability to gene flow. Divergently selected loci in two populations can be combined by recombinant hybridization. This can lead to a new species or produce adaptive introgressions in the original population. Adaptive and neutral variation can be exchanged between all populations via gene flow (Modified from Abbott *et al.* 2013).

Domesticated crops have experienced strong human-mediated selection during improvement, aimed at developing high-yielding varieties. Traditional breeding programs tend to concentrate on specific genotypes, which combine traits of interest and may be used as progenitors in several crosses. However, high-throughput SNP genotyping in crops, such as maize or wheat (Cavanagh *et al.* 2013), has evidenced small differences in terms of the amount of genetic diversity between modern cultivars and landraces, constraining our ability to expand the cultivation of domesticated species into environments beyond those in which domestication occurred, e.g., into more extreme climates, marginal soils, degraded agricultural landscapes, or into sustainable systems with reduced agricultural inputs. At the same time, subsequent to domestication, most crops spread from centres of origin into new habitats, often encountering locally adapted populations of their wild progenitors and closely related species. Usually, domesticated plants and their wild progenitor can hybridize, giving a first step towards the formation of weedy populations that combine traits of domesticated and wild types. Such hybridizations can result in adaptive introgressions, as has been documented between maize and wild teosinte (*Zea mays* ssp. *mexicana*), where the incorporation of adaptive *mexicana* alleles into maize during its expansion allowed this crop to grow in the highlands of central Mexico. More recently, whole genome scan of introgression signal was documented for cassava (*Manihot esculenta*) cultivars, whose domestication started around 6,000 years ago in the Amazonian basin. Sequencing wild (*M. esculenta* ssp. *flabellifolia*) and domesticated cassava genomes and comparing them to related species (*M. glaziovii*), not only evidenced a strong maternal bottleneck, but interspecific introgressions were shown to introduce variation into the nuclear genome, particularly in farmer varieties in Africa, were it was introduced only 500 years ago and spread by undocumented crosses (Bredeson *et al.* 2016). These crop expansions provide compelling opportunities to study evolution through introgressive hybridization.

Normally the introgression of traits from wild or weedy germplasm is difficult in modern breeding programs due to the prevalence of non-domesticated traits governed by dominant genes (Beebe *et al.* 1997). However, the fact that traditional farming systems have made of domestication a dynamic process resulting from selection, hybridization and reselection over many years, open the possibility that the variability so generated could be useful beyond the site where it occurs by continuous screenings to recover promising recombinants and introgressants that would complement modern breeding programs. Unfortunately, the use of wild relatives as a genetic resource has been taken into account from an old fashion optic, just by looking for particular phenotypes of agromorphological interest. Once a population with a desirable characteristic is identified, breeders cross them with modern varieties or cultivars, in order to introduce such traits from the wild donor. This strategy can potentially work with efficiency if the selected trait is monogenic, that is, one or only a few genes in proximity control it, such as pathogen resistance. Indeed, a survey of the use of wild germplasm in crop improvements over the last decades (Hajjar and Hodgkin, 2007), including rice, wheat, maize, barley, sorghum, millet, cassava, potato, chickpea, cowpea, lentil, soybean, bean, pigeonpea, banana and groundnut, revealed that over 80% of the reported beneficial traits conferred by genes derived from wild relatives, are involved in pest and disease resistance. Similarly, the stabilized hybrid *Helianthus annuus* ssp. *texanus* captured alleles that provide herbivore resistance from wild *H. debilis* (Whitney *et al.* 2006), and tomato cultivars introgressed several chromosomal segments from wild *S. pimpinellifolium*, enhancing fruit colour (Tomato Genome Consortium, 2012).

Several traits, however, rely on the additive action of more than one locus, epistatic interactions and by tuning gene expression by other types of regulatory elements on the genome. Finding such genes and regulatory elements is a great challenge for plant breeders. Thus, although wild germplasm is perceived to be a poor bet for the improvement of most traits based on phenotypic examination, it is quite possible that some favourable alleles are hidden in unexplored accessions. Massive genomic screenings, including SNP detection through individual genome sequencing and comparison of transcriptomic profiles and co-expression networks of wild and domesticated populations, are indispensable tools for finding those loci and construct more accurate genetic maps reflecting recombination hotspots and barrier loci for introgression. Implementing such strategies requires a major shift in the paradigm for using our genetic resources but should accelerate targeted breeding programs in the short term.

2.4 Going back to the wild inside the *Phaseolus* genus

Several *Phaseolus* species reproduce by self-pollination; however, there are examples of intermediate outcrossing in the genus. This is the case of *P. coccineus*, a species that is usually pollinated by bees and hummingbirds and *P. lunatus* that uses bees as natural pollinators. Not surprisingly, opposite to tepary and common beans, different populations of Mesoamerican *P. coccineus* sampled in central Mexico and Chiapas, display high and similar levels of genetic variation (determined with seven electrophoretic markers) without differences among wild and cultivated populations (Escalante *et al.* 1994). The same was concluded while comparing several SSRs from European and American populations of *P. coccineus* (Spataro *et al.* 2011).

In spite of its preferential autogamy, *P. vulgaris* cannot be considered as a closed reproductive system, as it maintains outcrossing rates that have been estimated between 1 and 17%, depending on the experimental protocol (Ferreira *et al.* 2006). In the case of common bean, intra-species outcrosses corresponds to a primary gene pool (GP-1; Figure 6a), however, inter-species hybridizations have also been reported within the Vulgaris clade (Figure 8). The secondary gene pool (GP-2), in which hybridization is possible but hybrids are weak with low fertility, has been observed in *P. coccineus, P. vulgaris, P. costaricensis, P. dumosus* (Blair *et al.* 2006); the tertiary gene pool (GP-3) in which embryo rescue is needed since hybrids are lethal or sterile, is possible in *P. parvifolius* and *P. acutifolius*. Even though it has been shown that no outcrossing events occur between Lunatus and Vulgaris groups, it is possible to obtain viable descendants by crossing *P. lunatus* and *P. polystachios* plants. Thus, it is possible that successive hybridizations leading to introgression events could have taken place even before *P. vulgaris* domestication, and that it has been an on-going phenomenon that occurs naturally and under human influence all along the domestication process.

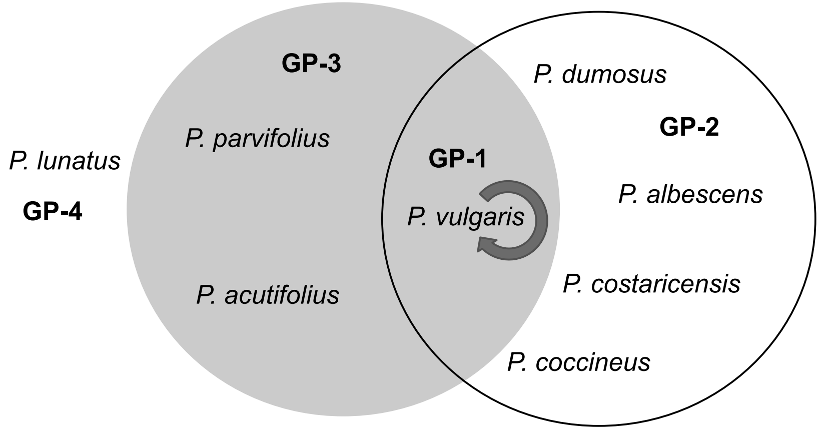


Figure 8. *Phaseolus* gene pools evaluated for their hybridization ability.

Genetic flow within the gene pool 1 (domesticated-wild) has been studied in Mesoamerica and the Andean region (Papa *et al.* 2005). It is well known that introgression does occur between wild, weedy and domesticated individuals. In different geographic zones, farmers that still maintain traditional cultivating systems usually exchange seeds and plant several different landraces in the same complexes in order to ensure some harvest, regardless of the annual growing and environmental conditions. Therefore, it is possible to maintain a high diversity that increases through spontaneous crossing among landraces. Indeed, higher molecular diversity within domesticated seeds planted under traditional cultivating systems than the one obtained in the local wild populations or the original breeding lines has been observed in several regions in Mexico, like Oaxaca (Worthington *et al.* 2012), Yucatán (Martínez-Castillo *et al.* 2006), Guanajuato and Michoacán (Payró *et al.* 2005; Zizumbo-Villarreal *et al.* 2005), and Peru and Colombia (Beebe *et al.* 1997). Furthermore, the protection of wild populations in the plots by traditional farmers can lead to hybridization of wild and domesticated populations, thereby generating weedy plants. In the same way, this protection favours backcrossing of weedy with domesticated plants and subsequently the establishment of segregants with high morphological similarity to the domesticated individuals. Measuring AFLP diversity, it has been proposed that differentiation of sympatric wild and domesticated populations is higher around domestication genes than in other loci in the genome; these observations suggest that selection in the presence of introgression is a major evolutionary factor maintaining the identity of wild and domesticated populations (Papa *et al.* 2005). Even though gene flow can occur in both directions, from domesticated to wild populations and vice versa, it has been observed that genetic introgression is three to four times more common from domesticated beans to their wild relatives than the other way around (Papa and Gepts, 2003). Taken together, these observations imply that genetic admixture and a possible mosaic genomic structure might be more frequent than expected following the preferential autogamy of the species. However, the possible mosaics need to be proven by genome sequencing.

Successful use of wild common bean relatives to introduce resistance markers into commercial varieties has been documented. In this regard, wild accessions have been used to develop varieties possessing different alleles of arcelin, which confers moderate levels of resistance to bruchids (*Acanthoscelides obtectus* and *Zabrotes subfasciatus*); cultivars resulting from crosses of elite lines (BAT93) and wild beans collected in Mexico (PI 417662) are web blight and common bacterial blight resistant, caused by *Thanatephorus cucumeris* and *Xanthomonas axonopodis* pv. *phaseoli*, respectively. Other inbred backcross populations show higher nitrogen, iron, and calcium seed content, or display higher yields than the recurrent elite parent (reviewed in Acosta-Gallegos *et al.* 2007). Efforts to increasing drought tolerance in common bean commercial varieties have been a priority for breeders, face to important and quick climate changes (Beebe *et al.* 2013), however, abiotic stress tolerance has been difficult to introduce. Given that wild *P. vulgaris* populations are distributed in a wide range of altitudes, different precipitation regimes and soil types, combining ecogeographical information, population structure, genomic and transcriptomic data could be useful for genome wide genetic associations that could accelerate the selection of wild individuals to be included in breeding programs (Cortés *et al.* 2013).

Moreover, the evaluation of morphoagronomic traits of the species belonging to each *Phaseolus* gene pool highlights the need to integrate them as genetic resources for breeding programs in the short term. Two cultivated species from GP-2, *P. coccineus* and *P. dumosus*, as well as wild *Phaseolus costaricensis,* are vigorous vines with perennial or semi-perennial tendencies. Even though three incompatibility barriers in crosses between common beans and runner beans have been identified (blocked cotyledon lethal, crinkle leaf dwarf and dwarf lethal), runner beans and year-long beans are often found in cloud forests of Central America and Mexico where climatic conditions are favourable for the development of fungal diseases such as rust, anthracnose (caused by *Colletotrichum lindemuthianum*) and web blight, and thus have been employed as sources of resistance to a wide array of bean pathogens, although their use for other traits has been very limited (reviewed by Porch *et al.* 2013). Using hydroponic systems, some accessions of *P. coccineus* were also shown to be very tolerant to aluminium-toxic acid soils (Butare *et al.* 2011). Field observations and subsequent green-house studies of root systems have revealed that runner beans have thick roots that might have a better potential to penetrate compacted soil than common beans. These traits could well contribute to drought resistance and merit further investigation. Moreover, wild populations of common bean and scarlet runner beans are often found growing together. The *P. vulgaris* × *P. coccineus* hybrid occurs naturally and can be easily made by controlled pollinations whereas reciprocal crosses have met with limited success due to unidirectional compatibility, post zygotic barriers and F1 hybrid sterility.

Tepary beans (*Phaseolus acutifolius)* are native to the desert highlands of northwest Mexico and the southwest of the USA. As such, they are extremely resistant to drought, heat and cold, and have been viewed as a potential source of drought resistance for common beans. Their roots are very long and thin, giving them the ability to penetrate soil rapidly to access limited soil water reserves (Butare *et al.* 2011). Additionally, comparative transcript profiling under water deficit of common and tepary beans revealed a very high number of responsive genes in *P. acutifolious*, some of them with functional annotations directly associated to drought tolerance (Michelletto *et al.* 2007). Despite crossing difficulties given that selection for common bean phenotype imposed by breeders eliminates much of the tepary bean introgressions during simple backcrosses, tepary beans have been used as a source of resistance for biotic constraints, especially common bacterial blight. The introduction of a novel congruity crossing method however, enhances recombination to reduce the elimination of the tepary bean large introgressions (Haghighi and Ascher, 1988), and thus, the observation of higher introgression rates estimated by AFLP sharing suggests that the use of *P. acutifolius* as a source of drought resistance alleles might be attainable (Muñoz *et al.* 2004). Tepary bean accessions have been identified with several other traits of potential value to common bean breeders including ashy stem blight and *Fusarium* wilt (*F. oxysporum)* resistance, BGYMV and bean rust resistance. Finally, lima beans (P. *lunatus)* grow over an even wider range of environments than common beans, since they are very tolerant to heat and edaphic problems. It is thus tempting to introgress traits from lima beans into common beans. However, efforts to date to cross lima beans with common beans have resulted in no more than totally sterile F1 plants.

Systematic exploration of the biodiversity of plants promises to facilitate traditional breeding and biotechnology based improvement of vegetable crops in key characteristics. In this regard, even though marker-assisted breeding programs have been successful in generating several common bean cultivars, the lack of biotechnological tools to manipulate *Phaseolus* species requires the design of more efficient strategies to incorporate a wider range of adaptations for disease resistance, abiotic stress tolerance, and other agronomic challenges, that are required in order to increase their resiliency and productivity. The identification of protein-coding genes directly affected by selection and a better understanding of how transcriptional networks are rewired following adaptation processes is needed. However, it is also necessary to explore the genomes of wild relatives that represent immediate sources of genetic innovations. Consequently, more elaborated and complementary sequencing protocols at the genomic and transcriptomic levels are required to distinguish key regulatory elements in the genomes of agronomic advantageous species that could be targeted by introgression strategies.

3 Justification

Currently, breeders' efforts revolve around the generation of new crop varieties able to grow under many different types of stress conditions, particularly resistant to drought and high temperatures. This is also true for common beans, for which traditional improvement programs are limited without a deeper understanding of the genetic diversity contained in the wild germplasm. Several studies have highlighted the importance of the genetic reservoir of wild populations, where new alleles behind adaptive traits rely. Given that, until now, it has been almost impossible to apply genetic engineering tools in common bean, the introduction of new adaptive traits hidden in the wild genetic reservoir depends on efficient hybridization strategies that result in the introgression of such loci. Normally the introgression of traits from wild or weedy germplasm is difficult due to the prevalence of non-domesticated traits governed by dominant genes. However, knowing that traditional farming systems have made of domestication a dynamic process resulting from selection, hybridization and reselection over many years, I can suggest that the variability so generated could be useful beyond the site where it occurs by continuous screenings to recover promising recombinants and introgressants.

Even though hybridization events between domesticated varieties and wild relatives growing in sympatry have been documented, no systematic screenings have been performed to accurately measure the efficiency of genomic introgression. The richness of *Phaseolus* species and wild *P. vulgaris* populations in Mexico offers the perfect scenario for such analyses that should ultimately contribute to the identification of populations prone to hybridize with cultivars that need to be adapted to new environments. Such screenings require the use of sequencing protocols, aimed at defining, in the first place, candidate genes and polymorphisms associated to domestication and improvement traits. Second, shared polymorphisms and more generally, haplotype clusters that define signals of ancestry and introgression between populations. This information should be translated, in the short term, into genetic markers and target populations that could be exploited in breeding programs to accelerate the development of new common bean varieties.

The reported screenings of selection in common bean converge to one important observation: domestication has affected, intentionally or by hitchhiking, protein coding genes and many different kinds of regulatory elements contained in intergenic segments with selection signatures that, all together, have produced the phenotypes we observe in cultivated lines. However, a more detailed description of the biological processes involved in the emergence of domestication traits requires the generation of an additional reference genome from the Mesoamerican gene pool, as well complete genomes from different populations of wild and cultivated lines.

4 Hypothesis

The geographic overlap of *Phaseolus* populations in Mesoamerica, particularly in Mexico, promotes hybridization events and the introgression of different loci that have facilitated the adaptation of cultivars to a wide range of environmental conditions. Genomic introgression signals can be distinguished from the effects of artificial selection, as domestication genes act as barrier loci for recombination.

5 Objective

Determine the rate of intra- and inter-species genomic introgression and the overlap of such signals with selective sweeps resulting from the domestication process of *P. vulgaris*.

5.1 Secondary objectives

- Produce a reference genome of a Mesoamerican *P. vulgaris* variety that complements the genomic resources available for the Andean gene pool.
- Reconstruct a phylogenomic profile of the Genus *Phaseolus* and the Vulgaris group in order to identify the closest sister clades that could be prone to hybridization events.
- Identify those genomic regions that result from introgression events between *P. vulgaris* subpopulations and between sister *Phaseolus* species.
- Compare the effects of artificial selection and genomic introgression in terms of genome structure and functional categories associated to the targeted loci.

6 Materials and Methods

6.1 Plant material.

As our reference genotype, we chose *Phaseolus vulgaris* BAT93 (Figure 9), a breeding line developed at the International Centre for Tropical Agriculture (CIAT, Cali, Colombia) and derived from a double cross involving four Mesoamerican genotypes: (Veranic x Tlalnepantla 64) x (Negro Jamapa x Tara). The biological material collected for genome resequencing included other important *P. vulgaris* accessions (Suppl. Table 1, Appendix A): eight wild Mesoamerican genotypes, selected according to their geographical distribution along the Mexican territory (Figure 10); one landrace from Chihuahua (Mexico); Jalo EEP558, a selection from the Andean landrace Jalo obtained from the Estação Experimental de Pato de Minas (Guazelli, Minas Gerais, Brazil); Faba Andecha, an Andean cultivar selected based on its domesticated traits; a wild accession from Argentina (G19901); five accessions from Peru and Ecuador considered by other authors as the ancestral form of the species because of its phaseolin isoform (PhI), all of them collected in the constrained location of the Amotape-Huancabamba deflection. Outside the *P. vulgaris* species, we selected eleven additional species covering most of the clade diversity of the genus, according to (Delgado-Salinas *et al.* 2006; Figure 11). These species correspond to the Tuerckheimii group (*P. hintonii*) and the unclassified group (*P. microcarpus*) from clade A and, from clade B, to the group Filiformis (*P. filiformis*), Lunatus (*P. lunatus* – lima bean), Polystachios (*P. polystachios* and *P. maculatus*), Leptostachyus (*P. leptostachyus*), and Vulgaris (*P. coccineus, P. dumosus, P. costaricensis* and *P. acutifolius*). Plants were grown under greenhouse conditions and young trifoliate leaves were collected for DNA extraction. For total RNA extraction, the breeding line BAT93 was growth at ±25ºC, 80% humidity, and 16h light: 8h dark photoperiod.

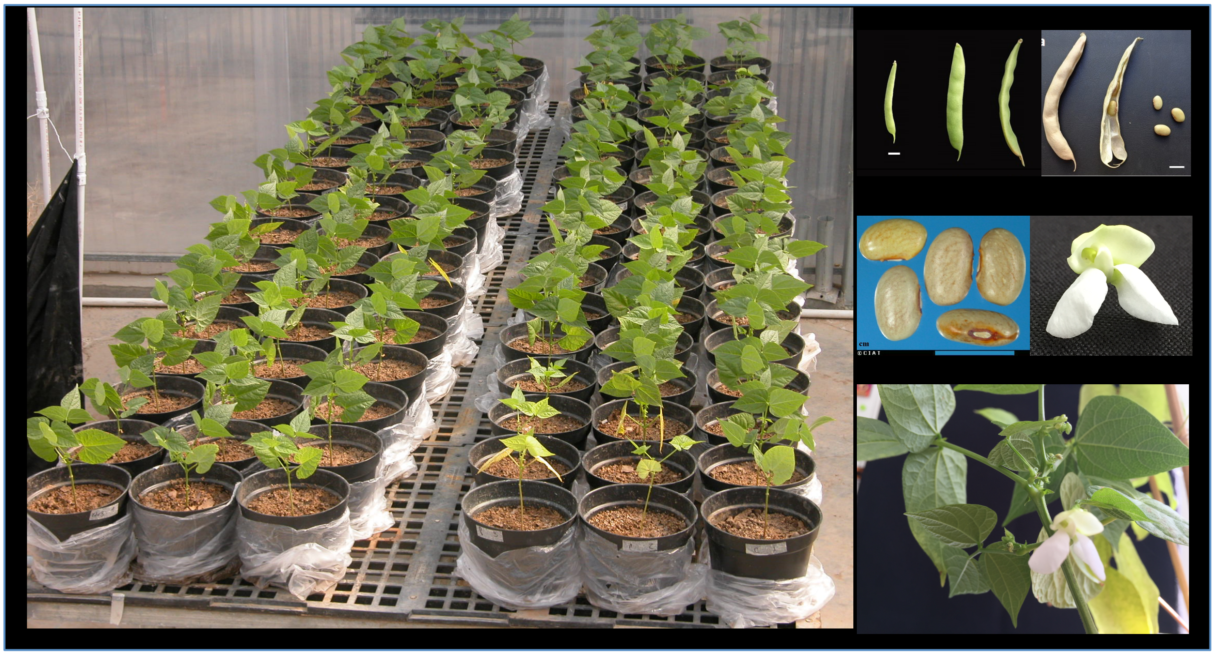


Figure 9. *P. vulgaris* BAT93 morphology.

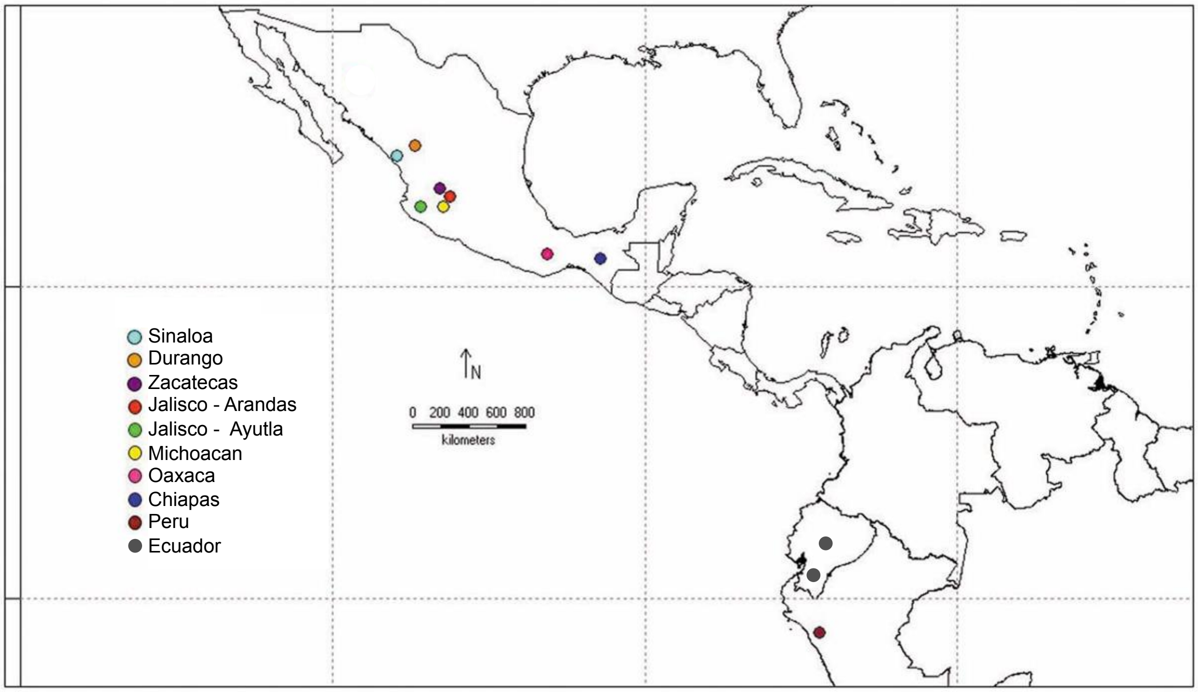


Figure 10. Geographic origin of Mesoamerican *P. vulgaris* accessions used in this work.

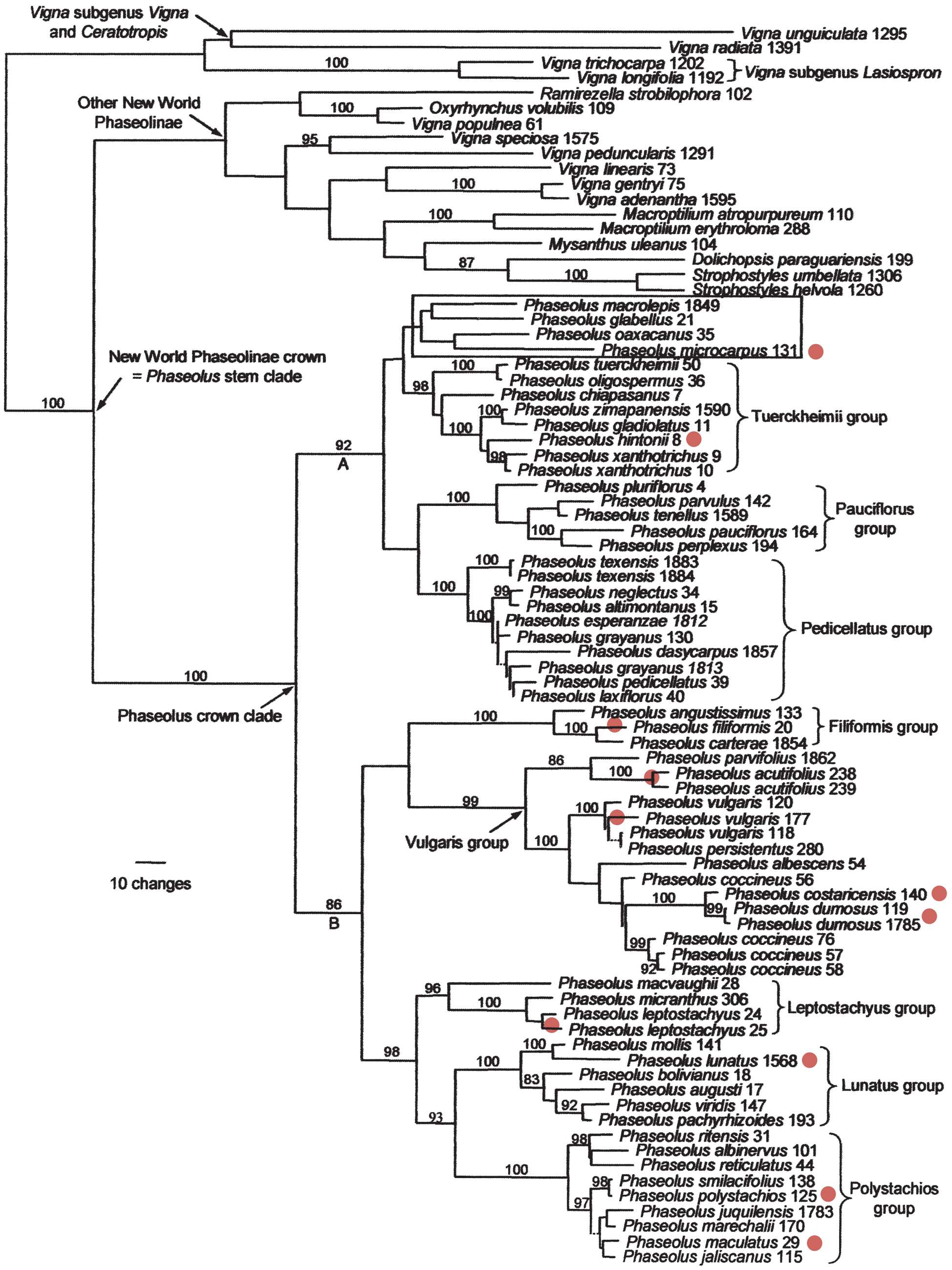


Figure 11. Maximum parsimony tree of the genus. The topology was derived from a combined analysis of trnK and ITS sequences sampled from *Phaseolus* and outgroups (taken from Delgado-Salinas *et al.* 2006). The species selected for genome sequencing are shown in red.

6.2 Genome/transcriptome sequencing and assembly.

6.2.1 Reference genome

Single-read and mate-pair libraries for BAT93 were prepared for sequencing on Roche, Illumina, SOLiD and Sanger platforms. A BAC library derived from the BAT93 line was sequenced at the Arizona Genome Institute (AGI, USA) using the automated sequencing platform ABI3730xl® (Applied Biosystems). TruSeq libraries were run on a HiSeq2000 instrument on five lanes of paired end 100 bp sequencing reads. Reference genome sequence from BAT93 was assembled based on Roche/454, SOLiD and Sanger reads using Newbler v2.6 (Roche). Assembly improvement, verification and chromosomal anchoring utilized genotyping-by-sequencing data, generated on the Illumina sequencing platform from 60 progeny of an F5 advanced intercross of BAT93/Jalo EEP558.

The available *P. vulgaris* reference obtained in this study were uploaded and locally aligned with LastZ in CoGe (http://genomevolution.org/CoGe/index.pl; Lyons *et al.* 2008). Synteny analyses and reference-guided pseudoassemblies were performed using the SynMap genomic tool (http://genomevolution.org/CoGe/SynMap.pl; Lyons *et al.* 2008).

6.2.2 BAT93 transcriptional atlas

BAT93 RNA-Seq libraries were prepared using the Illumina TrueSeq RNA-Seq library preparation protocol. Pooled sequencing of indexed libraries was performed on the Illumina HiSeq with v3 sequencing chemistry and approximately 50 million read pairs (2 x 75 nt sequencing protocol) were generated per sample. Small RNA sequencing on the same samples was carried out with non-fragmented RNA. We used the Illumina small RNA v1.5 protocol and selected inserts of size 20-100 nt. Pooled sequencing of indexed libraries on the HiSeq resulted in 7-11 million reads per sample (50 nt single reads).

Furthermore, RNA was extracted from different BAT93 samples under more than 100 biotic and abiotic stress conditions, as well as different developmental stages. Equimolar quantities of RNAs from each condition were pooled to create two normalized libraries that were sequenced using the 454-titanium platform and assembled with Newbler v2.5 (http://454.com/products/analysis-software/index.asp, default parameters).

6.2.3 Functional annotation and repeat detection.

For the de novo predictions of repeat elements, the REPET pipeline was used (Flutre *et al.* 2011). The predicted LTR retrotransposon family was further refined using the programs LTRharvest and LTRdigest (Ellinghaus *et al.* 2008; Steinbiss *et al.* 2009). The final prediction for LTR retrotransposons is the union of this procedure and REPET-based predictions. Homology-based TE identification was performed using RepeatMasker against plant-specific repeat families in RepBase v. 17.11 (Jurka *et al.* 2005). Additionally, we ran RepeatMasker v3.2.8 against plant-specific repeat families and *G. max* repeat library from RepBase to identify interspersed repeats. For the Protein-coding gene annotation RNA-Seq reads from 33 tissues were aligned with GEM to the reference genome. Cufflinks models derived from these alignments, along with isotigs assembled from a pyrosequenced normalized cDNA library and ESTs/mRNAs present in Genbank, were aligned and assembled on the genome by PASA. *Ab initio* gene prediction software [GeneID (Blanco *et al.* 2007), SGP2, AUGUSTUS (Stanke *et al.* 2006) and GlimmerHMM (Majoros *et al.* 2004)] were first trained using a set of PASA training set candidates filtered by BLAST search against nr for full-length coding sequences and then run on the reference assembly. Proteins from Uniprot were aligned to the genome. Functional annotation was performed by using in-house developed pipeline which performs an electronic inference of function that is based in the sequence similarity between the bean predicted proteins and known proteins in different public repositories: InterPro, KEGG, Reactome, SignalP, PhylomeDB and Blast2GO (Hunter *et al.* 2012; Kanehisa *et al.* 2012; Croft *et al.* 2014; Petersen *et al.* 2011; Götz *et al.* 2008).

Resistance genes were identified using the Disease Resistance Analysis and Gene Ontology (DRAGO) pipeline (Sanseverino *et al.* 2013). Gene Ontology enrichments among genes with preferential/specific expression patterns, falling within introgressed genomic windows and with domestication haplotypes were performed using the topGO package implemented in Bioconductor (Alexa and Rahnenfuhrer, 2016), using the classic Fisher's exact test with a maximum p-value of 0.05.

6.2.4 Non-coding RNA analysis.

Homology-based long non-coding RNAs were predicted taking *A. thaliana* lncRNA transcripts as templates. These were blasted against the bean assembly using RepeatMasker and the hits were then used as anchor points to realign the *A. thaliana* queries with surrounding genomic regions using exonerate as a split aligner. Final conservation was estimated on T-

Coffee (Notredame *et al.* 2000) pairwise re-alignments between the query and its predicted spliced model. *Ab initio* lncRNA models were predicted using Cufflinks and then using Cuffmerge to combine transcript models from all samples into a single set of consensus models. Sets of overlapping transcripts (>=1nt) were clustered into 1,226 gene models. LncRNA transcript expressions were obtained using the Flux Capacitor (Sammeth, 2016). For co-expression analysis we calculated Pearson correlation coefficients of all lncRNA genes with all protein-coding genes having sufficient expression across libraries.

CMsearch tool from the Infernal package (v. 1.1rc2, Nawrocki *et al.* 2009) was used to *de novo* prediction of small structured non-coding RNAs in the bean genome. We scanned the genome looking at every RNA model stored in the Rfam database (v. 11). An E-value cut-off of 0.01 allows to detect 2,529 non-overlapping hits; of these 258 are in contigs and 2,271 in scaffolds. Small RNA sequencing libraries were made after a size selection step. Reads from each small RNA libraries were independently aligned to the assembly v.10 using Bowtie2. Resulting mappings and *de novo* predicted small RNAs were used as an input to htseq-count, HTSeq v.0.6.1 to quantify small RNA features. We checked for the presence of sequences similar to rRNA by using the riboPicker tool.

6.2.5 Re-sequenced accessions.

DNA libraries were constructed and sequenced -from both ends (paired-end reads)- using the HiSeq (Illumina) technology at the Genomic Services Laboratory of LANGEBIO-CINVESTAV, Mexico. Reads of high quality (FastQC and FastxToolkit) were mapped with BWA v0.7.9a (Li *et al.* 2014) using default parameters against the *P. vulgaris* BAT93 reference genome, as well as to a synteny-based pseudoassembly produced with SynMap at CoGe (http://genomevolution.org/CoGe/SynMap.pl; Lyons *et al.* 2008) of BAT93 taking the G19833 genome as scaffold with at least four contiguous syntenic CDSs between assembled tracks determined with LastZ local alignments. This pseudoassembly was produced in order to construct longer chromosomes with more certainty of the order and sense of the scaffolds that, at their current state in the BAT93 genome version, were partially anchored into eleven linkage groups.

6.2.6 SNP calling and depth adjustment of *P. vulgaris* accessions

For each sequenced accession, individual-specific consensus sequences were generated and small variants (single nucleotide polymorphisms) were identified using the samtools mpileup implemented in ANGSD v0.614 (Korneliussen, 2014). Depth adjustments for SNP calling and consensus sequence reconstruction were done taking into account the sequencing depth of each accession: for all but 4 *P. vulgaris* accessions (Zacatecas, Oaxaca, Michoacán, Jaliso-Arandas) for which the depth threshold was set at 5 reads, a minimum of 10 reads was required. Called SNPs in positions that were covered in all accessions were considered for further analyses.

Reads with a unique hit to the reference were kept for the SNP calling step. SNPs were called using the mpileup command from SAMtools [0.1.18 (Li *et al.* 2009)], bcftools with the output from mpileup and subsequently filtered using vcfutils varFilter for the following criteria: 1) minimal depth of 5, 8 or10 (we adjusted the minimum read depth according to the coverage of each of the accessions); 2) maximal depth of 100; 3) minimum frequency of 0.3. In order to reduce false SNPs caused by misalignments, we used the option –B at the mpileup step. From the bam files, we constructed the consensus sequence of each of the accessions using the mpileup command and bcftools.

In order to normalize the heterogeneous coverage of the sequenced *P. vulgaris* genotypes and to avoid SNP enrichment at a larger coverage, we defined a novel strategy to evaluate the number of polymorphisms recovered at different read depths. From the SAM files of the accessions with the largest genome coverage that were derived after BWA mapping, we randomly picked aligned reads to approximate the coverage to ~6, 10, 16 and 20X, corresponding to the lowest and highest coverages we obtained for the *P. vulgaris* accessions. For this purpose, we used the DownsampleSam command from Picard. For each sample, SNPs were called as described before and the number of shared SNPs was evaluated. From the down sampling step it was clear that at low coverage, a read depth= 5 was equivalent in terms of the number of recovered SNPs to a read depth=10 for higher genome coverage. Additionally, we quantified the number of shared SNPs between random samples at their corresponding depths and observed that 75% of the SNPs were identical when random sampling was approximated to 10X with a read depth=5, or 20X with read depth =10. Based on these results, we were able to adjust the minimum read depth for SNP calling according to the coverage of each of the accessions.

6.2.7 Transcriptome analysis (RNA-Seq).

Reads were independently aligned to the reference *P. vulgaris* assembly v.10 using the GEMtools RNA-Seq pipeline v.1.6.2 (Griebel and Marco-Sola 2016). On average, 89±5% of the reads was mapped across samples, 69±10% of the reads mapping uniquely. Flux Capacitor v1.2.4 was used to quantify genes, transcripts, exons and splice junctions in each sample separately. To identify the preferentially expressed and organ-specific PCGs we calculated Z–scores. Differential expression was estimated with the software package edgeR (R v. 3.0.1, edgeR v. 3.2.4). For the differential expression analysis and co-expression network construction we normalized read counts into counts per million (CPM). Coefficient of variation (CV) value was used to identify putative housekeeping genes (top 10% of the genes with lowest CV). In total, we assign 2811 genes into housekeeping category.

The libraries were classified into organ groups by its phenotype: root, stem, leaf, flower, axial meristems, pods and seeds. Also we grouped the libraries by developmental stage of the plant - 5 vegetative stages: V0 (Germination), V1 (Emergence), V2 (Primary leaves), V3 (1st. trifoliate leaf), V4 (3rd. trifoliate leaf) and 5 reproductive stages: R5 (Preflowering), R6 (Flowering), R7 (Pod formation), R8 (Pod filling), R9 (Maturation). Hierarchical clustering analysis of the PCG expression profiles was performed using the hclust command in R and default complete linkage method. The Gene Ontology (GO) and enrichment analyses were performed using the Blast2GO and goseq with a FDR ≤ 0.05.

6.3 Phylogenomic profiles

6.3.1 Phylome

The database used for the phylome reconstruction contained 30,405 unique protein sequences for *P. vulgaris* BAT93. The resulting phylome comprises 27,986 gene trees, representing 92% of the predicted proteins. To build the gene trees, a Smith-Waterman search was used to retrieve homologs of each bean protein. These homologous sequences were aligned using MUSCLE v3.8 (Edgar, 2004), MAFFT v6.712b (Katoh and Toh, 2008), and KAlign v2.08 (Lassmann *et al.* 2009) and then the resulting alignments were combined using M-Coffee (Wallace *et al.* 2006) and trimmed with trimAl v1.4 (Capella-Gutierrez *et al.* 2009). Phylogenetic trees based on the maximum likelihood approach were inferred from these alignments. Maximum likelihood trees were reconstructed using the two best-fitting evolutionary models. The evolutionary models best fitting each protein family were selected using BioNJ (Gascuel, 1997) and PhyML v3 (Guindon *et al.* 2010). Orthology and paralogy relationships among *P. vulgaris* genes and those encoded by the other considered genomes were inferred using a phylogenetic approach, implemented in ETE v2 (Huerta-Cepas *et al.* 2010). The resulting orthology and paralogy predictions can be accessed through http://phylomedb.org/

6.3.2 Gene dating

Briefly, this analysis relies on a comparison in terms of homology relationships between *P. vulgaris* and 12 other plant species doing a BLAST search of all against all species to retrieve homologous sequences with a cut-off e-value of 1e-5 and a minimum coverage of 50% between query and target sequences. The considered species include Asterids (*S. lycopersicum*), Rosids (*V. vinifera, P, trichocarpa, A. thaliana, T. cacao, F. vesca, P. persica, C. melo*) and Legumes (*C. arietinum, M. trucantula, C. cajan, G. max*). Single-gene trees from BAT93 phylomes were scanned to detect and date duplication events using a previously described algorithm (Huerta-Cepas and Gabaldon, 2011). Duplications events were assigned to four different relative evolutionary periods: basal to *P. vulgaris*, basal to legumes, basal to rosids, and basal to the split of rosids and asterids. Only events including the seed protein of each gene tree were considered for downstream analyses. Finally, speciation events detected for single-gene trees in the BAT93 phylome were used to date bean proteins. The furthest orthologous sequence, according to the previously mentioned ages, was selected as the age of each seed protein. We dated 24,098 proteins (~79%) using this approach. For the remaining proteins, the relative age was assigned after detecting the most distant homologous sequence among the BLAST results.

6.3.3 Nuclear markers

From the collections of SNPs for each chromosome, singletons (unique SNPs for a particular genotype) were removed to avoid noisy signals derived from long-branch attraction effects. The filtered polymorphisms were then used to reconstruct phylogenetic trees based on the Maximum Likelihood (ML) approach. We used the best-fitting evolutionary model selected with PhyML v.3 (Guindon *et al.* 2010) and aLRT non-parametric SH branch support.

6.3.4 Chloroplast markers

A 55Kb chloroplast sequence was derived from scaffold00910 of the current BAT93 assembly, which was blasted against the available genomic sequence of the plastid from *P.*

vulgaris Negro Jamapa (Guo *et al.* 2007), displaying 99% identity. The consensus sequence of this scaffold was obtained as described above for the accessions belonging to the Vulgaris group and *P. hintonii*, as the outgroup. The 55Kb plastid tracks were aligned and cleaned with TrimAl; the corresponding tree topology was also constructed with the ML approach implemented in PhyML, using aLRT non-parametric SH branch support.

6.3.5 Coalescent simulations

In order to have a temporal frame of the divergence between AHZ genotypes and the *P. vulgaris* clade, we conducted coalescent simulations using the chloroplast sequenced fragment of 55Kb to avoid noisy signals from recombination events in the nuclear markers. We used the Bayesian approach implemented in BEAUti and BEAST v2.3.0 (Drummond *et al.* 2012), considering only 5 genotypes – BAT93 and Jalo EEP558 as the representative genotypes of the MA and Andean gene pools, respectively; one accession from Peru (G21245), *P. dumosus*, *P. costaricensis* and *P. coccineus*. An uncorrelated lognormal relaxed clock was selected using two different priors, one for the divergence between MA and AN gene pools of 165 Kya (Schmutz *et al.* 2014) and the second corresponding to the emergence of the Vulgaris group of 3 Mya (Delgado-Salinas *et al.* 2006). Based on these priors, we set the BEAUti parameters as follows:

a) MA/AN divergence of 165Kya: BAT93-Jalo $_{[dXY]=}$ 0.00026875; = d_{XY}/(2*165Kya) = 8.144e-4;

b) *P. coccineus*/*P. vulgaris* divergence of 3Mya: *P. coccineus*-*P. vulgaris* $_{[dXY]}$= 0.0068; = d_{XY}/(2*3Mya) = 0.001133;

In both cases, the monophyly of BAT93/Jalo EEP558 and *P. dumosus*/*P. costaricensis* were set a priori. The XML files were fed as input to BEAST, to perform 3 MCMC runs with 10,000,000 steps (log every 1000). Log and tree files were combined with LogCombiner; the consensus trees were obtained with TreeAnnotator and drawn with FigTree for each case.

6.4 *Phaseolus* subpopulations analyses.

6.4.1 Population genetics estimates.

To distinguish the demographic and selective effects in *P. vulgaris* genotypes during the domestication process, we performed genome-scale and a gene-by-gene screenings of neutrality deviations for which, eight sequenced accessions were clustered in two independent groups (N=4), one containing Faba Andecha and Jalo EEP558 of Andean origin, and BAT93 (Mesoamerican) together with a landrace from Chihuahua, Mexico; a second group with four wild Mesoamerican *P. vulgaris* genomes from Arandas (Jalisco), Oaxaca, Sinaloa and Zacatecas, Mexico. The genomes of both groups were aligned and divided into 50Kb sliding, overlapping windows (10Kb steps); at the gene level, we used the gene models defined for BAT93 and their coordinates in the genome to trace the full gene sequences in each of the genomic consensus of the selected *P. vulgaris* accessions.

Genes and windows were aligned with MUSCLE, and all gaps were removed with trimAl to avoid miscalculations of neutrality deviations due to non-covered positions. Pairwise differences (π), the number of segregating sites (θW) and Tajima's D (Hurst *et al.* 2009) values were calculated using the Bio::PopGen Statistics bioperl module.

6.4.2 Haplotype association.

We used the complete collection of SNPs of each *P. vulgaris* accession, including the genotypes form Peru and Ecuador as part of the wild subpopulation (19 genotypes in total) that were identified as described in section 1.3. The lists of non-unique SNPs from each chromosome were converted into tped files and then to bimbam format using Plink (Purcell *et al.* 2007). The resulting files were used as input for hapQTL (Xu and Guan, 2014), a haplotype association method that relies on a hidden Markov model, and is suitable for large data sets to infer ancestral haplotypes and their loadings at each marker for each individual. With this algorithm, the local haplotype sharing (LHS)—the probability of two diploid individuals descending from the same ancestral haplotypes and thus a natural extension of identity by descent —can be quantified using the loadings. By testing whether the genetic similarity is associated with a particular phenotype, hapQTL is able to identify associations at each (core) marker between local haplotypes and phenotypes. For all hapQTL independent runs at each chromosome, we used 2 upper-layer clusters, 2 lower-layer clusters and 20 steps in the EM runs using linear approximation; the rest of the parameters were kept as default. Two combinations of phenotypes were defined, 1) BAT93, Negro San Luis and Chihuahua labelled as "cases" of domestication in MA (DMA) and the other 16 genotypes (wild MA, AN and AHZ) as "controls"; 2) BAT93, Negro San Luis, Chihuahua, Jalo EEP558 and Faba Andecha labelled as "cases" of domestication in both COD, and the rest (wild MA, wild AN and AHZ) as "controls". For each domestication phenotype, we permuted case–control labels once and computed Bayes factors, treating these as Bayes factors under the null. Based on the permutation tests, Bayes factors (bf1 and bf2) were filtered as follows: both COD, bf1 ≥ 3 and bf2≥3.5; DMA, bf1≥3.3 and bf2≥3.

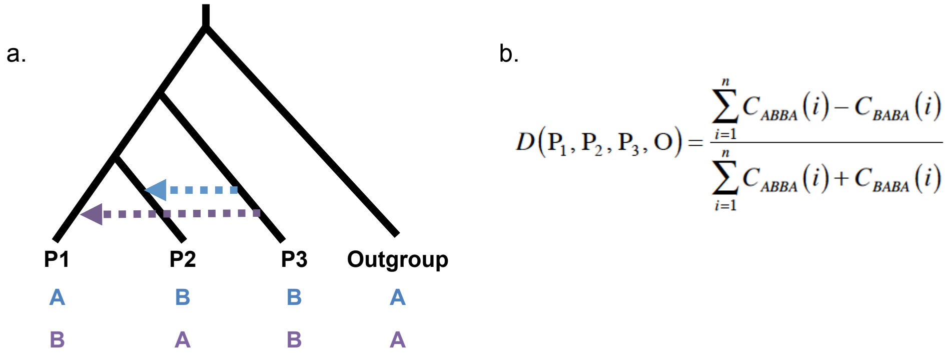
Once selected based on their Bayes factors, SNPs were evaluated with SnpEff (Cingolani *et al.* 2012) to identify those markers located in the coding sequences (exons), regulatory regions (5'/3' UTRs) or introns. We selected as domestication candidates those genes that contained at least 2 SNPs with high association factors to any domestication phenotype that were affecting regulatory regions, had non-synonymous effects on the coding sequence, altered splicing sites or stop codons.

6.4.3 Introgression analysis

The availability of statistical tests to quantify the extent of ancient admixture in the genomes of present day populations is very limited. Some of them require geographic modelling and forward-in-time population simulations (Lohmueller *et al.* 2011). However, a formal test for introgression based on the direct comparison of DNA sequences from Neanderthals and modern human populations was recently developed, exploiting the asymmetry in the frequencies of the two non-concordant gene trees in a three-population species tree (Durand *et al.* 2011). This statistical proposal represents an *ad hoc* approach for the purpose of our analysis and will be briefly described.

Let's assume we have sequenced one particular genomic region from two present-day *Phaseolus vulgaris* populations, that we denote P1 and P2; we have also sampled the orthologous region from a different *Phaseolus* species (or a wild relative), which we denote P3, and one more from an outgroup population, denoted O (Figure 12a). These four sequences will need to be accurately aligned. The null hypothesis that we wish to test is a demographic scenario in which P1 and P2 descend from a common ancestral population that diverged from the ancestors of P3 at an earlier time, without any gene flow between P3 and P1 or P2 after they split. The alternative hypothesis is that P3 exchanged genes with P1 or P2 after these two populations diverged. To test these hypotheses, we first restrict to positions in the genome where we have coverage for P1, P2, P3 and O. We denote the outgroup allele as "A", and restrict our analysis to bi-allelic sites at which P1 and P2 differ and the alternative allele "B" is seen in P3 (Durand *et al.* 2011).

For the ordered set {P1, P2, P3, O}, we call the two allelic configurations of interest "ABBA" or "BABA". The pattern ABBA refers to biallelic sites where P1 has the outgroup allele, and P2 and P3 share the derived copy (Figure 12). The pattern BABA corresponds to sites where P1 and P3 share the derived allele, and P2 has the outgroup allele. The ABBA-BABA test can then be used for each pair of *Phaseolus* populations to determine the differences in admixture rates between them, as proposed by Liang and Nielsen (2011).



$$D(\mathrm{P}_1,\mathrm{P}_2,\mathrm{P}_3,\mathrm{O})=\frac{\sum_{i=1}^{n}C_{ABBA}(i)-C_{BABA}(i)}{\sum_{i=1}^{n}C_{ABBA}(i)+C_{BABA}(i)}$$

Figure 12. ABBA-BABA basic introgression test. a. Topology representing an ABBA (blue) or BABA (purple) configuration of alleles. b. Patterson's D statistics. C_{ABBA}(i) and C_{BABA}(i) are counts of either 1 or 0, depending on whether the pattern ABBA or BABA are observed at site I in the genomic block. P_1/P_2: receptor populations; P_3: donor population; O: outgroup species.

Next, we need to implement a statistic corresponding to the difference in the counts of ABBA and BABA sites across the n base pairs for which we have data of all four samples, normalized by the total number of observations. In this statistic, C_{ABBA}(i) and C_{BABA}(i) are indicator variables; they can be 0 or 1 depending on whether an ABBA or BABA pattern is seen at base i. This statistic, denoted D, is represented in figure 9b. Under the null hypothesis, the two concordant gene trees should occur with equal frequencies, and D should equal zero. There are different classes of events that can produce a significant deviation from the null hypothesis. First, P3 exchanged genes with P1 or P2. Alternatively, P1 or P2 could have received genes from an unsampled population that needs to be at least as diverged as P3 from (P1, P2) for D to differ significantly from zero. It is important to keep in mind that gene flow between P1 and P2, or between P3 and the ancestor of P1 and P2, is not expected to produce a deviation from the null hypothesis (Durand *et al.* 2011). Quartets for ABBA-BABA were defined according to the classification of the *P. vulgaris* accessions (wild/domesticated) and the phylogenetic profile of the species. The outgroup was fixed as *P. hintonii*.

We then combined two different parameters (Martin *et al.* 2014), the dynamic estimator of the degree of introgression between subpopulations (fd) and the absolute genetic distance [d_{XY} (Eq 1a)]. In principle, as described by Martin *et al.* 2014, genomic regions that behave as f_d outliers, can be distinguished as introgressed from ancestral variation if the absolute genetic distance d_{XY} is also reduced between a donor and a receptor population, given that in the presence of gene flow, genomic windows coalesce more recently than the species split, so the magnitude of reduction in P2-P3 d_{XY} is greater than in the absence of recombination and hybridization. The f^ estimator was derived from the ABBA-BABA D statistic (Figure 12b), and it assumes unidirectional gene flow from P3 to P2 (i.e., P3 is the donor and P2 is the recipient). In the case of the dynamic estimator fd, the denominator is calculated by defining a donor population (PD) for each site independently. For each site, PD is the population (either P2 or P3) that has the higher frequency of the derived allele, thus maximizing the denominator and eliminating f estimates greater than 1 (Eq. 1b):

a.

$$d_{xy} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ix}(1-\hat{p}_{iy}) + \hat{p}_{iy}(1-\hat{p}_{ix})$$

b.

$$\hat{f}_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_{\mathrm{D}}, P_{\mathrm{D}}, O)}$$

Eq. 1. Introgression estimators. a. Absolute genetic divergence. p_x and p_y refer to the reference allele frequency in taxons *x* and *y*, respectively, in a genomic window of *n* base pairs. b. Dynamic estimator of introgression. S: the difference between sums of ABBAs and BABAs, calculated using the frequency of the derived allele at each site in each population rather than binary counts; P_D: the population (either P_2 or P_3) with the higher frequency of the derived allele and maximizes the denominator.

Based on the geographic origin of the accessions we selected for this study, we defined subpopulations of two to three individuals each, to place them as potential donors of receptors in the ABBA-BABA tests: MA-North (Durango, Sinaloa); MA-South (Oaxaca, Chiapas); MA-Central (Zacatecas, Ayutla); MA-West (Michoacán, Arandas); DMA (BAT93, Negro San Luis, Chihuahua); AN (Jalo EEP558, Faba Andecha, Wild Andean); AHZ

(Peruvian and Ecuadorian accessions). Several triads were considered to estimate the f_d parameter, permuting the receptor subpopulations and fixing *P. hintonii* as the outgroup. We defined introgressed blocks as those windows that belong to the top 5% f_d outliers that, at the same time, display d_{XY} values smaller than the average d_{XY} across the whole collection of f_d outliers. Genomic windows displaying such traits were condensed using a costume R script to define larger introgressed blocks of at least three 5kb neighbouring blocks. The parameters fd, d_{XY}, pi and D were calculated for 5Kb non-overlapping windows along the 11 linkage groups of the synteny-based pseudoassembly of BAT93, using the pipeline reported by Martin *et al.* (2014) and available at: http://datadryad.org/resource/doi:10.5061/dryad.j1rm6

6.5 Metabolomic profile

6.5.1 Sample Preparation and Extraction

Young trifoliate leaves from *P. vulgaris*, *P. pseudovulgaris* (which denotes a group of accessions from the AHZ – see further sections) and *P. coccineus* were collected and immediately frozen in nitrogen liquid. Then, the leaves were lyophilized and finely ground (<300 µm) using a Mixer Mill MM 400 (Retsch®). Subsequently, extracts were prepared mixing 50 mg of plant powder in 1,000 µL of a methanol and formic acid solution (75 % v/v and 0.15 % v/v respectively). The mixture was sonicated for 15 min in a water bath at maximum frequency and centrifuged at 10,000 g for 10 min at 4 °C. The supernatant was filtered through a 0.22 µm filter before the analysis by direct-injection electrospray mass spectrometry (DIESI-MS). All samples were prepared by triplicate and analysed immediately.

6.5.2 Mass Spectrometry Conditions

For DIESI-MS analysis, the methanolic extracts of *Phaseolus* leaves were injected directly (flow rate 10 µL·min-1) to a mass spectrometer equipped with an electrospray ionization source and a single quadrupole analyser (Micromass ZQ, Waters Corps. Mexico). Mass spectra were acquired in positive mode with the following settings: Capillary voltage 2.75 kV, cone voltage 35 V and extractor voltage 4 V. The desolvation gas was set to 400 L·h-1 at a temperature of 250 °C. The cone gas was set to 50 L·h-1, and the source temperature to 120 °C. Continuum data were acquired in a range of 50–1300 m/z during 1 min with a scan time of 10 s and an inter scan time of 0.1 s.

6.5.3 Data Analysis

Prior to data analysis, the .raw native data format of spectra was transformed to standard mass spectrometry. mzML format employing msconvert (Chambers *et al.* 2012). Then, the spectra data were processed using a workflow designed in R (http://www.rproject.org) with the package MALDIquant (Gibb and Strimmer**,** 2012) as follows: .mzML data import, summarizing all scans of each sample, smoothing by an Savitzky-Golay filter, and peak alignment/detection for comparison of peaks across different spectra. In total, 318 high quality intensity values of ions were used for statistical analysis. We employed a hierarchical clustering analysis (HCA) approach for the generation of metabolic heat maps to evaluate the differences in the fingerprinting data. To find the most important ions, we generated a Random Forest Tree model for classification in the R package 'Rattle' (Williams, 2011).

6.5.4 High-resolution mass metabolic profiling

To identify compounds from non-targeted metabolite profiling, methanolic extracts were reconstituted in a mixture of methanol/ de-ionized water/ formic acid, 75:24.85:0.15 (*v/v/v*) and analysed on an Acquity UPLC System (Waters, USA, BEH C18 2.1 x 50 mm, 1.7 um column), coupled to an orthogonal QTOF Synapt G1 (Waters, UK) high-resolution mass spectrometer. LC-MS/MS data were analysed using MS-DIAL software v2.06 (Tsugawa *et al.* 2015). Peak annotation was performed comparing fragment mass spectra with MassBank, ReSpect ESI and MS/MS libraries in positive ion mode.

7 Results

7.1 *Phaseolus vulgaris* BAT93 reference genome

7.1.1 Genome assembly

We assembled the reference genome of *P. vulgaris* BAT93 using a hybrid sequencing strategy that included 454 single reads and 8, 10, and 20 kb mate pair libraries; 3 and 5 kb SOLiD mate pair libraries; and Sanger bacterial artificial chromosome (BAC)-end and genomic read pairs. After redundancy removal, reads were assembled with Newbler and Illumina reads (45 × coverage) were used to correct homopolymer errors and close or reduce gaps within scaffolds. Illumina genotyping-by-sequencing (GBS) data from a set of 60 F5 lines of a BAT93 × Jalo EEP558 advanced intercross (6.7× coverage per line on average), together with 827 public marker sequences, were used for assembly correction and scaffold anchoring. Discontinuous genotype profiles were corrected by breaking scaffolds at the misassembly points. Markers were aligned to the assembly and GBS profiles of these scaffolds were used as seeds to place other scaffolds with the same -or similar profiles- onto chromosomes, followed by genetic map calculation. The final BAT93 genome sequence encompassed 549.6 Mb, close to previous size estimates (Arumuganathan and Earle**,** 1991), with 81 % of the assembly anchored to eleven linkage groups (Table 1). The assembly included 97 % of the conserved core eukaryotic genes (Parra *et al.* 2007), thus reflecting its completeness. In addition, we identified transposable elements by combining de novo and homology-based approaches, finding 35 % of the *P. vulgaris* BAT93 genome assembly to be covered by repeats, mostly long terminal repeats (LTRs).

7.1.2 Gene model prediction

To aid in gene prediction and to obtain a global view of the transcriptional universe of common bean plants during development, we sequenced two normalized libraries derived from 162 RNA samples from plants grown under optimal and stress conditions with the 454 pyrosequencing platform. These were assembled (Newbler v2.5) into 21,628 isogroups that include 28, 601 isotigs with an average length of 1047 bp; when compared to the genomic scaffolds of BAT93, 99.6% of the isotigs could successfully map a genomic region. Additionally, 61 RNA samples from 34 different organs and/or developmental stages from healthy plants were sequenced using Illumina technology. This information, together publicly available expressed sequence tags (EST) and cDNA sequences, were combined with *ab* *initio* predictions to produce an initial gene set of 66,634 transcripts (54,109 proteins) in 30,491 protein-coding genes (PCGs; Figure 13b).

In addition to PCGs, we identified and annotated small RNA (sRNA) and long non-coding RNA (lncRNA) sequences. *In silico* homology modelling based on sRNA sequencing led to the identification of 2529 sRNAs belonging to plant known families. LncRNAs were identified by combining *Arabidopsis thaliana* homology-based predictions and computationally predicted transcript models based on RNA-Seq data. Once filtered from single exon models, putative open reading frames (ORFs), and transcripts mapped within 1 kb of annotated PCGs (Ørom *et al.* 2010), we obtained 1033 intergenic lncRNAs (38 inferred from *A. thaliana*), coding for 1858 transcripts.

Table 1. Summary of *P. vulgaris* cv. BAT93 genome assembly

	Whole genome	Scaffolds only
Assembly		
Total length	549,604,264	494,957,111
Number of scaffolds/contigs	68,379	9,047
N50(size/number)	433,759 / 324	526,483 / 267
N90(size/number)	2,023 / 8,894	35,958 / 1,484
Range (min-max)	500-3,177,954	2,000-3,177,954
% of Ns	34.96 %	36.99 %
G + C content	38.43 %	36.64 %
Annotation		
Number of protein coding (PC) genes	30,491	29,569
Number of PC transcripts	66,634	65,685
Number of small RNAs	2,529	2,271
Number of long non-coding genes	1,033	870
G + C content exonic (for PC genes)	47.57 %	47.70 %
Number of functionally annotated transcripts	62,713 (94.12 %)	62,594 (95.2 %)

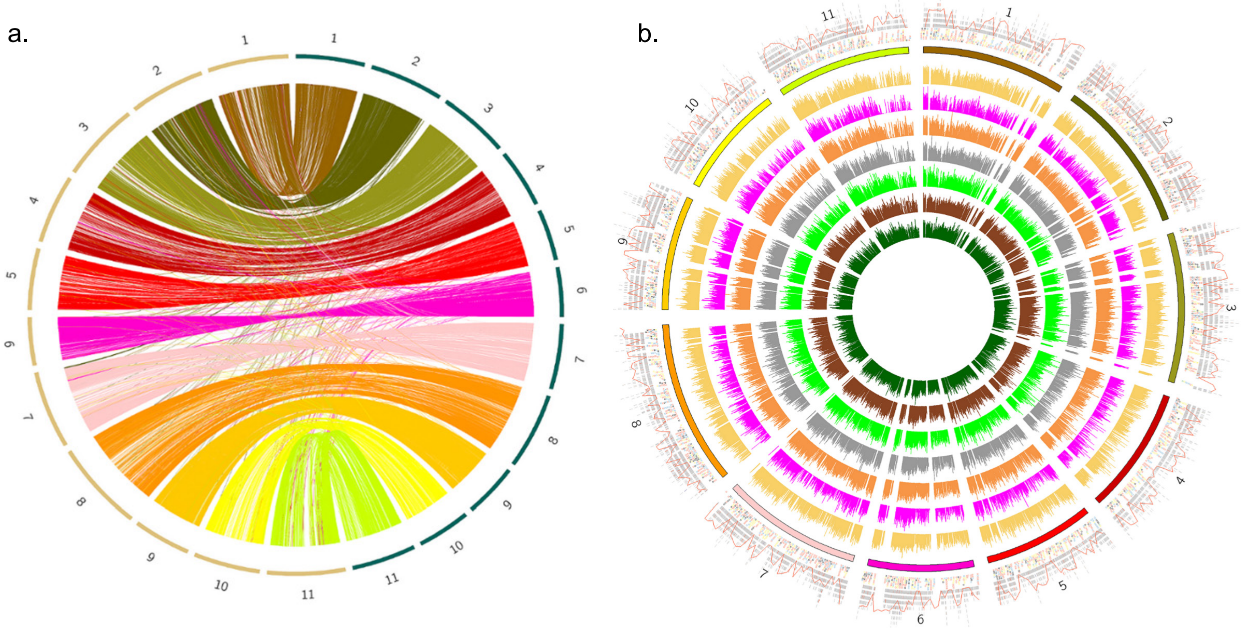


Figure 13. Overview of BAT93 genome assembly and transcriptome coverage. a. Synteny-like comparison of one-to-one orthologs between BAT93 (green) and G19833 (brown) linkage groups. b. Circos plot representing the gene content and transcriptome maps of the linkage groups of *P. vulgaris*. The outer ring represents the localization of genes across bean linkage groups. Grey regions are meant to contain genes and white regions depleted from annotated genes. The red line shows the repeat coverage across the linkage groups. Below, squares of different colours represent different types of genes: red, smallRNAs; blue, lncRNAs; yellow, legume-specific; black, resistance. The inner rings below the horizontal bar delineating the linkage groups represent RNA-Seq coverage for the different organs: axial meristem, flower, pod, seed, leaf, root and stem (modified from Vlasova *et al.* 2014).

7.1.3 Genome comparison of the Mesoamerican and Andean accessions

Given the availability of a full genome sequence of an Andean variety of *P. vulgaris* (G19883), we compared the assembly of BAT93 against the linkage groups deposited in Phytozome. Out of the 25,991 BAT93 PCGs that could be placed in linkage groups, 20,617 were uniquely mapped to 20,618 PCGs in the Andean genome (Figure 13a). A Syntenic Path Assembly (SPA) was performed at CoGe using a LastZ nucleotide-nucleotide search, for which the minimum number of aligned pairs was defined as 4 genes (Figure 14). This indicates that 4 is the minimum number of gene-pairs that DAGChainer needs in a collinear gene set to keep a syntenic block. Choosing this number of gene pairs produced a pseudo-assembly of the BAT93 genome containing 723 scaffolds, in which 17,776 PHASIBEAM PCGs are encoded. This pseudo-assembly was used for downstream analyses, given the importance of having long DNA tracks for introgression tests, as will be discussed in subsequent sections.

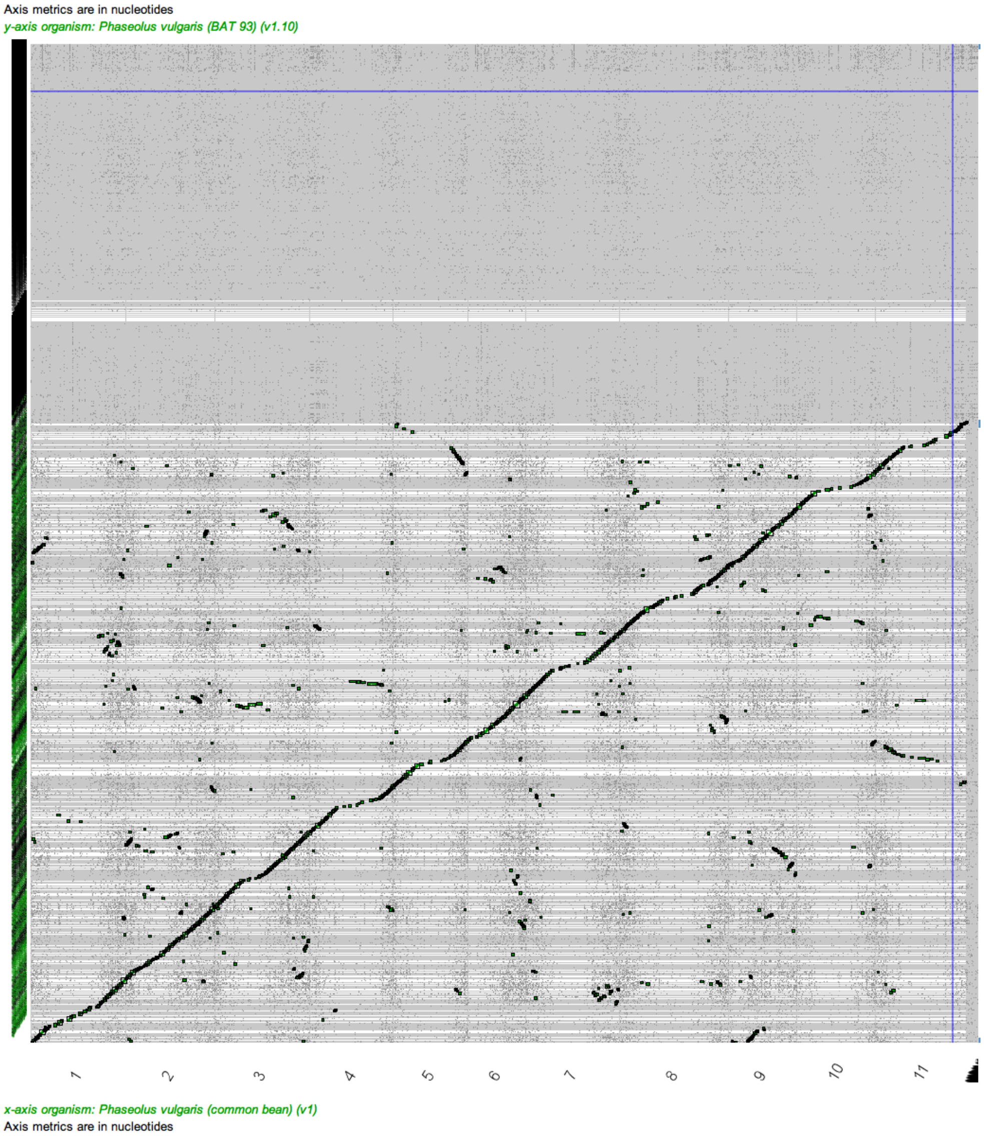


Figure 14. SPA of BAT93 (y-axis) against the eleven linkage groups (x-axis) of the Andean *P. vulgaris* G19833 variety deposited in Phytozome.

In addition, we compared our PCG model predictions with that of the Andean *P. vulgaris* G19883 genome using a combination of synteny and phylogeny-based orthology assignment between both genomes. Out of the 25,991 BAT93 PCGs that could be placed in linkage groups, 20,617 were uniquely mapped to 20,618 PCGs in the Andean genome (Figure 13a). When considering both placed and unplaced PCGs, 21,600 BAT93 PCGs were mapped to 21,604 PCGs in the G19833 genome. We then aligned the protein coding sequences of these equivalent genes and found that 1186 PCG pairs have sequence identity lower than 95 %

when gaps are not considered. Additionally, when BAT93 Illumina reads were mapped to the G19833 assembly we identified 10,193 regions of 1 kb or longer with zero coverage containing a total of 314 PCGs. We found 94 % of the lncRNAs in the Mesoamerican genome were also present in the Andean genome. Homology profiling against 12 other complete plant genomes revealed 526 bean-specific lncRNA genes and only five lncRNAs conserved in all twelve plant genomes.

7.1.4 Functional annotation

Functional annotation based on sequence similarity between bean predicted proteins and known proteins in different public repositories revealed that 62,713 (94.12%) transcripts and 26,635 (87.35 %) genes had some type of functional annotation.

Given that the Mesoamerican BAT93 line has been described as less susceptible to diseases such as bean common mosaic virus rust, angular leaf spot, anthracnose or common bacterial blight compared with its Andean counterpart, we also characterized 852 resistance genes. These include 234 belonging to the cytoplasmic NBS-LRR class. In comparison, G19833 had been predicted to contain 376 cytoplasmic NBS-LRR class genes, of which 316 could be mapped to 220 BAT93 genes. Interestingly, we were unable to find resistance-gene clusters that were specific to either of the two varieties, indicating that the genomic clustering of resistance genes predates the split of both gene pools and suggest that the differences in pathogen susceptibility might be due to polymorphisms in these loci, rather than a gene presence–absence effect.

7.1.5 Phylome

To gain insight into *P. vulgaris* genome evolution, we reconstructed its phylome, i.e., the complete collection of evolutionary histories of bean genes, using PCG sets derived from either BAT93, G19833 or both genomes. We obtained 27,986 trees for the BAT93 phylome (available through PhylomeDB; Huerta-Cepas *et al.* 2014) and scanned them to detect and date gene duplication events, delineate orthology and paralogy relationships, and annotate functions. We reconstructed a species phylogeny using two complementary approaches: i) the analysis of 172 sets of widespread groups of one-to-one orthologs, and (ii) a super-tree reconstruction using 82,365 single-gene trees from the three phylomes above. Both approaches yielded an identical topology (Figure 15.), which provides an evolutionary framework for downstream comparative genomics analyses.

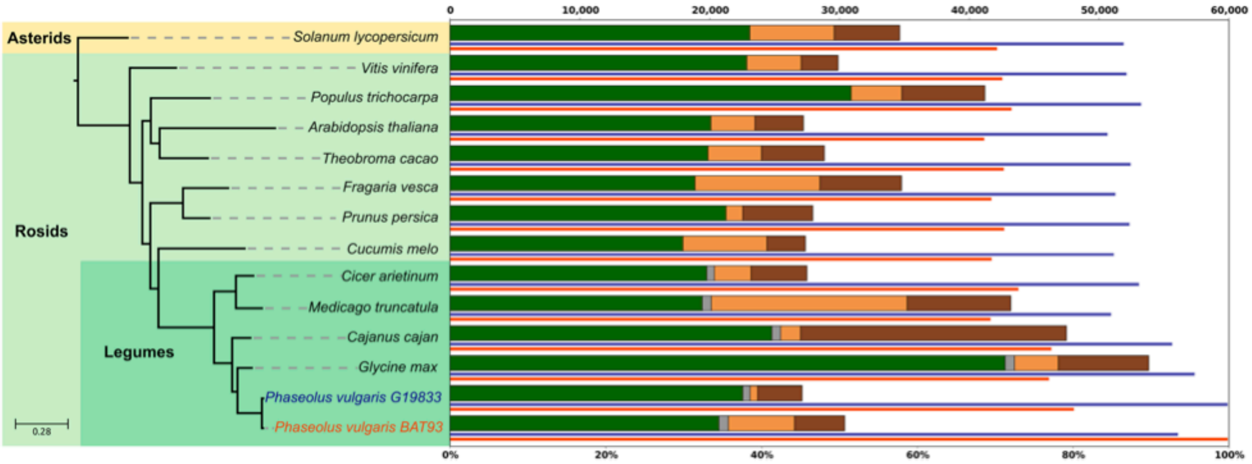


Figure 15. Phylogenomics analysis. The species phylogeny is based on maximum-likelihood analyses of a concatenated alignment of 172 widespread, single-copy orthologous genes. The two different *P. vulgaris* accessions used in this phylogeny are colored differently. Bars represent the total number of genes for each species (scale on the top) and are divided to indicate different types of phylogenetic profiles: green, widespread proteins which are found in at least 12 of the 14 species; grey, widespread but legume-specific proteins which are found in at least four of the six legumes species; light-orange, genes without a clear phylogenetic profile; brown, species-specific genes with no (detectable) homologs in other species. The thin blue line under each bar represents the percentage of *P. vulgaris* G19833 genes that have homologs in a given species. Conversely, the thin orange line represents the percentage of *P. vulgaris* BAT93 genes which have homologs in a given species (Vlasova *et al.* 2014).

From this phylogeny we defined four evolutionary periods as the lineages preceding the divergence of *Phaseolus*: basal to *Phaseolus*; basal to legumes; basal to rosids; and basal to the split of rosids and asterids. We then assigned the duplications inferred from gene trees to each of these periods. The resulting pattern of duplication densities is consistent with the proposed wave of whole genome duplication events at the split of rosids and asterids (Jiao *et al.* 2011) and at the base of legumes (Yang *et al.* 2015; Cannon *et al.* 2006). However, in contrast to what has been observed in soybean (Schmutz *et al.* 2010), we found no footprints that recent whole genome duplication occurred in any of the two sequenced *P. vulgaris* lineages. We assessed functional enrichment among genes restricted to specific clades or specifically duplicated in the lineages described above. The largest gene family expansion specific to BAT93 corresponded to putative cellular receptors with extracellular domains. We found two additional BAT93-specific expansions that were functionally enriched in seed development and the ubiquitin pathway. We found several gene family expansions common to BAT93 and G19833 in which the gene tree topologies suggested that duplications preceded the divergence of the two lineages. These duplications are enriched in genes involved in defence and stress response. Genes widespread in legumes but absent from other species were enriched for functions related to symbiosis with soil microorganisms and pathogen response. Interestingly, functions related to response to nematodes, which often parasitize leguminous plants, and regulatory response to auxin and oxygen were enriched among families duplicated at the base of legumes.

7.1.6 Transcriptomic atlas of BAT93

We used RNA-Seq libraries from 27 organs/developmental stages for which we have technical replicates (7 of the 34 conditions only had one sample) to generate a gene expression atlas across organs and during plant development. Libraries were classified into seven organs (root, leaf, seed, pod, stem, flower and axial meristem) and into developmental stages (V0–R9, expanding from 48 hours to 86 days) (Fernández *et al.* 1986; García Mendoza**,** 2009). Hierarchical clustering of the samples based on PCG expression recapitulates tissue types, the main separation being between the root and aerial samples (Figure 16). At a threshold of gene expression of 1 RPKM, we identified 20,525 (67 %) PCGs, and 521 (52 %) lncRNAs expressed in at least one organ, and 12,261 (40 %) PCGs and 99 (10 %) lncRNAs were expressed in all organs. On average, we detected 64 % of PCGs and 28 % of lncRNAs expressed per organ.

PCGs were putatively classified as house-keeping genes when they were within the top 10 % of the expressed genes with lowest coefficient of variation across all samples**.** This resulted in 2811 genes that, according to a GO enrichment analysis, mostly carry out functions related to fundamental cell processes. Similarly, we identified a core set of 25 lncRNA genes that are both ubiquitously expressed in all organs and evolutionarily conserved in at least seven of the twelve species used for comparative analysis and thus may play crucial roles similar to those played by housekeeping PCGs. In general, highly conserved lncRNAs tend to have a higher level of expression.

We performed differential gene expression analysis for PCGs and lncRNAs across all pairs of samples, both in individual samples as well as in sets of samples grouped into organs and developmental stages. We found that 937 (4%) PCGs and 171 (17%) lncRNAs had organ-specific expression. About half (84) of the latter are fruit-specific, in contrast with organ-specific PCGs, which are enriched in roots (32 % of organ-specific PCGs are root-specific; Figure 17).

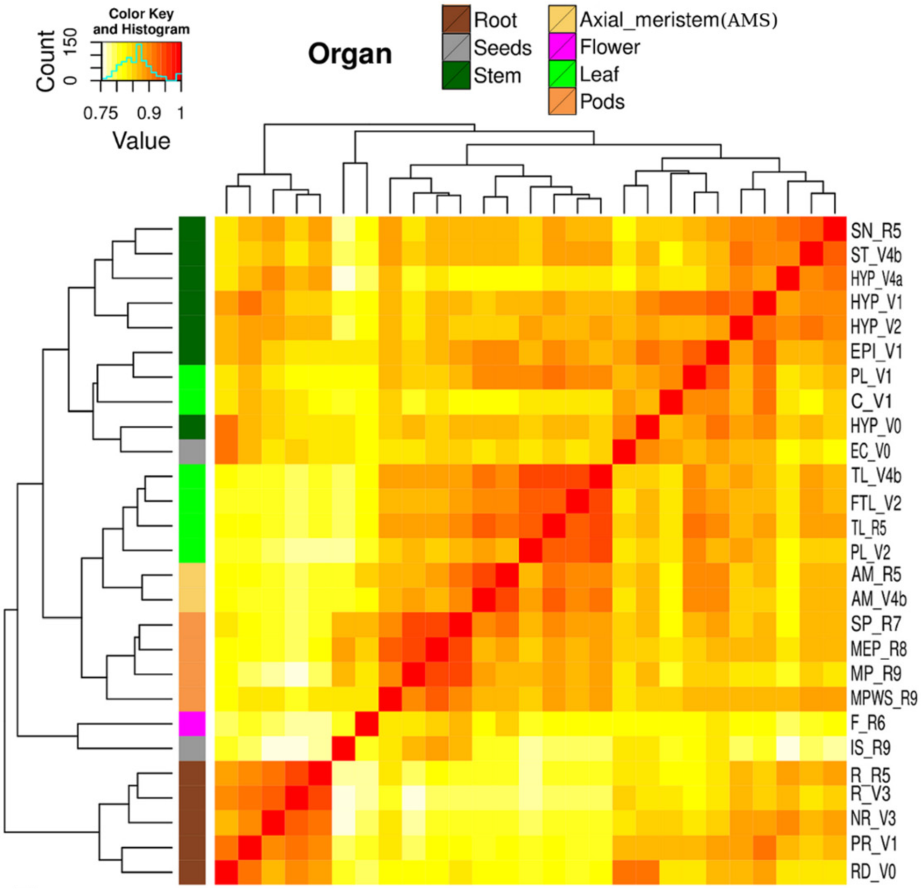


Figure 16. Hierarchical clustering of bean samples based on expression levels of PCG.

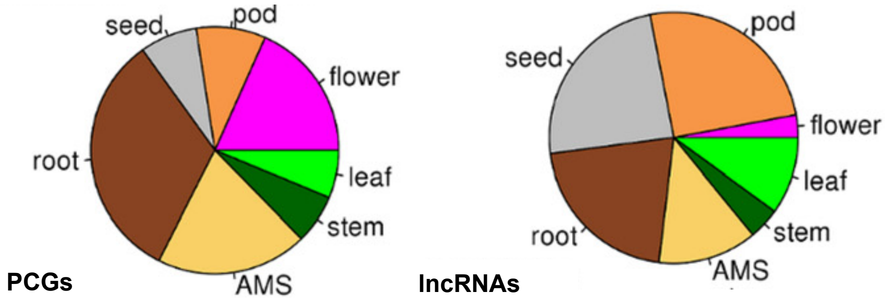


Figure 17. Distribution of organ-specific PCG and lncRNAs across organs.

We also compared gene expression in each stage of plant development with the previous stage globally, as well as independently in each of the four organs where we had sufficient numbers of samples at different stages: root, leaf, stem and pooled flower/pod/seeds, referred here to as fruits. Overall, a larger number of transcriptional changes occur during the vegetative as compared with the reproductive stage for both PCGs and lncRNAs (Figure 18). For instance, during the establishment of primary leaves, over 1000 genes are differentially expressed, including 20 lncRNAs, while this number drops to less than 120 when comparing leaves during the later stages. We found similar numbers of differentially expressed genes during root, leaf and stem development (2165, 2220 and 2859, respectively), and a larger number (4869) during fruit formation. The functions enriched in genes that are differentially expressed between different stages in each organ are consistent with the physiological changes associated with the development of that organ.

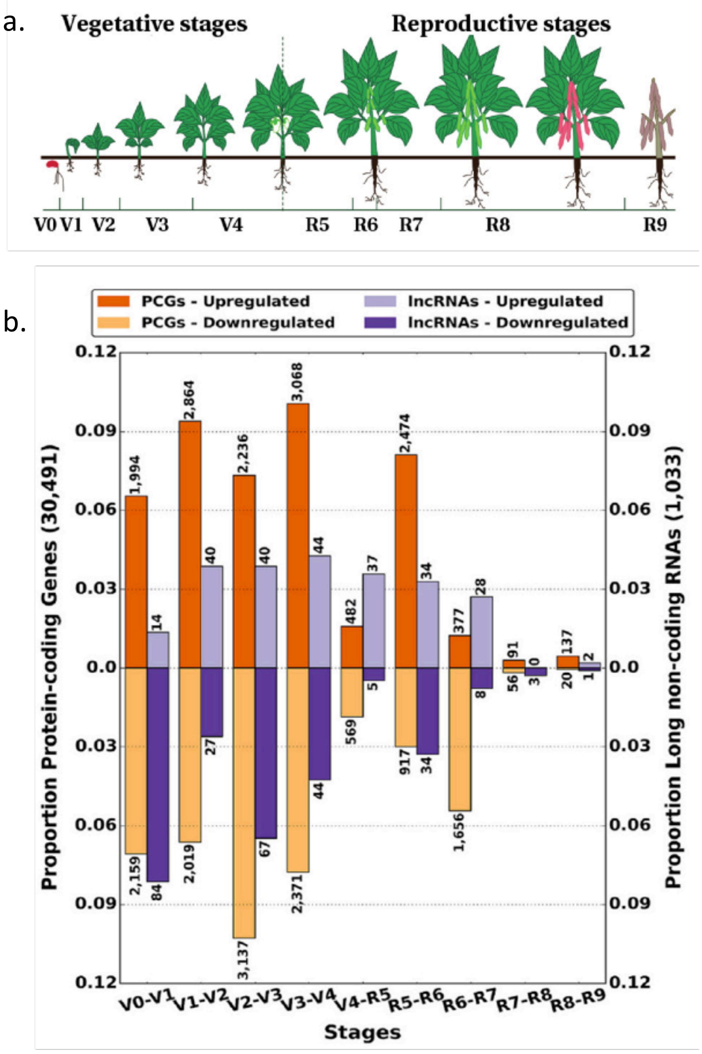


Figure 18. Transcriptome dynamics. **a.** Development stages of the common bean. Modified with permission from the technical guide for the bean growing by the "Instituto Interamericano de Cooperación para la Agricultura" (IICA; Garcia Mendoza**,** 2009). **b**. Differential PCG and lncRNA expression during development. Each bar corresponds to the number of genes differentially expressed in a given developmental stage compared with the previous one. Values above and below zero indicate the proportion of up-regulated and down-regulated genes, respectively; the number of regulated genes is shown at the tip of the corresponding bar.

7.2 Early speciation in the Vulgaris group

7.2.1 Phylogenetic profile of 30 *Phaseolus* genomes

Thirty *Phaseolus* genomes, selected to represent most of the species diversity in the genus, were re-sequenced using the Illumina HiSeq platform at different coverages, ranging from 8X to 20X (Suppl. Table 2, Appendix A). According to the phylogenetic classification proposed by Delgado-Salinas and co-workers (Delgado-Salinas *et al.* 2006), which divides *Phaseolus* species in two sister clades, our sampling covers one out of the three well defined groups from clade A (Tuerckheimii) and has at least one representative species covering all groups from clade B (where all domestication events have taken place and has a broader distribution in America), with an intended bias to the Vulgaris group. Raw reads were filtered and mapped against the *P. vulgaris* BAT93 reference genome, as well as to a synteny-based pseudoassembly of BAT93 using the G19833 genome as scaffold. From genome sequencing of the thirty bean accessions and sister species, we identified ~6,735,000 SNPs that were analysed by two approaches**,** i) reconstruction of their main evolutionary relationships and ii) assessment of genetic exchange between accessions. To avoid potential artefacts such as long-branch attraction, SNPs that were unique for any given accession or species were removed. The reconstructed phylogenetic profile of the species using a random sampling of over 460,000 SNPs, was consistent with the previously proposed phylogeny (Delgado-Salinas *et al.* 2006), in which *P. lunatus*, *polystachios* and *maculatus* belong to a tight group of species with incipient domestication; *P. microcarpus* and *P. hintonii*, all from clade A are more distant species and were used to root the tree. The Vulgaris group, while consistent in terms of phylogenetic proximity of sister species such as *P. acutifolius, P. coccineus, P. dumosus* and *P. costaricensis*, highlighted a novel pattern. In contrast with previous reports in which wild accessions from Peru and Ecuador formed a clade derived from MA wild subpopulations (Bitocchi *et al.* 2012), our tree topology placed them as a separate clade, sister to all *P. vulgaris* genotypes (Figure 19).

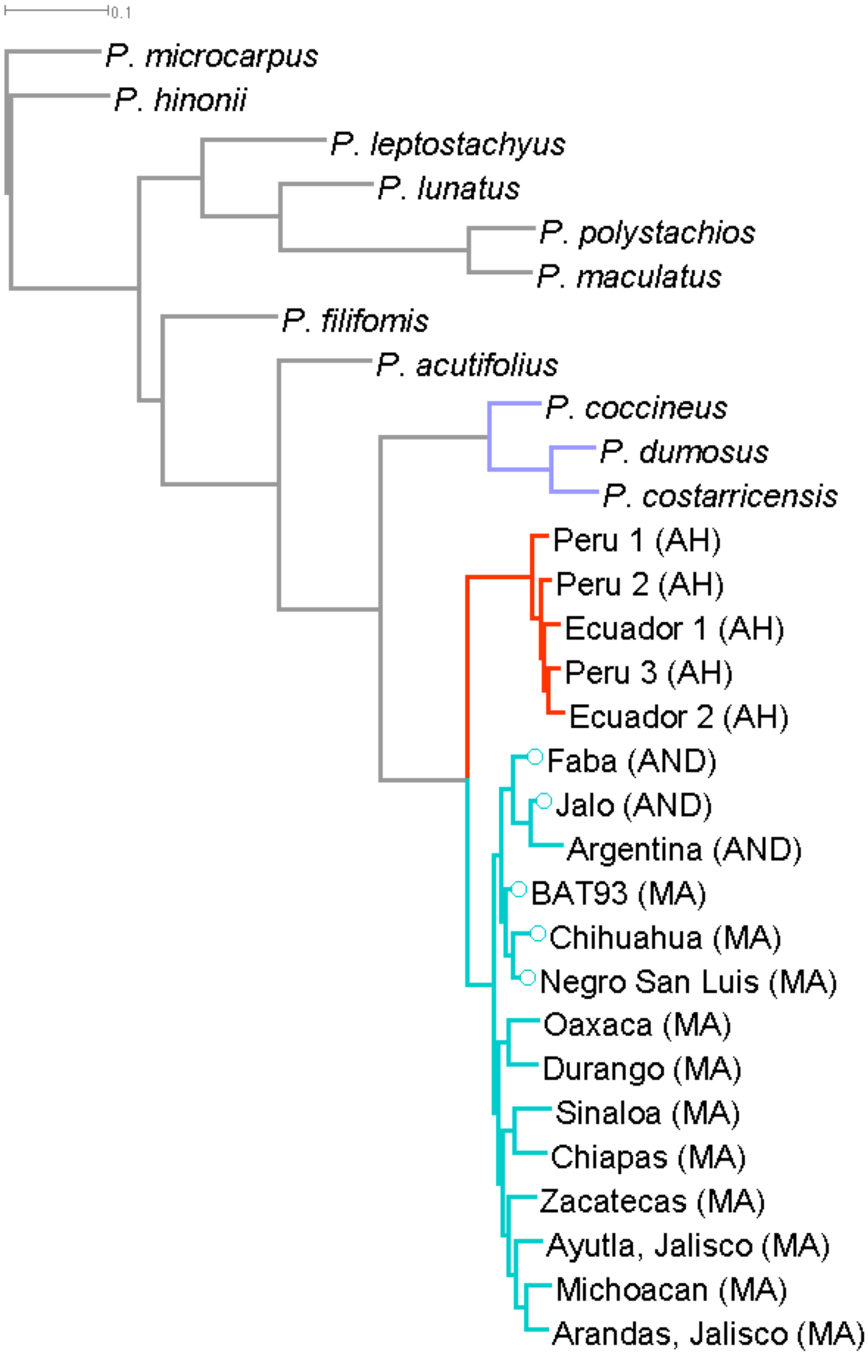
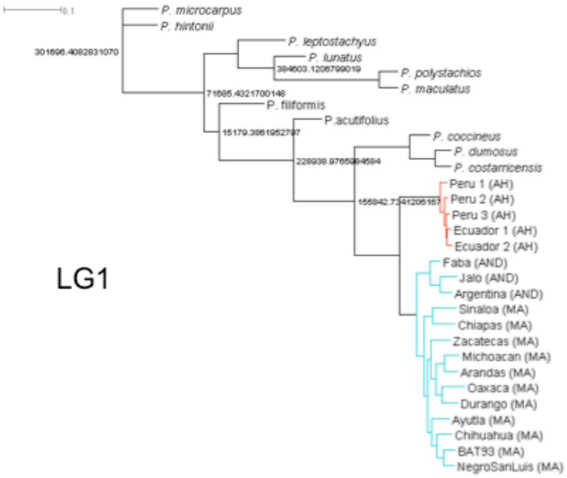


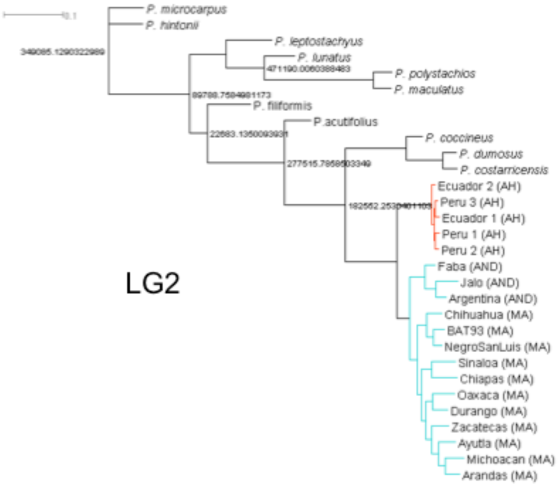
Figure 19. Phylogenetic tree of *Phaseolus* species. 460,000 non-unique SNPs were concatenated to produce this ML tree. Species belonging to the Vulgaris group are highlighted in colours; circles indicate domesticated genotypes. All clades have aLRT non-parametric SH branch support >0.7.

Moreover, non-unique SNPs from each linkage group were concatenated and aligned to produce independent topologies that could reflect the evolutionary history of each *Phaseolus* chromosome. As depicted in figure 20, the phylogenetic trees that were generated, follow the same topology at basal branches as the one in figure 19, but displayed different grouping of the *P. vulgaris* accessions that can only be explained by recombination events between subpopulations that have taken place all along the Mesoamerican area.

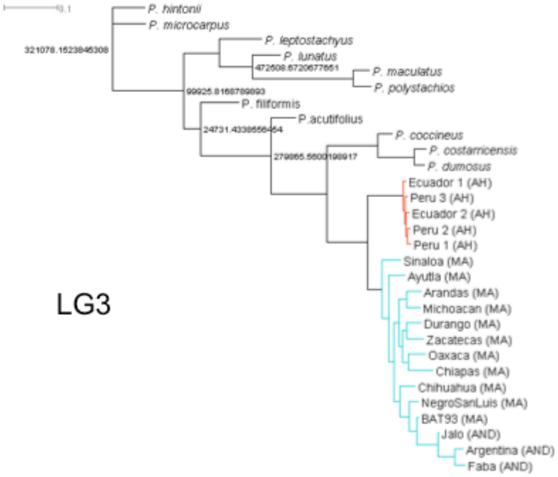
One of the still unanswered questions revolving around *P. vulgaris* is its actual centre of origin. As mentioned above, for many years, people studying this group of legumes suggested its origin to be in the area of Peru and Ecuador. However, as discussed in section 2.1, more recent molecular evidence has pointed Mesoamerica as the most feasible centre of origin for *P. vulgaris*. The idea of Peru and Ecuador as the geographic point from common bean radiated, calls again our attention, as in all the phylogenies obtained the SNPs collected from the genomes of the accessions G21244, G21245, G23587, G23724 and G23582, place them at the root of the vulgaris accessions. In contrast to the rearrangements inside the vulgaris clade, the Peruvian accessions remain at the same place, no matter which chromosome is taken into account (Figure 20). From this approach, it is no possible to claim that Peru-Ecuador are actually the centre of origin of *P. vulgaris*; we cannot claim either that these genomes correspond to the ancestral forms of *P. vulgaris*; what we can conclude from these observations, is that the genotypes extracted from this area have diverged enough to be consistently maintained outside the *P. vulgaris* clades, as what we can tag as a "sister species". This idea will be further discussed and approached using different genomic and metabolomics tools.



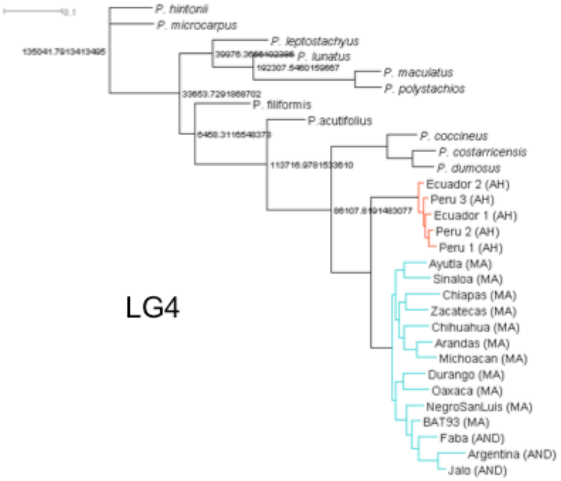
LG1



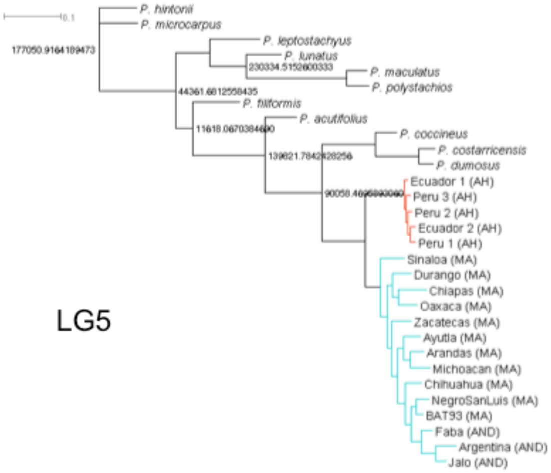
LG2



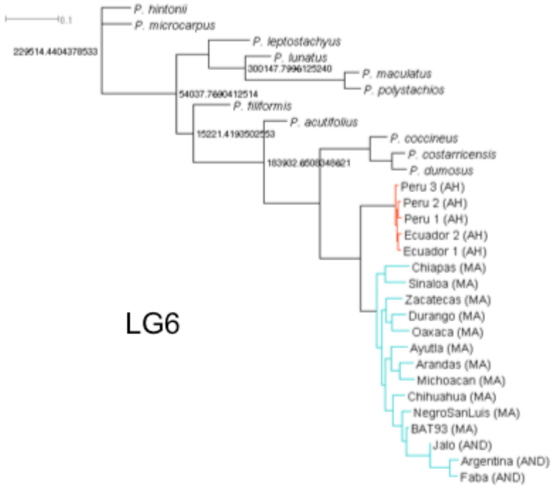
LG3



LG4



LG5



LG6

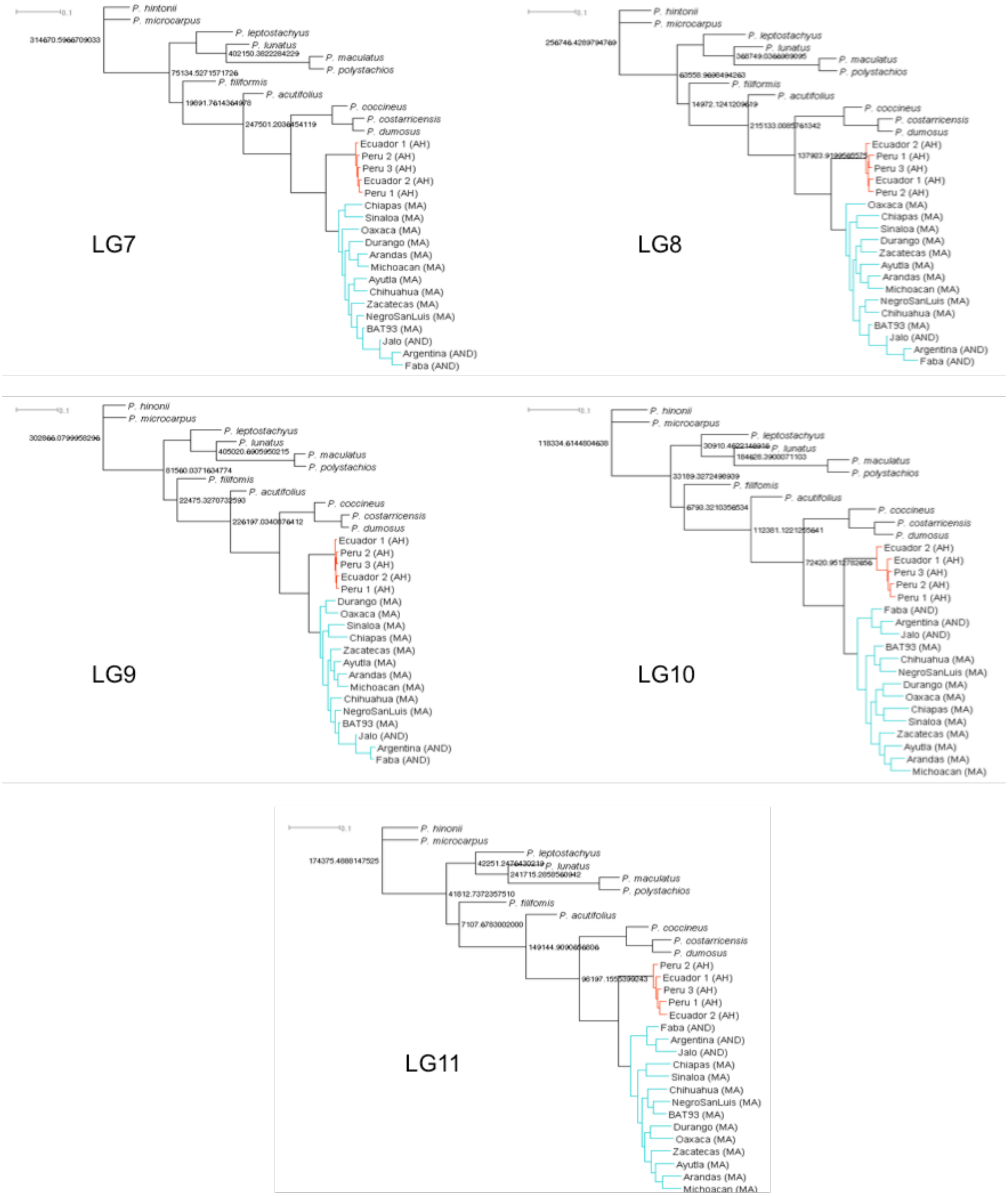


Figure 20. Tree topologies of each linkage group. SNPs for each chromosome were aligned and ML phylogenies reconstructed using aLRT implemented in PhyML.

In order to confirm this evolutionary relationship, we constructed a maximum likelihood topology using one third (55kb) of the chloroplast genome. The phylogenetic signal of the plastid genome was consistent with the one observed with nuclear markers (Figure 21), which implies a divergence of the AHZ genotypes and the *P. vulgaris* lineage that precedes the split of the Mesoamerican and Andean gene pools.

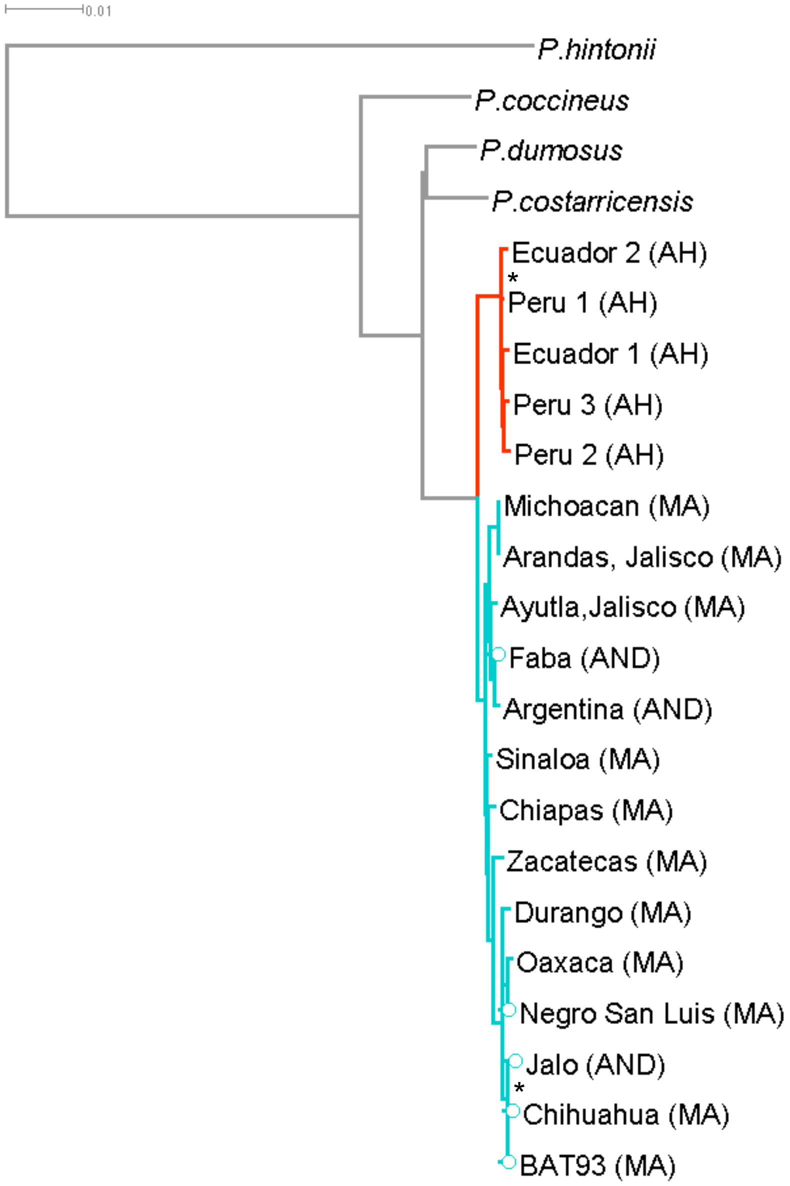


Figure 21. ML tree using 55Kb of the chloroplast genome. Clades with (*) have aLRT non-parametric SH branch support <0.7.

7.2.2 Genetic divergence discriminates between *Phaseolus* species

Based on the collection sites of the accessions, we constructed subpopulations of *P. vulgaris* genotypes that became useful for further calculations that require allele frequencies instead of binary polymorphisms. These subpopulations are referred here to as: WMA-North (Sinaloa; Durango); WMA-West (Michoacan; Jalisco); WMA-Centre (Zacatecas; Jalisco); WMA-South (Chiapas; Oaxaca) – the four of them composed of wild genotypes; DMA (BAT93; Negro San Luis; Chihuahua) –domesticated genotypes of Mesoamerican origin; AN (Jalo EEP558, Faba Andecha, G19901)- a group composed by two domesticated genotypes and a wild accession from Argentina; AHZ (Peru: G21244, G21245, G23587; Ecuador: G23724, G23582) – wild accessions from the Amotape-Huancabamba Deflection. A simple but highly informative parameter we were able to calculate with these subpopulations, was the pairwise absolute genetic divergence (d_{XY}) according to (Smith and Kronforst, 2013) (Figure 22). A remarkable but, to a certain extent expected observation, was the difference between intra and inter-species d_{XY} values. We observed average distances below 0.009 for *P. vulgaris* MA intra-species subpopulations and, consistent with the phylogenetic signal, the d_{XY} value for inter-species comparisons (*P. vulgaris* against its sister species *P. coccineus, P. dumosus* and *P. costaricensis*) ranged between 0.026-0.03. However, given the *P. vulgaris* intra-species d_{XY} range, two contrasting results called our attention. First, accessions from the AH Zone belong to very restricted populations, both at the geographic and genetic levels, as they are the least divergent accessions from all our comparisons (d_{XY}=0.0023). Second, the AHZ subpopulation and any other *P. vulgaris* group are equally divergent (d_{XY}≈0.014) as are the two well defined sister species, *P. dumosus* and *P. costaricensis* (d_{XY}=0.011). Not only d_{XY} values within *P. vulgaris* subpopulations and between *P. vulgaris* and AHZ accessions are different (Kruskal-Wallis p-value=0.014), but the comparison of inter and intra-species d_{XY} values as depicted in Fig 1c indicates they are all derived from independent populations (Kruskall Wallis test, H=14,0784, 4 d.f., *P*=0.007). These observations provide further support to our phylogenomic results, indicating that the AHZ group should be considered an independent evolutionary unit within the Vulgaris group, that given its morphological resemblance of *P. vulgaris* we suggest to be denoted as “*Phaseolus pseudovulgaris*”.

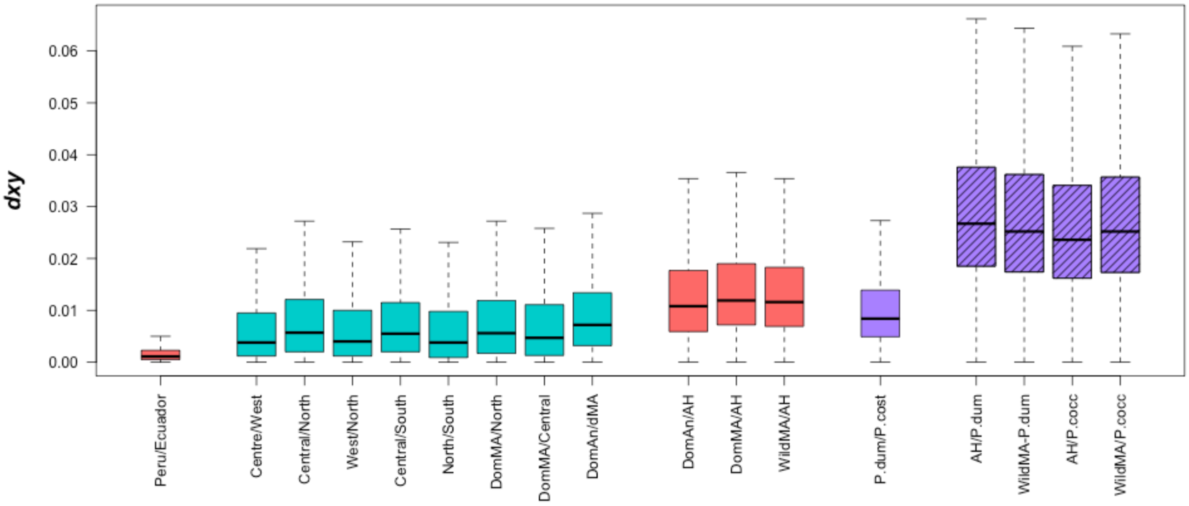


Figure 22. Absolute genetic divergence inside the Vulgaris group.

7.2.3 Dating of *Phaseolus* lineage divergence

In order to have a temporal frame of the divergence between AHZ genotypes and the *P. vulgaris* clade, we conducted coalescent simulations using the chloroplast sequenced-fragment of 55Kb to avoid noisy signals from recombination events in the nuclear markers. Using two different priors, one for the divergence between MA and AN gene pools of 165 Kya (Schmutz *et al.* 2014) and the second corresponding to the emergence of the Vulgaris group of 3 Mya (Delgado-Salinas *et al.* 2006), we corroborated an early splitting of the AHZ lineage that occurred around 700 Kya – 1Mya, previous to the separation of the MA and AN gene pools (Figure 23; Tables 2,3).

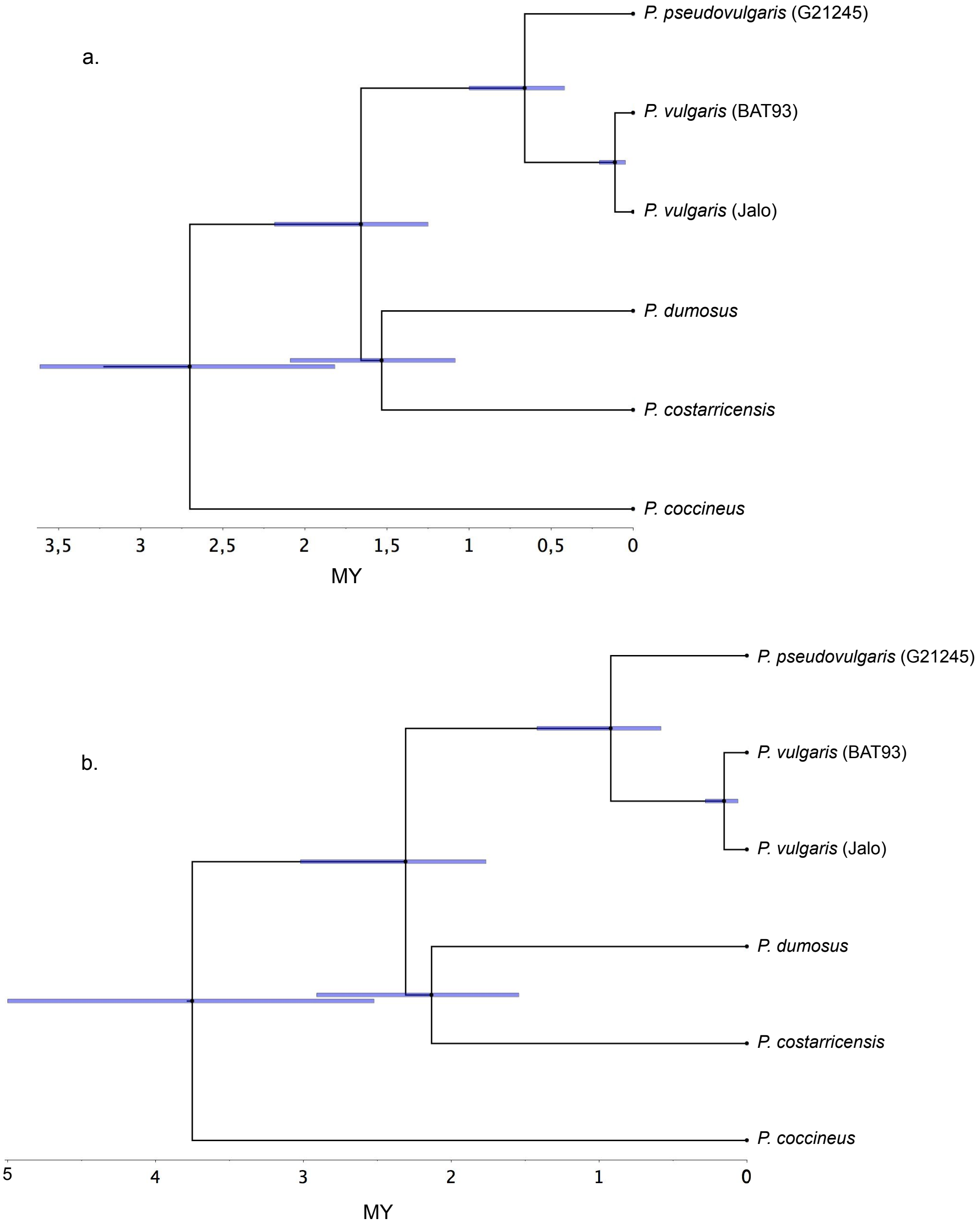


Figure 23. Tree topologies resulting from coalescent simulations. a. TMRCA using 165Kya of divergence between MA and AN gene pools as prior. b. TMRCA using 3Mya of divergence between *P. vulgaris* and *P. coccineus* as prior.

Table 2. Coalescent simulations (165 Kya prior)

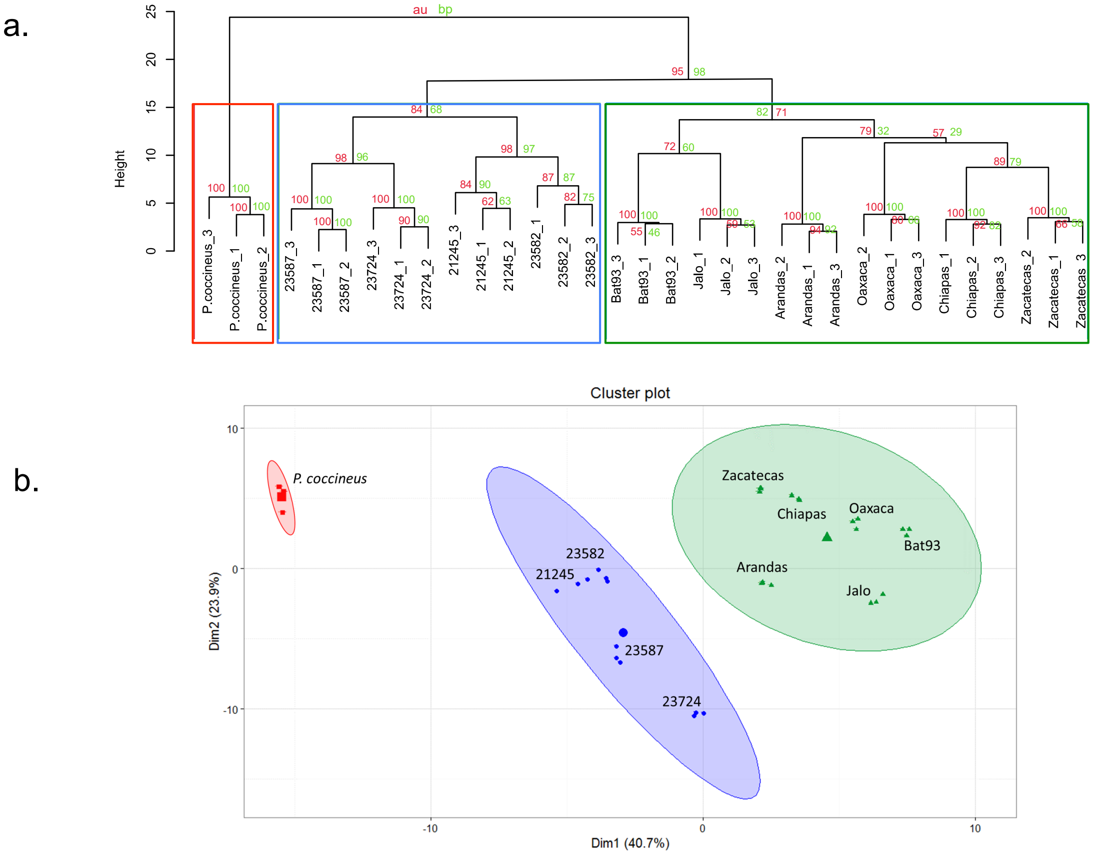
Summary Statistic	tmrca (*P. dumosus* / *P. costaricensis*)	tmrca (MA/AND gene pools)	tmrca (*P. vulgaris* / *P. pseudovulgaris*)	tmrca (*P. vulgaris* / *P. dumosus*)	tmrca (*P. vulgaris* / *P. coccineus*)
Mean	2.1804	0.1711	0.9808	2.4202	3.8029
Stderr of mean	0.0101	2.7663E-3	0.0117	0.0224	0.0153
Stdev	0.6211	0.3108	0.8779	2.1294	1.1571
Variance	0.3858	0.0966	0.7706	4.5342	1.3389
Median	2.1338	0.1548	0.922	2.3264	3.7525
Geometric mean	2.1477	0.1563	0.9441	2.3698	3.7415
95% HPD Interval	[1.5403, 2.9119]	[0.0663, 0.2887]	[0.582, 1.4424]	[1.705, 3.2055]	[2.5401, 5.0339]
Auto-correlation time	7209.8729	2139.4536	4765.0878	2990.8166	4710.2285
Effective simple size	3745.1423	12620.9795	5666.6323	9028.3036	5732.6306

Table 3. Coalescent simulations (3 Mya prior)

Summary Statistic	tmrca (*P. dumosus* / *P. costaricensis*)	tmrca (MA/AND gene pools)	tmrca (*P. vulgaris* / *P. pseudovulgaris*)	tmrca (*P. vulgaris* / *P. dumosus*)	tmrca (*P. vulgaris* / *P. coccineus*)
Mean	1.5656	0.1269	0.7021	1.7319	2.731
Stderr of mean	4.7861E-3	7.2413E-3	0.0114	0.0141	0.0114
Stdev	0.4613	1.1871	1.5387	1.8349	1.497
Variance	0.2128	1.4092	2.3676	3.367	2.2411
Median	1.5318	0.1112	0.6613	1.6662	2.6993
Geometric mean	1.5438	0.1119	0.6725	1.6953	2.6841
95% HPD Interval	[1.0849, 2.0854]	[0.0484, 0.2062]	[0.423, 1.0038]	[1.2216, 2.2757]	[1.8173, 3.6215]
Auto-correlation time	2906.9868	1004.7639	1494.6592	1592.8139	1556.4576
Effective simple size	9288.6559	26873.9738	18065.6569	16952.3886	17348.3689

7.2.4 Metabolomic profiling differentiates *Phaseolus* species

High-throughput, non-targeted mass fingerprinting has been shown to be a powerful tool that allows inter and intra species discriminations (Montero-Vargas *et al.* 2013; Sotelo-Silveira *et al.* 2015). Therefore, combining direct-injection electrospray mass spectrometry (DIESI-MS) and hierarchical cluster analysis (HCA), we constructed the metabolic profiles of *P. vulgaris*, *P. pseudovulgaris* and *P. coccineus* accessions from young trifoliate leaves. More than one thousand different mass to charge signals (*m*/*z*) were recovered, representing the 'metabolic space' of each accession. After mass error removal and signal filtering, 318 high quality mass signals of metabolites were kept for further analyses. The HCA of the one hundred most abundant metabolites correctly isolated *P. coccineus* as the outgroup and discriminated *P. vulgaris* accessions into wild or domesticated varieties. Most importantly, the *P. vulgaris* accessions were separated from its cryptic sister species from AHZ, placing them in two independent clades (Figure 24).



c.

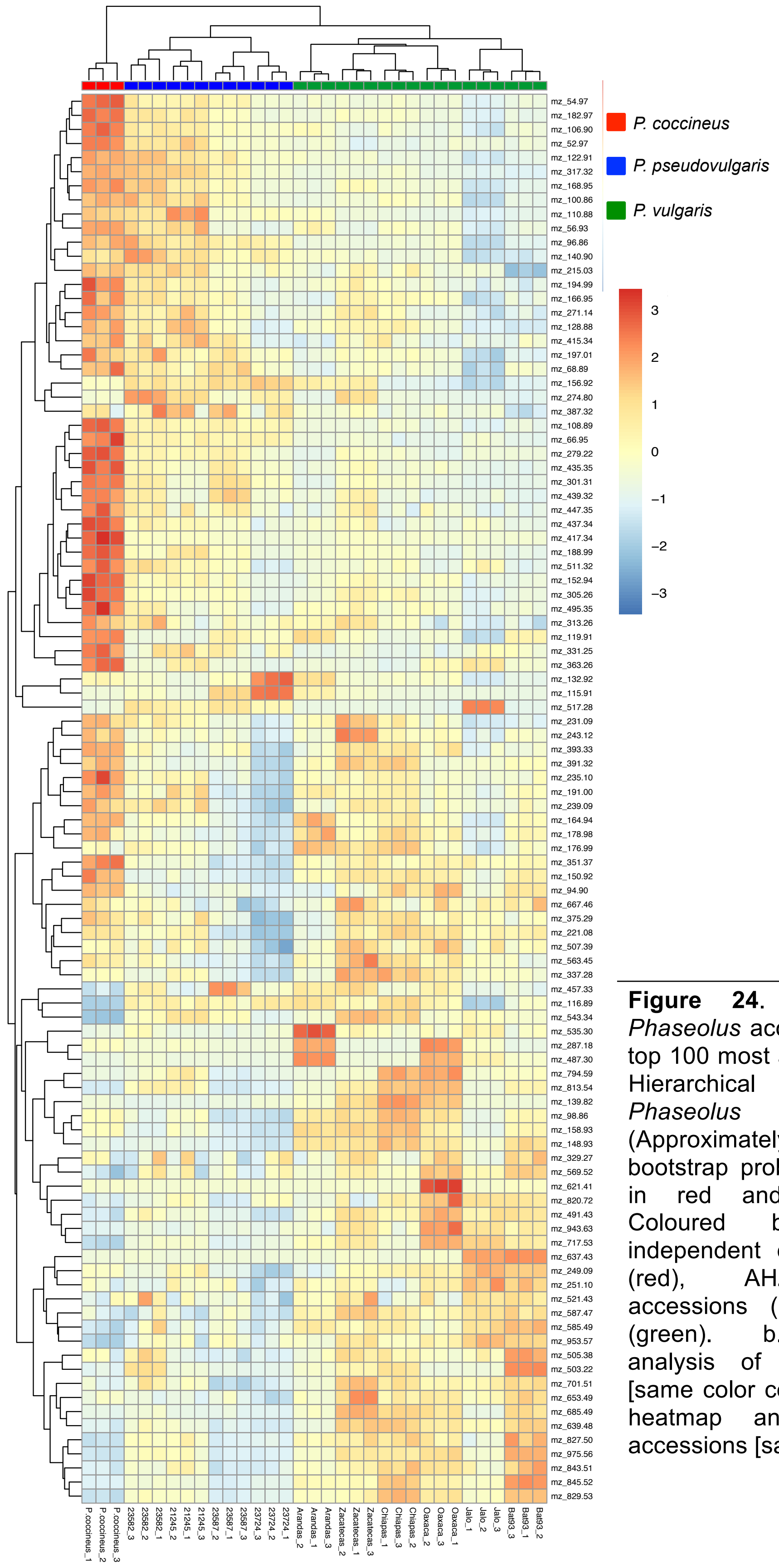


Figure 24. Metabolic profile of *Phaseolus* accessions considering the top 100 most abundant metabolites. a. Hierarchical clustering tree of *Phaseolus* accessions; AU (Approximately Unbiased) and bootstrap probabilities are highlighted in red and green, respectively. Coloured boxed enclose the independent clades of *P. coccineus* (red), AHZ-*P. pseudovulgaris* accessions (blue) and *P. vulgaris* (green). b. Principal component analysis of *Phaseolus* accessions [same color code as (a)]. c. Metabolic heatmap and clustering of the accessions [same color code as (a)].

Using a data mining approach (Winkler, 2015) we identified the thirty most important variables that best explain the metabolic differences between common bean populations. The dendrogram constructed from those variables reassembled the phylogeny described in the previous section, with bootstrap and approximately unbiased probabilities (AU) supporting the topology (Figure 25).

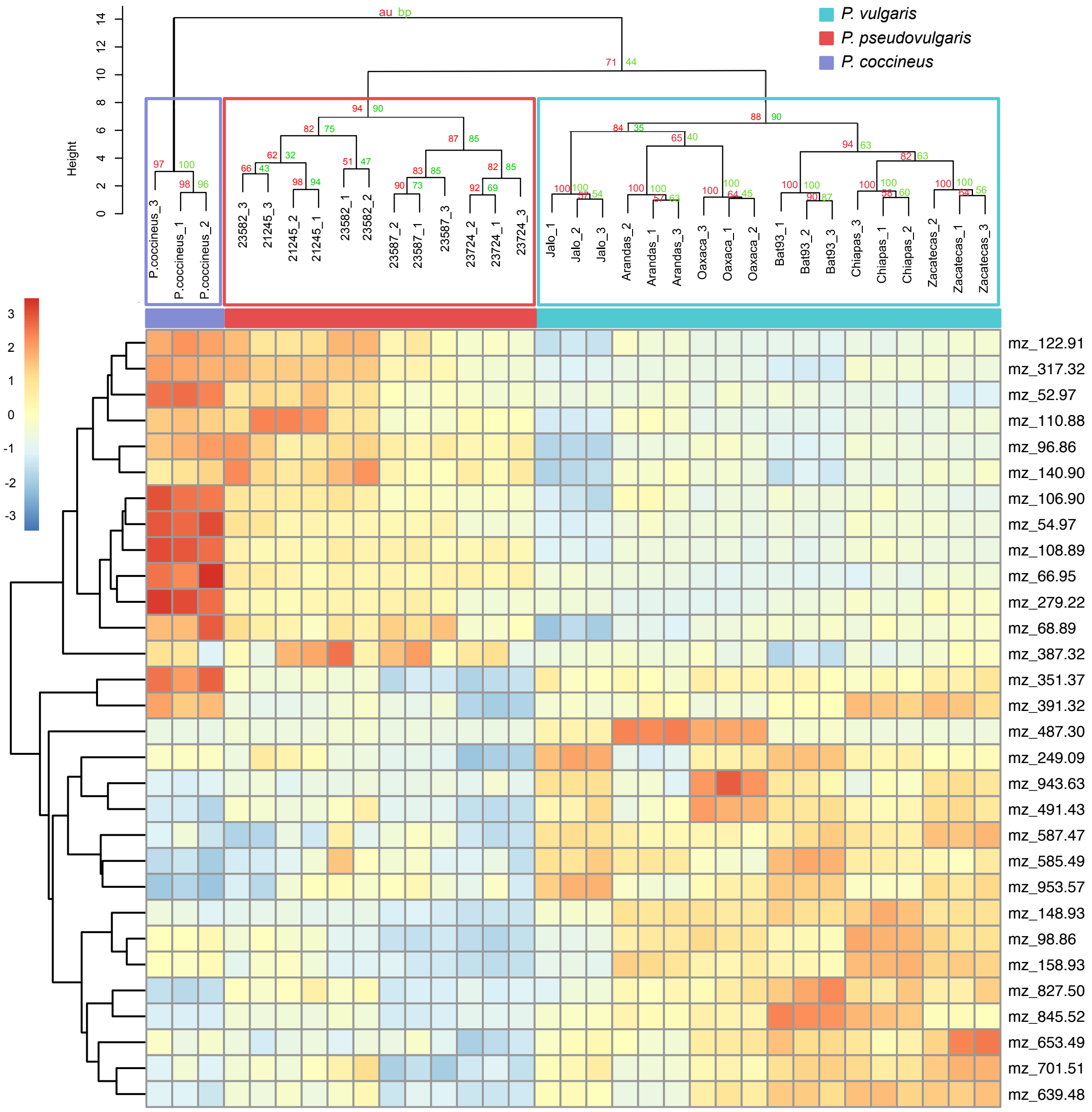


Figure 25. Metabolomic profile of *Phaseolus* species. The heatmap of the thirty most important mass signals from extracts of young trifoliate leaves of *P. vulgaris*, *P. pseudovulgaris* and *P. coccineus,* as well as the associated horizontal dendrogram reconstruct the phylogeny of the accessions.

Based on high-resolution LC-MS data we identified 44 metabolites, 25% of which belong to the 100 most important variables that explain inter-species differences. Most of the metabolite diversity in this set corresponds to flavonoids, such as the isobars of luteolin and kaempferol, or the coumarin derivative 4-methylumberlliferone (Suppl. Table 3, Appendix A). In legumes flavonoids play a crucial role during legume-microbe interactions in the rhizosphere (reviewed by Reddy *et al.* 2007). Luteolin in particular is a strong inducer of Nod gene expression (Peck *et al.* 2006), chemo-attractant and growth regulator of rhizobia (Caetano-Anolles *et al.* 1988); kaempferol is a flavonol involved in the regulation of auxin transport in response to rhizobia (Ng *et al.* 2015), while 4-methylumberlliferone is implicated in controlling lateral root formation (Li *et al.* 2011).

7.3 Recent evolution of *Phaseolus* species

7.3.1 Genomic introgression between *Phaseolus* species.

As discussed in section 2.2, several lines of evidence converge in the establishment of two geographically and genetically isolated gene pools, one in Mesoamerica and one in the Andes, from which, two independent domestication events took place followed by local adaptations and further expansions. Along these processes, not only the genetic diversity of the domesticated varieties has been compromised due to the domestication bottlenecks, but also, hybridization events between wild and domesticated populations, as suggested by morphological variation and microsatellite diversity, have occurred (Beebe *et al.* 1997; Payró *et al.* 2005; Zizumbo-Villarreal *et al.* 2005; Martínez-Castillo *et al.* 2006; Worthington *et al.* 2012), displacing the original genetic diversity in these regions.

Taking advantage of several tools [Patterson's D statistic for ABBA-BABA tests (Green *et al.* 2010; Durand *et al.* 2011; f estimators (Martin *et al.* 2014)] recently developed to determine introgression events between populations, we looked for such signals within and between *Phaseolus* species.

7.3.1.1 Unbalanced intra- and inter-species introgression

As our first approach to determine how frequently recombination events were taking place inside the genus *Phaseolus*, we conducted the calculation of Patterson's D estimator (ABBA-BABA test) using biallelic sites and the binary form of the formula by taking each genome as an independent sample. Windows of 100kb were chosen, and each chromosome was studied independently. Important observations were derived from this screening. In the first place, introgression events are much more frequent inside *P. vulgaris* as a species, with only a few blocks shared between closely related species, such as *P. vulgaris*, *P. coccineus*, *P. dumosus* and *P. costaricensis*. The frequency of recombination is particularly high between wild genotypes from Mesoamerica, as one could expect given the outcrossing rates of wild *P. vulgaris*. Recombination is also more frequent inside each gene pool, which is also expected given the geographic isolation of the accessions. In the particular case of BAT93, its genome showed several shared blocks with wild Mesoamerican genotypes, reflecting its hybrid origin.

Interestingly, each chromosomes displayed different recombination patterns, as depicted in the CIRCOS plots of Figure 26. The introgression signal between particular genotypes is consistent with the differences in the tree topologies of each chromosome described in previous sections and highlights the sensitivity of the ABBA-BABA approach to differentiate between alleles shared by ancestry or by recombination.

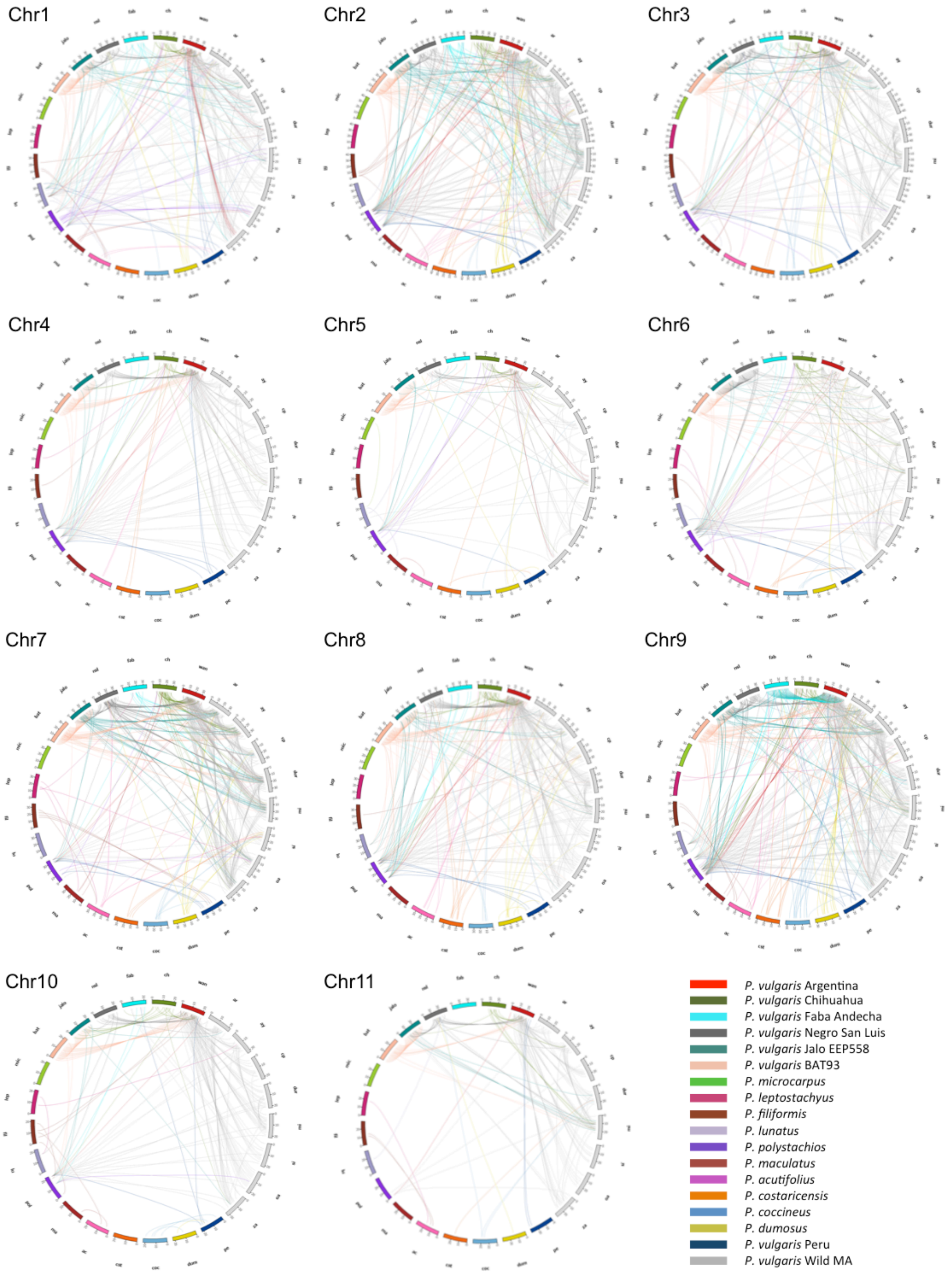


Figure 26. Introgression signal determined using Patterson's D estimator (ABBA-BABA) with bialleic sites.

However, more recent revisions of the ABBA-BABA method to determine introgression events, have also developed new improvements to the Patterson's D statistics, that take into account not only binary characters at bi-allelic sites (0/1), but allele frequencies when populations are being sampled. For this purpose, we combined two different parameters (Martin *et al.* 2014), the dynamic estimator of the degree of introgression between subpopulations (fd) and the absolute genetic distance (d_{XY}), both of them calculated in non-overlapping 5Kb windows across the eleven linkage groups. Several triads were considered to estimate the f_d parameter, permuting the receptor subpopulations and fixing *P. hintonii* as the outgroup.

Overall, there is a clear tendency to increase the introgression signal as we compare close subpopulations, in other words, introgression occurs with higher frequency within a species (Figure 27) as f_d values are close to 0.3 between *P. vulgaris* subpopulations, regardless of their wild or domesticated traits, whereas inter-species f_d values drop to 0.05-0.1.

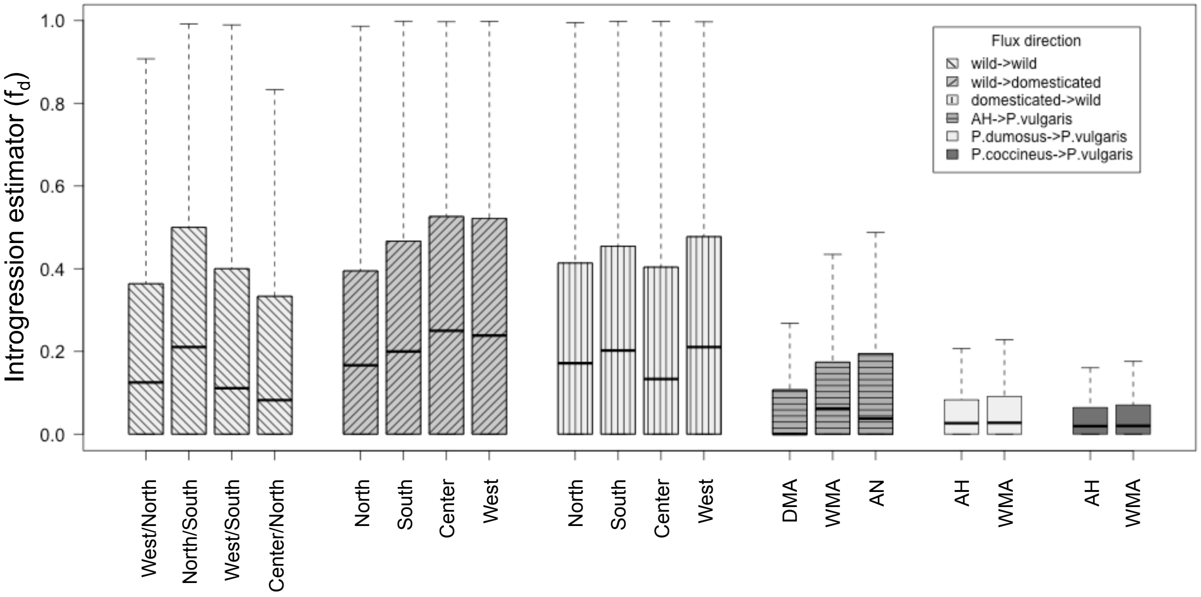


Figure 27. f_d estimator of introgresison between *Phaseolus* subpopulations.

Our genome-wide introgression analyses confirmed that there is a remarkable asymmetry of genetic flow between wild and domesticated common bean subpopulations, as already measured trough AFLP diversity and phenetic discrimination of wild, domesticated and weedy germplasm (Papa and Gepts, 2003). While it can occur in both directions, the genomic contribution in terms of the total length of introgressed tracks is much more limited from wild into domesticated genotypes than from domesticated into wild subpopulations. Indeed, condensed introgressed blocks account for 4.1 – 8.2 Mb when wild subpopulations are tested as donors and domesticated subpopulations as receptors, but in the opposite direction, the introgression signal spans from 5.7Mb up to 17.1 Mb. This is particularly evident when we take the Central subpopulation of wild MA genotypes as the receptor in the f_d triad (Figure 28); according to local records (Sagarpa, 2014) Zacatecas and Jalisco are among the most important states that produce common bean in Mexico – Zacatecas is the first producer during spring and summer, while Jalisco produces it all along the year- therefore, it is not surprising that genetic flow from domesticated genotypes occurs with such frequency in the area.

Even though the strongest introgression signal is maintained within MA *P. vulgaris* genotypes, the length of single introgressed genomic tracks between wild subpopulations or from wild into DMA does not reach recombination units as most of the fd+d_{XY} signal falls in small blocks of less than 50Kb with a maximum length of 200Kb (Figure 28). In the case of wild accessions, this might be a direct consequence of the frequency of hybridization events between populations in MA that have maintained high levels of genetic diversity (*pi*South= 0.0057; *pi*North= 0.0066; *pi*West= 0.0033; *pi*Centre= 0.0054; Figure 29). In domesticated accessions introgression blocks from wild neighbours might be broken as a consequence of selection against hybrids that are easily recognized by farmers. In contrast to these observations and in line with the asymmetric gene flow between wild and domesticated genotypes, we detected longer introgressed blocks from domesticated into wild genotypes, reaching up to 500Kb that could be the result of more recent hybridization events, particularly in geographic areas of high common bean production.

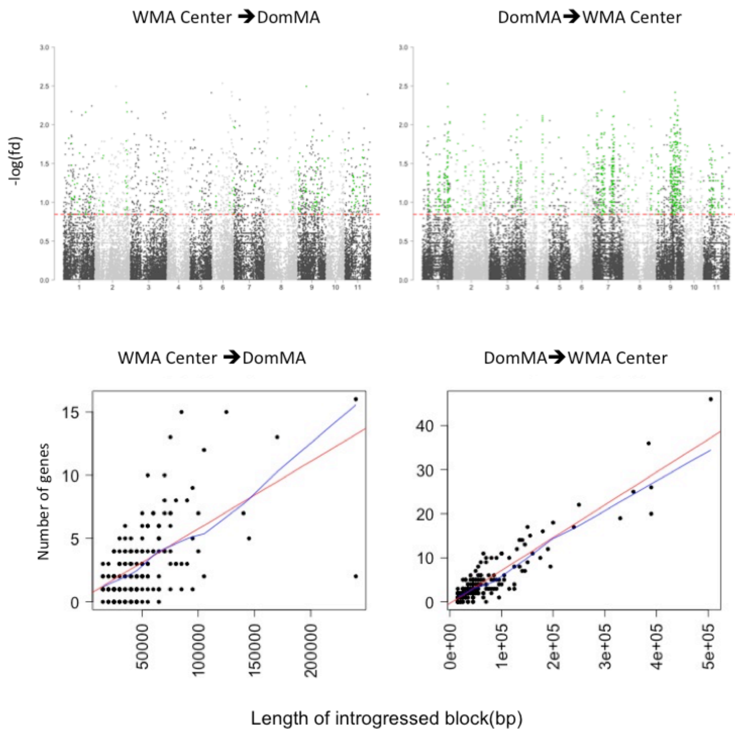


Figure 28. Unbalanced introgression signal between wild and domesticated subpopulations in MA. a. Global f_d estimations across the linkage groups divided in 5Kb non-overlapping windows is represented in Manhattan plots (top panels); the red threshold lines show the top 5% f_d outliers in each comparison, and strong signals of introgression (fd+d_{XY}) are highlighted in green. The number of genes encoded in each introgressed block is represented as scatter plots [bottom panels – colour lines: linear (red) and local (blue) regressions].

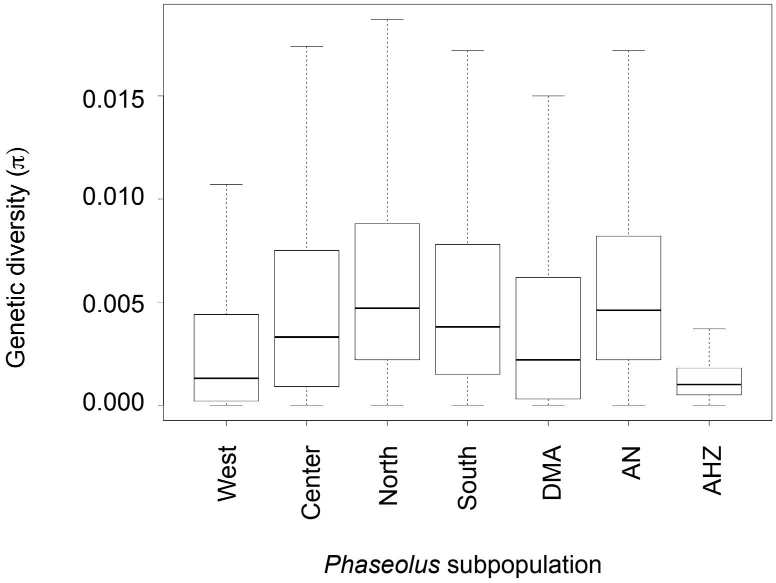


Figure 29. Genetic diversity within different *Phaseolus* subpopulations.

The unbalanced genetic contribution of wild and domesticated populations as a consequence of hybridizations is also measurable in terms of the number of genes that are transferred between subpopulations (Figures 28 and 30). The number of PCGs is highly dependent on the length of the introgressed block and, with a few exceptions as in the case of introgression from the Southern wild population into DMA, in most triads we observed a direct proportionality between gene content and window length (Figure 30), with more genes moving from domesticated genotypes into wild subpopulations, reaching up to 40 genes per track. However, we should not discard that the detection of high fd+d_{XY} signals between hybrid subpopulations could underestimate the introgression spans, which might cover larger portions of the genome if homogenous populations were tested.

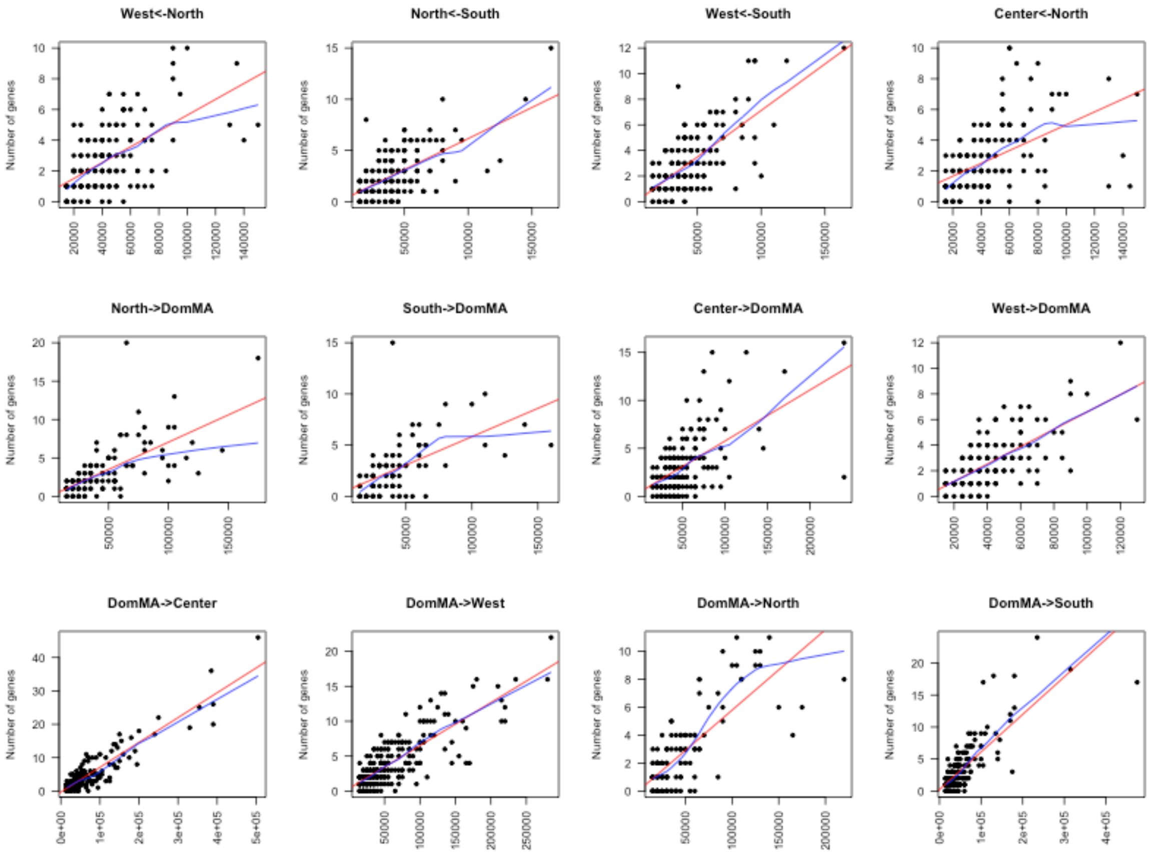


Figure 30. Number of genes encoded in each introgressed block. Colour lines: linear (red) and local (blue) regressions.

We also detected an unbalanced genetic flow between the Northern and Southern American hemispheres taking the AHZ genotypes as an intermediate subpopulation. Even though introgression signal could be detected in both directions, from AHZ to MA and from AHZ to AN, it is stronger towards the Andean accessions (Figure 31), a fact that can be explained both by the phenology of the plants and by the lower levels of genetic diversity in the Andean region that contribute to the maintenance of long (up to 335Kb) introgressed blocks (Bitocchi *et al.* 2012). Moreover, the AHZ subpopulation seems to be preferentially autogamous based on two important observations: first, the level of genetic diversity in the Andean deflection is lower than in any other tested subpopulation (*pi*AHZ=0.0017); second, we detected weaker introgression signals when these genotypes are permuted as receptors and AN as donors in the test triad.

This means that while *P. vulgaris* plants growing in the Southern hemisphere can be cross-pollinated by their AHZ wild neighbours, this does not occur in the opposite direction.

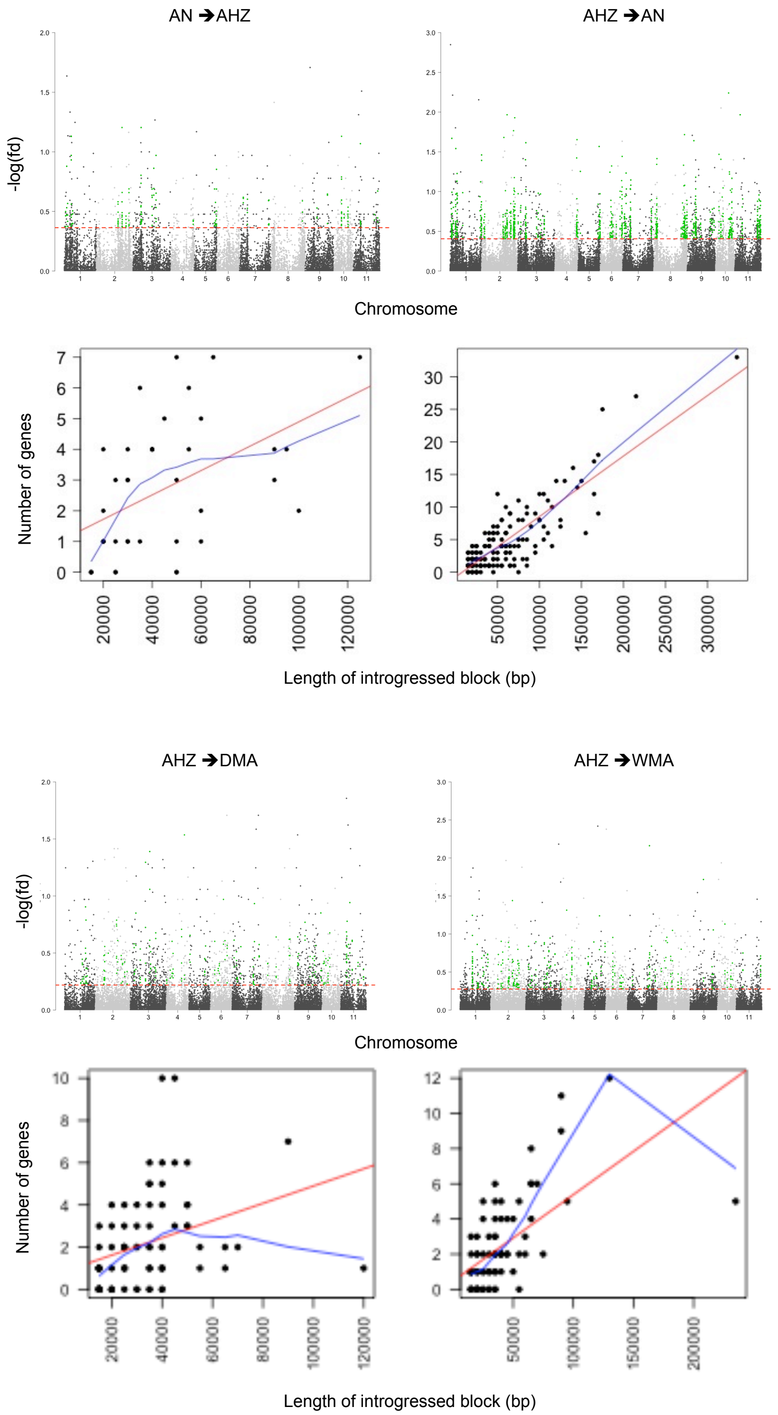


Figure 31. Unbalanced introgression signal across Northern and Southern hemispheres. Global f_d estimations across the linkage groups divided in 5Kb non-overlapping windows is represented in Manhattan plots; the red threshold lines show the top 5% f_d outliers in each comparison, and strong signals of introgression (f_d+d_{XY}) are highlighted in green. The number of genes encoded in each introgressed block is represented as scatter plots [colour lines: linear (red) and local (blue) regressions].

Finally, it is possible that the genetic contribution of sister species that have reached the southern hemisphere, such as *P. coccineus* and *P. dumosus* (Figure 32), also contributed to the differentiation of *P. pseudovulgaris* with respect to *P. vulgaris*.

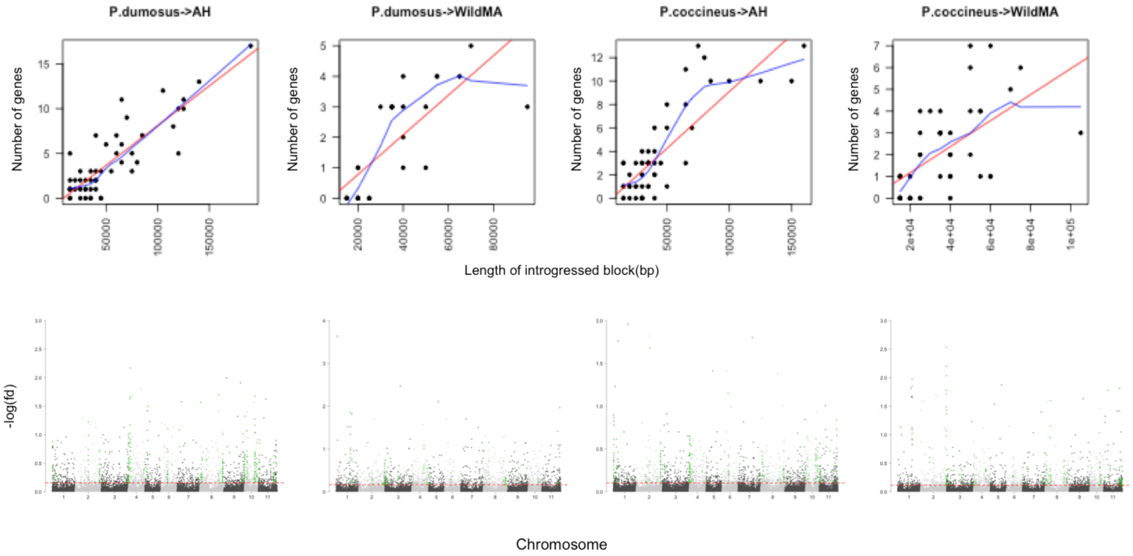


Figure 32. Introgression signal between sister species belonging to the Vulgaris group. Colour lines: linear (red) and local (blue) regressions.

7.3.1.2 Functional description of PCG within introgressed genomic blocks

The functional description of PCG dragged by introgression events unveiled interesting pathways that should be highlighted. First, GO terms associated to hormone-mediated signalling pathways, reproductive processes, post-embryonic development and the formation of reproductive organs are enriched between *P. vulgaris* subpopulations. Moreover, several categories related to purine and pyrimidine metabolism, ATP catabolism and other energy sources were also detected. The functional implication of such terms are relevant, as purine and pyrimidine are not only the building blocks for nucleic acid synthesis, but have been also implicated in developmental processes that include embryo maturation, germination, dormancy, fruit ripening, and leaf senescence (reviewed by Stassolla *et al.* 2003). Interestingly, as previous reports in other crops have shown (discussed in section 2.2), genes involved in biotic and abiotic stress response were enriched in most of the *P. vulgaris* triads in both directions, from wild into domesticated subpopulations, from domesticated into wild subpopulations and also between wild genotypes, implying that the continuous movement of such loci favours the adaptation of common bean to different habitats. Genes within these categories correspond to WRKYs, leucine-rich repeat receptor kinases, pathogenesis related proteins, glutathione S-transferases and SnRK2 protein kinases, among others (Figures 33 and 34).

Contrary to the mobility of genes behind reproductive processes within *P. vulgaris* subpopulations, such categories were not statistically enriched when *P. pseudovulgaris* and *P. vulgaris* are evaluated in the introgression triads. In these cases, categories related particularly to transport (i.e. vesicle docking and organization) and response to a wide variety of stimuli and stress were enriched. This is a very important observation that gives further support to the reproductive isolation of accessions collected in the AH Deflection. As discussed in section 2.2, those alleles that determine species phenotypes tend to introgress at a very low frequency, therefore, the fact that genes involved in reproductive processes are not transferred according to our analyses, suggests that these loci are important in the establishment of reproductive barriers between these close species.

Finally, GO terms related to cell wall biogenesis and organization, pectin and cell wall polysaccharide metabolic processes were enriched among introgressed genes transferred from *P. coccineus* and *P. dumosus/P. costaricensis* into *P. vulgaris*, which could contribute to the acquisition of pathogen resistance in *P. vulgaris* (Miedes *et al.* 2014).

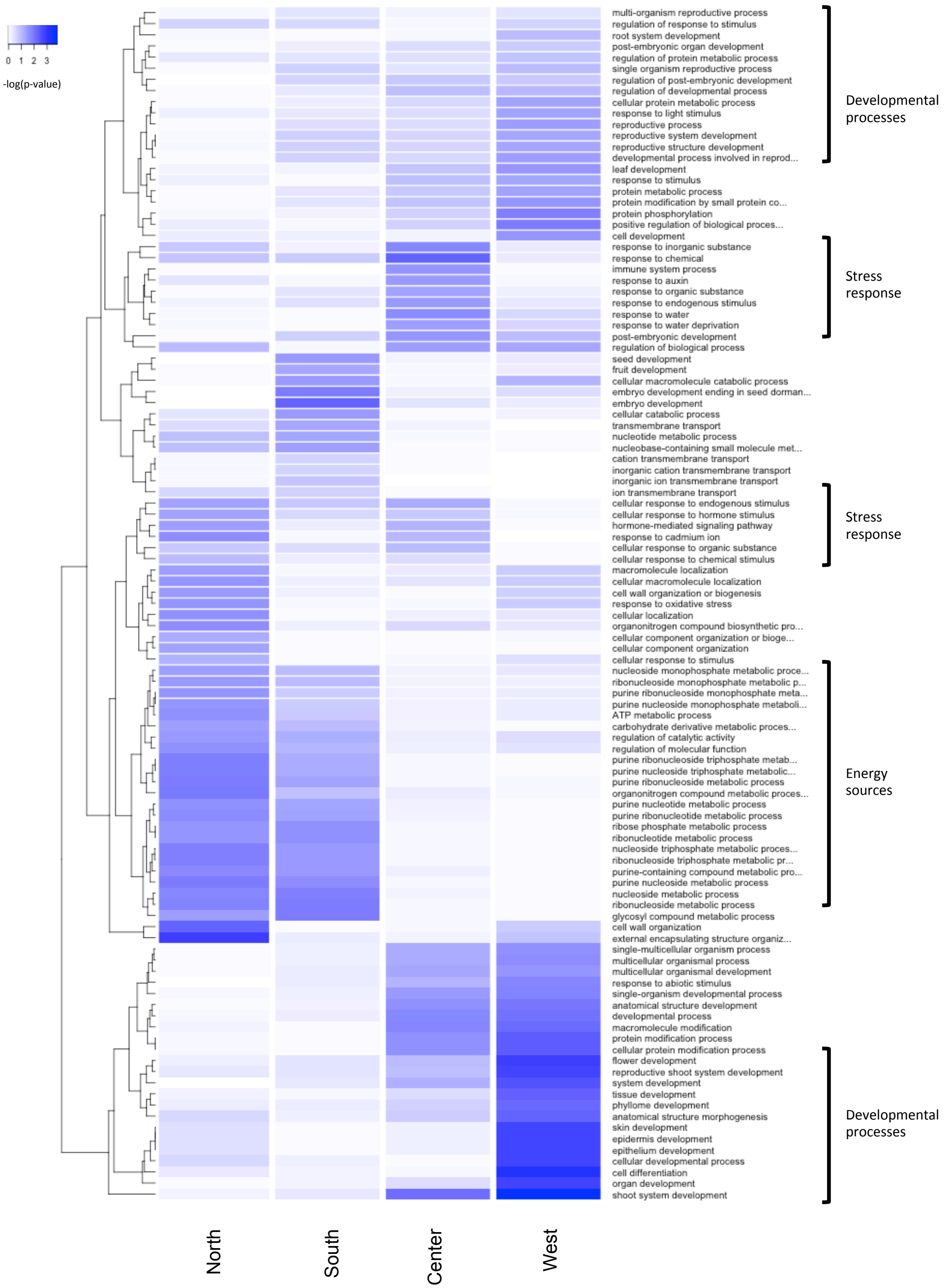


Figure 33. Enriched categories among PCGs introgressed from wild MA subpopulations into domesticated genotypes.

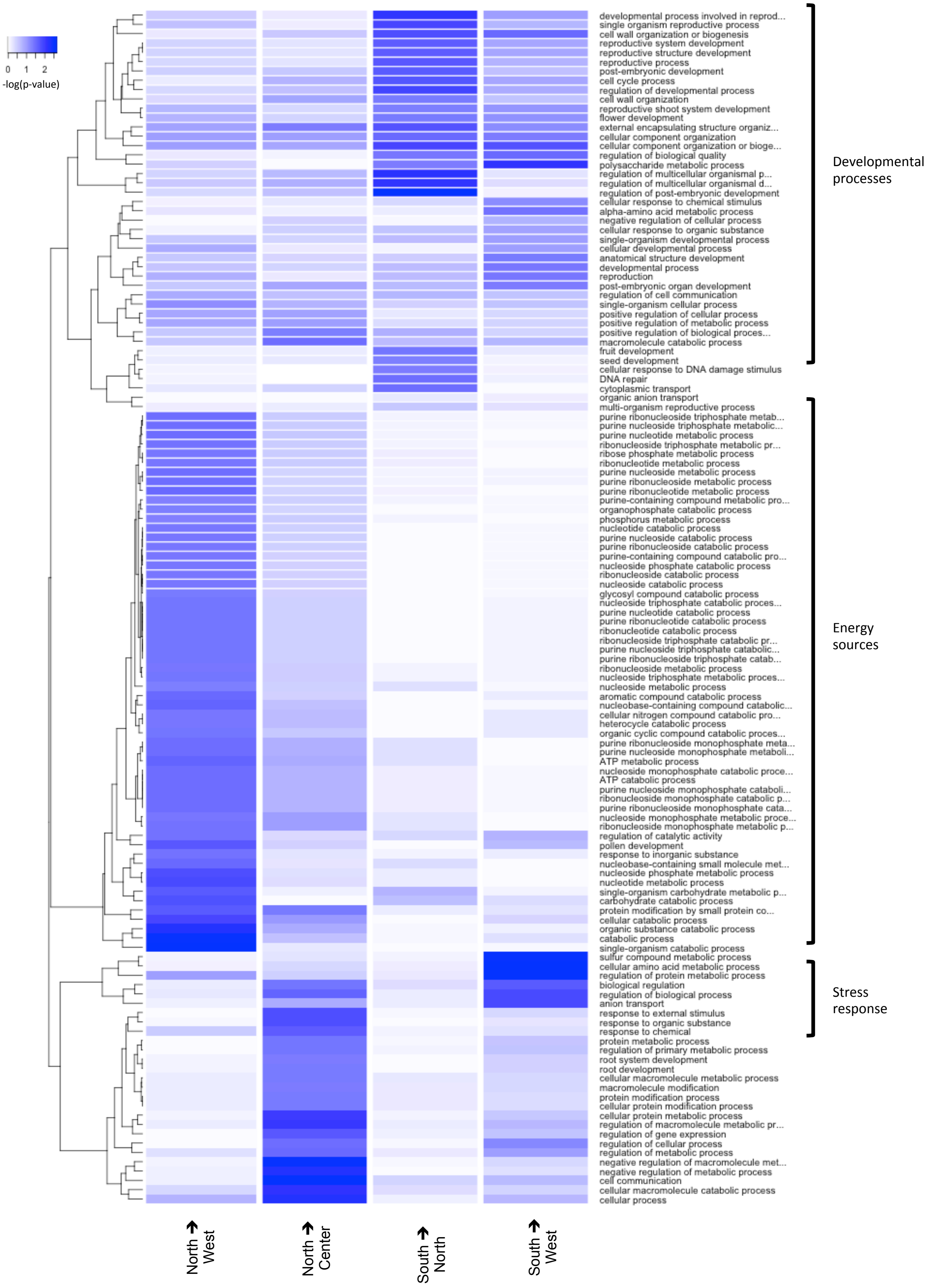


Figure 34. Enriched categories among PCGs introgressed between wild MA subpopulations.

7.3.2 Genomic landmarks for domestication syndrome

7.3.2.1 Selection signature through classic population genetics estimators

Given the importance of domestication as an accelerated evolutionary process that shapes the genomes of domesticated species, it is essential to trace the genomic regions and coding loci behind the phenotypic changes that altogether integrate the domestication syndrome. In several crops, certain genes have been directly associated to such traits, that is the case of maize and rice, however, in spite of years of research in the field, little is known about the "domestication genes" in *P. vulgaris* – even less in other *Phaseolus* species that have also gone trough the domestication process in America. An initial attempt to identify domestication gene candidates was performed using population genetics tools: Tajima's D and loss of genetic diversity estimated as the ratio of Watterson estimators θ(domesticated)/θ(wild).

Based on the phylogenetic profiles already described, we calculated Watterson estimator (θ_W) and Tajima's D (Hurst 2009) on two separate groups of *P. vulgaris* by pooling together: 1) four wild Mesoamerican genotypes, and 2) the four domesticated genotypes; each genotype is intended to represent geographically isolated subpopulations. If artificial selection occurring during both domestication events targeted the same DNA tracts, we could expect the pool of domesticated genotypes to display positive values of Tajima's D, since both Andean accessions share common sequence variants, as should the two selected Mesoamerican accessions. The same regions, if neutral, would display negative Tajima's D values in the pool of wild genotypes given that the emerging sequence variants are specific to each wild subpopulation. At the same time, the loss of genetic diversity in these DNA tracts should be evidenced by θ_W [domesticated] /θ_W [wilds] < 1.

We tested this hypothesis at whole genome-scale for 50Kb genomic sliding windows (Figure 35). As previous studies of natural populations show (Nolte and Schlötterer 2008), average D values were negative for both wild and domesticated accessions (-0.39 and -0.42, respectively), compatible with neutrality of the vast majority of polymorphisms. This analysis revealed large DNA tracts ranging from 50Kb up to 300Kb, accounting for 21.5Mb, spread across the eleven linkage groups, reflecting the effects of the domestication process (z-score on Tajima's D; $p<0.05$). As a consequence of the induced self-pollination during the domestication process, a high degree of homozygosis was observed in the elite accession

Jalo EEP58 (Figure 35 g) compared to the wild MA genotypes (Figure 35 a-d) or even the landraces from Chihuahua and Faba Andecha (Figure 35e,f). This observation is compatible with our introgression screenings given that, in spite of the predominant autogamy of *P. vulgaris*, the outcrossing rates, particularly among wild subpopulations, have introduced an important amount of genetic diversity and polymorphisms. Faba Andecha on the other hand, is a particular example of an introduced Andean genotype in Europe that has been in contact with other landraces and thus, displays more heterozygosis as a consequence of recent inter-gene pool hybridizations (Angioi *et al.* 2010; Gioia *et al.* 2013).

We performed a similar screening on ~20,000 PCGs with high sequence coverage in the eight selected accessions. We identified 3,342 genes that displayed a partial or complete loss of genetic diversity and were likely targeted during the generation of the landraces; the effect of the domestication bottleneck was further confirmed using coalescent simulations. In addition, 2,583 domestication genes reported by Schmutz and co-workers (Schmutz *et al.* 2014) were traced in the *P. vulgaris* BAT93 reference genome and contrasted with our gene candidates (Figure 35, outer circles). It is noteworthy that the two independent studies share 723 PCG (723/1,678 that were identified as one-to-one “orthologous” by BLAST bidirectional best hit, BBH), many of which have been associated to domestication syndrome traits.

In order to reduce false selection signal due to genetic hitchhiking and the slow decay of linkage disequilibrium, our domestication candidates were limited to those encoded within genomic sweeps to a final set of 328 PCG. Furthermore, to identify genes under similar selection pressure, we searched for genes affected by non-synonymous mutations (nsSNPs) in the four domesticated accessions, considering only SNPs that affect sites that were otherwise conserved in all wild accessions as well as in *P. coccineus*. We identified 130 such genes, carrying nsSNPs (not necessarily the same SNP) in all four domesticated accessions. Overall, we can distinguish 453 unique domestication candidates that include PCGs previously described in other crops as key regulators of domestication syndrome traits, such as the MADS-box genes *pistillata* and *sepallata3*, involved in organ development and flowering; seed dispersal and dehiscence gene *bel1*-like; lycopene cyclase (*lcyB*), associated to carotenoid accumulation in endosperm tissues (Bai *et al.* 2009), seed storage proteins glutelin type-A/B; a polyphenol oxidase involved in fruit discoloration (Yu *et al.* 2008); an omega-hydroxypalmitate O-feruloyl transferase, involved in the synthesis of aromatics of the suberin polymer, an important component of the water-impermeable seed coat (Smýkal *et al.*

2014); galacturonosyltransferase-4 and pectin acetylesterase, both involved in pectin structure and accumulation; an auxin-induced protein 5ng4 (nodulin like), as well as genes related to plant architecture such as scarecrow-like 8 and the homologue to teosinte branched 1, *tb1*. A remarkable observation was the significant enrichment (Fisher's test p<8.85^{-8}) of the sucrose/starch biosynthesis pathway among bean domestication candidates.

As the effects of selection are not restricted to PCGs and can also affect, directly and indirectly, all kinds of regulatory elements, we evaluated the association of non-coding RNAs with the domestication process. We detected signal of selection on 53 miRNA precursors, whose functions implicate them in stress/disease responses (Arenas-Huertero *et al.* 2009; Ding *et al.* 2013) (miRNA169, 2118) and organ development and patterning (Rhoades *et al.* 2002; Yu *et al.* 2010; Sieber *et al.* 2007) (miRNA164, 159, 156, 166), as well as on 143 lncRNA genes. Interestingly, these lncRNAs are enriched in organ-specific expression patterns (1.3 fold, P<0.04). This trend is even stronger when considering fruits, in which the 13 tissue specific lncRNAs represent a two-fold enrichment over expectation. We also used co-expression analysis to identify domestication related ncRNAs, and identified 83 lncRNAs having a Pearson cc > 0.95 with at least one PCG for a total of 22 PCGs. This lncRNA set is also highly enriched in organ specific lncRNAs, with the most significant enrichments in fruits (41 lncRNA highly correlated with domestication PCGs) and in flowers (9 lncRNAs). Similarly, we detected 49 small ncRNAs correlated with 28 domestication PCGs (Pearson cc >0.95). These observations are consistent with a regulatory role of non-coding genes in domestication.

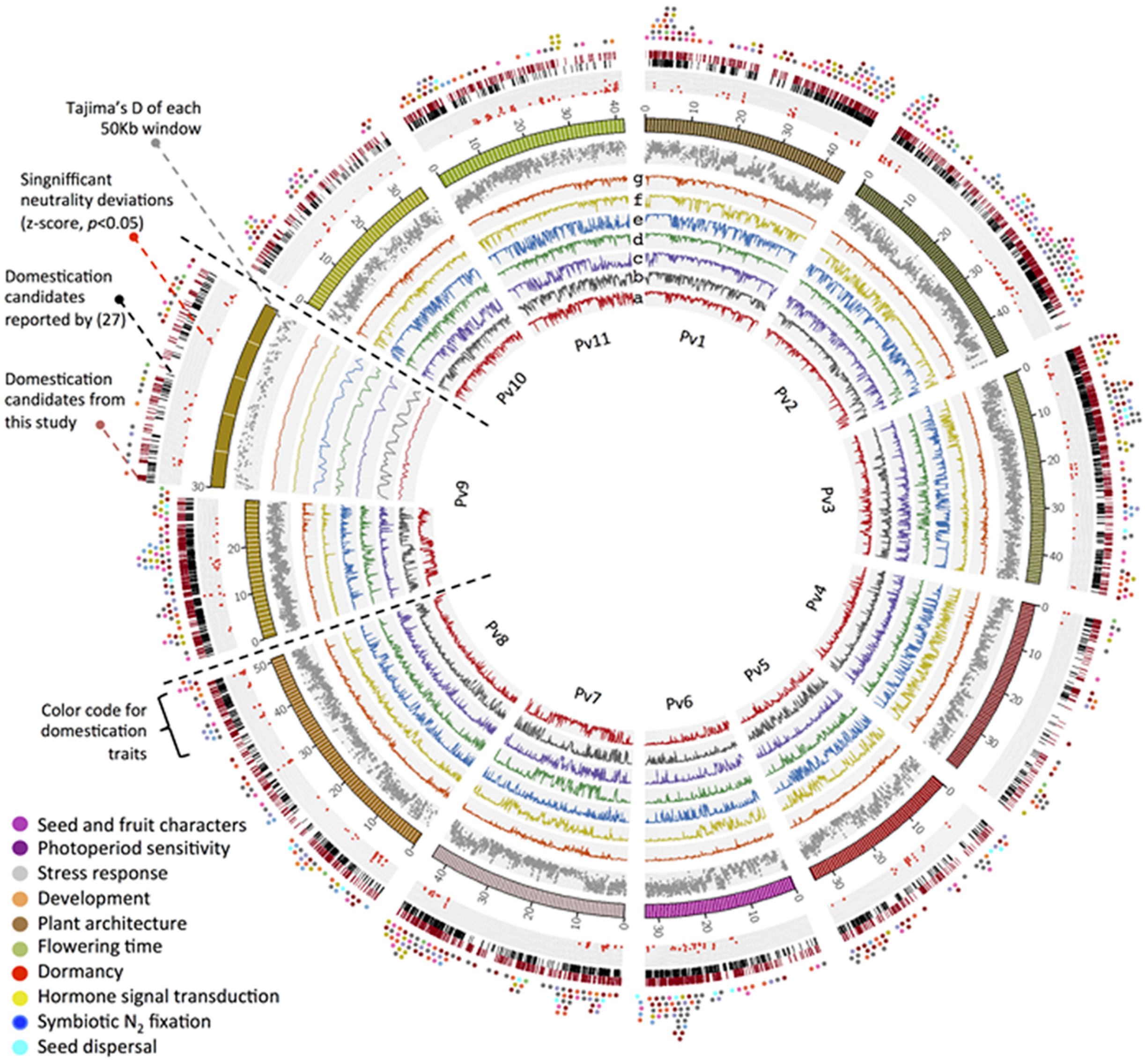


Figure 35. Genome scale screening of domestication effects. Level of homozygosis ($S_{hom}/S_{hom}+S_{het}$) in *P. vulgaris* accessions is depicted (inner circles): A. Zacatecas; B. Sinaloa; C. Oaxaca; D. Arandas; E. Chihuahua; F. Faba Andecha; G. Jalo EEP558. Windows with significant genetic losses (θ_W[domesticated]/θ_W[wilds] < 1) and strong Tajima's D signal in domesticated accessions (p<0.05) are highlighted with red dots. Linkage group size in Mbp; internal linkage group bands represent cM units (Fonseca *et al.* 2010).

7.3.2.2 Domestication signature through haplotype association

As our second approach to define gene candidates targeted by artificial selection, we looked for PCGs containing haplotypes associated to the “domesticated phenotype”. First, we defined haplotypes strongly associated to the domestication process in MA (absent in the Andean gene pool and in wild accessions from Mesoamerica and the AH Zone) and secondly, haplotype clusters present in both centres of domestication (CODs) but absent in all wild accession considered in our sampling. Given the lack of wild genotypes of Andean origin, we were not able to distinguish haplotypes associated exclusively to the domestication process in this area. Following this procedure, we identified 599 genes with common haplotypes to domesticated genotypes from MA and the Andes, and 619 genes with haplotypes specific to MA domesticated accessions (Figure 38a). Similarly, 52 and 45 lncRNAs with domestication-associated haplotypes were defined common to both CODs and MA, respectively (Figure 38b). Based on the clade specificity according to their level of conservation with other plant species, we can conclude that most of the genes with domestication signal are common to Rosids and Asterids (66% and 69% in both COD and MA, respectively), 15% are shared with Rosids in both cases, and 6% (COD) and 17% (MA) are legume specific genes – from this latter category in the MA gene set, 62.3% are *Phaseolus* specific genes, 32.1% are common to other legume species and 5.6% are only shared between common bean and soybean.

As a result of a global screening of protein definitions associated to domestication gene candidates, we identified 21 resistance genes and several GO categories statistically enriched that can be easily linked to the emergence of domestication traits (Figure 36). This is the case of the sucrose/starch biosynthetic pathway (directly related to the starch content in the seeds), the regulation of reproductive processes (involving transcription factors such as wox2 for embryonic patterning, or GTE1 that promotes seed germination), inflorescence development and meristem determinacy (e.g. homeobox gene knotted1), whose components displayed haplotypes common to both CODs. Also, the nodulation-signalling pathway 2 (NSP2) involved in Nod factor signalling in legumes and several genes related to dormancy and photoperiod sensitivity pathways were identified, particularly with SNPs at their regulatory regions (Figure 37). Other enriched categories such as chromatin assembly, nucleosome organization and the regulation of histone methylation, highlight an epigenetic control behind the domestication syndrome that should be further explored. Similarly, among genes with MA haplotypes, we identified enriched GO categories particularly related to the development of reproductive structures (e.g. carpel, gynoecium, flower and ovule), organ formation (including transcription factors KAN2 or AS1), and genes directly involved in auxin transport and homeostasis, or nodulation (early nodulin 93).

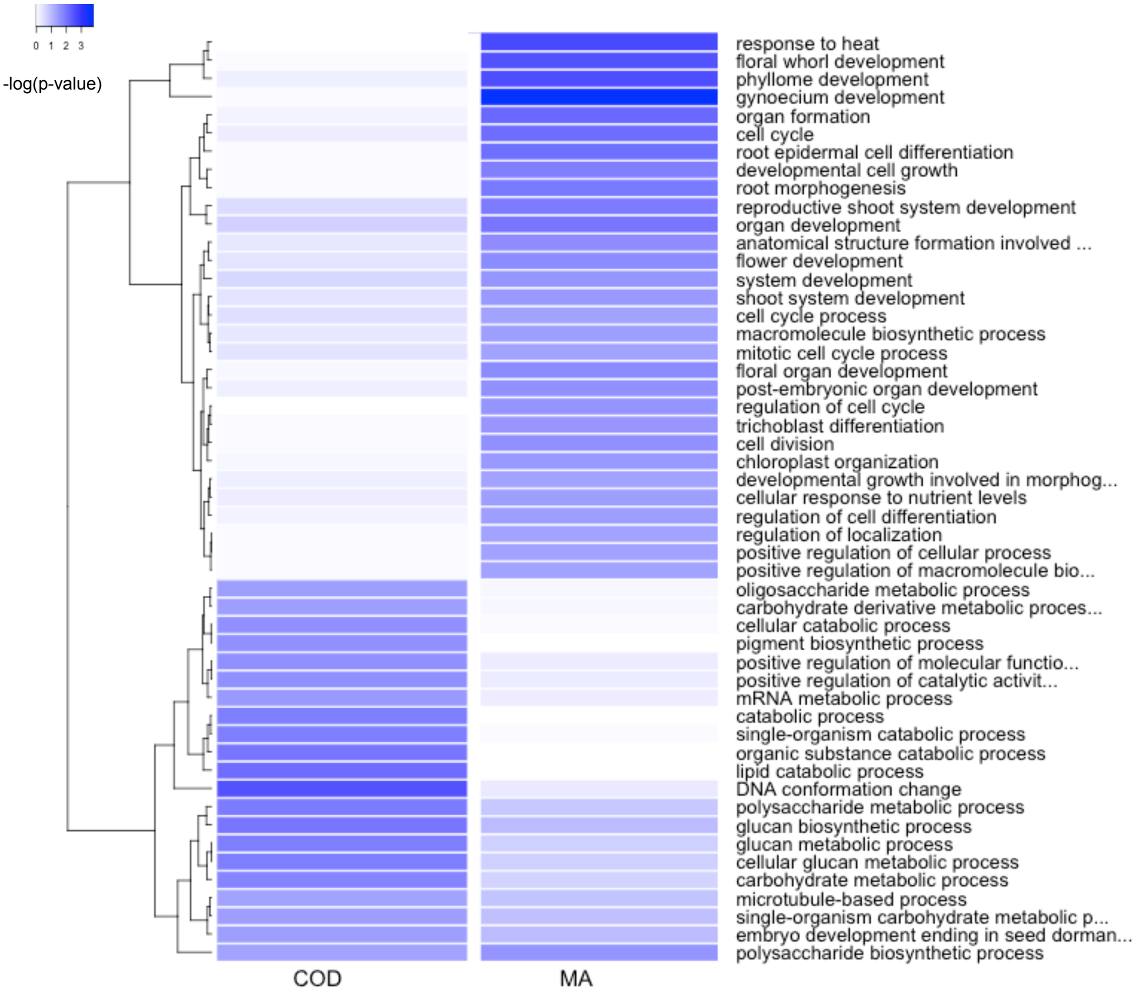


Figure 36. Functional description of domestication genes. Heatmap show GO enrichments from genes with domestication-associated haplotypes.

As expected based on previous findings (Papa *et al.* 2005), domestication gene candidates do not overlap with introgressed regions, as depicted in Figure 38a,b and c, as selective sweeps, i.e. regions that are rich in domestication-associated haplotypes, do not display signals of introgression.

In addition, we verified the expression pattern of domestication gene candidates in the transcriptional atlas of BAT93 recently published (Vlasova *et al.* 2016), and observed 115 housekeeping genes, as well as a small number of genes with organ (27 genes) and/or developmental stage (21 genes) specific expression. Among these, it was interesting to notice root specifically expressed genes such as a cysteine-rich RLK (resistance gene), a carotenoid cleavage dioxygenase (NCED1) that catalyses the first step of abscisic-acid biosynthesis from carotenoids in response to water stress, or a pollen specific leucine rich repeat.

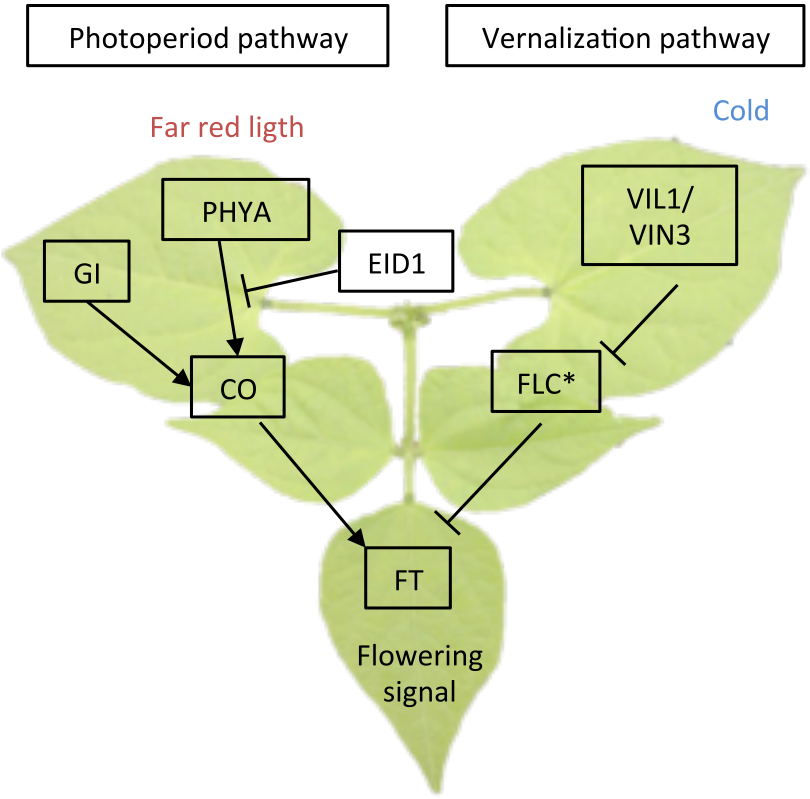


Figure 37. Photoperiod sensitivity and vernalization pathways. All genes associated to these pathways [except for (*)] share domestication-associated haplotypes in both CODs that differentiate them from wild individuals.

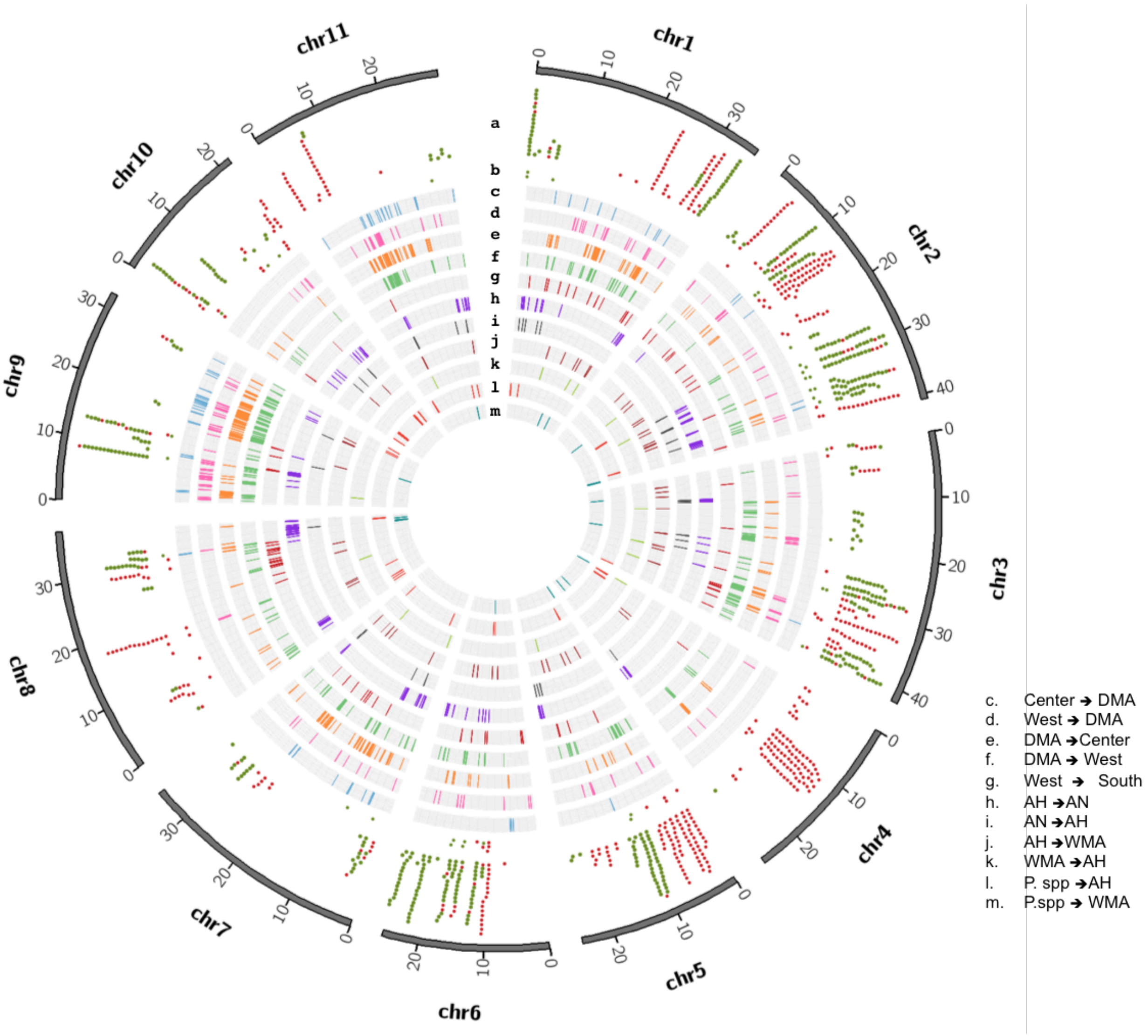


Figure 38. Introgression and domestication signals across *P. vulgaris* linkage groups. a. Domestication genes; green: common to both COD; red: MA-specific. b. LncRNAs with domestication-associated haplotypes (same colours as b). c-k. Introgressed blocks: c,d: wild➔domesticated; e,f: domesticated➔wild; g: wild➔wild; h-k: AHZ⇐➔*P. vulgaris*; l,m: *P. dumosus/P. costaricensis* ⇐➔ *P. vulgaris.*

8 Discussion

P. vulgaris migrations, bottlenecks and human-mediated artificial selection during domestication and improvement have shaped the genomes of wild and domesticated populations we observe nowadays. Such a complex history started to be disentangled many years ago, but it is until the emergence of high throughput sequencing technologies that we are really beginning to understand the evolutionary factors behind *Phaseolus* evolution. In the present study, we generated a high quality reference genome of a Mesoamerican elite variety BAT93 that, together with the previously published genome of an Andean variety of common bean (Schmutz *et al.* 2014), set the framework for deeper analysis on population dynamics and comparative genomics.

The AHZ genotypes correspond to the closest sister species to *P. vulgaris*

By means of genome resequencing and combining phylogenomics and metabolomics strategies, we clarified most of the discrepancies brought out by noisy phylogenetic signals of the genotypes collected in the Peruvian-Ecuadorian area in South America. Both phylogenetic profiles using nuclear and chloroplast markers evidenced that the Amotape-Huancabamba genotypes belong to the closest sister clade of *P. vulgaris*, and our coalescent simulations confirm that that AHZ genotypes emerged by allopatric speciation before the split of both common bean gene-pools. Moreover, following the original definition of ‘phenotypes’ as “all ‘types’ of an organism that are distinguishable by either direct inspection or only by finer methods of measuring or description” (Johannsen**,** 1911), our metabolic profiling data provide strong evidence of a phenotypic discrimination between *P. vulgaris* and the AHZ genotypes. The combination of such phenotypic signal and the genomic evidence pointing to a speciation event in the tropical Andes, strongly sustains the Mesoamerican origin of common bean and what we tentatively denote as “*Phaseolus pseudovulgaris*” could be considered as a cryptic species and an independent evolutionary unit in the Vulgaris group, close to the ancestor of wild *P. vulgaris* subpopulations in America.

Recently, it was shown that rhizobial strains isolated from common bean nodules in Mexico, Brazil and Ecuador belong to different taxonomic linages, native to each region in Mesoamerica or the Andes (Aguilar *et al*. 2004; Ribeiro *et al.* 2013); a subsequent, more extensive survey of nodule bacteria (Servín-Garcidueñas *et al.* 2015) revealed a clear preference of nodulation by *Bradhyrhizobium* in most *Phaseolus* species and a shift to *Rhizobium* nodulation in the Vulgaris clade. Similarly, coevolution of *P. vulgaris* and certain pathogens such as *Phaeoisariopsis griseola* and *Colletotrichum lindemuthianum* has been described, showing infection specificity by isolates from the Andean or Mesoamerican regions to cultivars from each gene pool (Guzmán *et al.* 1995; Geffroy *et al.* 1999 and 2000). These observations, combined with our metabolomic profiles that evidenced the abundance of secondary metabolites such as flavonoids as one of the variables that allows inter-species discrimination, suggest that different *Phaseolus* species have the capacity to select their symbionts from coexisting soil bacteria and that symbiont preference shifts have accompanied *Phaseolus* speciation and diversification processes in Mesoamerica.

Parallels and contrasts of two independent methods to define domestication candidates.

The use of two different strategies to define domestication candidates provided us with independent sets of genes that should be carefully examined. As discussed in previous sections, the results published by Schmutz and coworkers (2014), suggested that different genes, 1,835 in Mesoamerica and 748 in the Andean region, were selected for during both independent domestication events, with only 57 of them shared by both processes. Moreover, according to the authors, independent domestication genes were identified even between subpopulations belonging to the same gene pool, i.e. 1,424 domestication genes were affected by artificial selection in the Mexican landraces and 418 in the Central American landraces, with only 33 overlapping between both groups; in the Andean gene pool, the number of domestication candidates for each subpopulation (from the Northern and Southern Andes) is not given, but it is remarked that no intersection between both gene sets was observed. These results were achieved combining three population genetics estimators: Tajima's D, Fst and π. Tajima's D was required to be negative as an indicator of positive selection; genetic diversity losses were determined with the ratio $\pi_{Wilds}/\pi_{Landrace}$ >1 and high Fst values were indicators of high differentiation between wild genotypes and landraces. From an evolutionary perspective, it is viable to suggest that common bean was domesticated several times, as reflected by the interpretation of such population genetics estimators, but it seems little parsimonious that the same domesticated phenotype can be achieved through so many different metabolic and regulatory pathways.

We attempted to determine domestication gene candidates and selective sweeps following a similar strategy: we looked for those loci that displayed a loss of genetic diversity (θ_W [domesticated] /θ_W [wilds] < 1) and were subjected to a bottleneck (Tajima's D >> 0). This strategy led us to conclude that most of the domestication signal is contained in PCGs common to both centres of domestication, showing that the domesticated phenotype was achieved in Mesoamerica and the Andes through the selection of the same genes and pathways. The comparison of our domestication gene set (3,342) and the one reported by Schmutz and co-workers (2,583) had 723 PCGs in common (taking into account that only 1,678 were tagged as orthologous between BAT93 and G19833 by BBH). Such a small intersection points to an important conclusion: the use of population genetics estimators on heterogeneous populations (combining individuals from different locations), could lead to as many independent domestication gene sets as subpopulations are included in the screening.

Tajima's D is commonly used to identify DNA sequences evolving neutrally or following some type of selection (positive or balancing). This estimator is particularly sensitive to population structure, which can produce positive D values that would indicate that balancing selection or a population bottleneck are taking place. On the other hand, when heterogeneous populations are examined, they can produce a high number of rare variants that could be erroneously interpreted as population expansions or positive selection. The use of this estimator in our samples introduced an important bias: in a scenario where a particular locus has an Andean allele and a Mesoamerican one, not necessarily associated to domestication but fixed in the populations, the lack of other variants produced elevated Tajima's D. Such alleles could be easily interpreted as strong bottlenecks, exposing a large catalogue of PCGs with false domestication candidates. While Schmutz *et al.* (2014) sustain that their domestication candidates are subjected to positive selection based on negative D values, these are close to 0 (not significantly deviated from neutrality) and close to the D estimations in the wild subpopulations.

Given the above-described limitations that bias the interpretation of population genetics estimators, we explored a different strategy. A great advantage of the sampling in our study relies on the fact that whole genome re-sequencing allowed us to define haplotypes across the linkage groups that otherwise would be impossible to determine, particularly when pooled populations are used for selection screenings. Thus, we were able to use an unbiased approach based on haplotypes and their association to phenotypic traits to validate the domestication gene candidates targeted by artificial selection. Under these conditions, we could differentiate domestication haplotypes shared by both centres of domestication from those particular to Mesoamerica. The definition of Andean domestication haplotypes was not possible due to the lack of wild genotypes from South America. Nevertheless, our results suggest that domestication could target independent genes in each COD but several pathways were commonly selected for to achieve similar phenotypes.

If we contrast both approaches in terms of the number of shared domestication gene candidates, we obtain 220 PCGs that were identified with our first method (as having a complete loss of genetic diversity or a positive Tajima's D value indicating a strong bottleneck), and, in our second screening, having an haplotype cluster strongly associated to the domesticated phenotype. As we could expect, the highest overlap (153/220) corresponds to PCGs identified in our second screening that displayed θ_W=0 in our first protocol, as those candidates share the same variants among domesticated genotypes and differ from the alleles in the wild subpopulations. The overlap of these categories is not complete due to several reasons. In the first place, our first screening was restricted to four individuals per group in order to keep a balanced population size. Our second strategy was more inclusive, as we considered all possible haplotypes present in domesticated and wild genotypes, even including those from the AH Zone that could be closely related to the Andean background given the introgression signals we observed. The functional description of PCGs in the intersection of both methods, sustains the identification of genes directly associated to the emergence of domestication traits, such as phytochrome a involved in photoperiod sensitivity, vin3-like protein that participates in the vernalization pathway, gibberellin dioxygenase responsible for plant architecture and changes in growth habit, and several stress response genes (i.e. ethylene-response transcription factor ERF3, pollen specific lrr proteins, receptor-like protein kinase rhg1 that helps in recognition and defence against nematodes).

Still, the intersection of the haplotype-based domestication candidates and those reported by Schmutz and co-workers was very low, with only 65 shared PCGs. This observation led us to analyse in more detail how the published domestication gene set was obtained. In spite of the wide sampling of genotypes in the published study, certain biases have to be highlighted: even when Guatemalan and Colombian populations have been defined as independent gene pools, no wild genotypes were included to contrast the landraces with their wild relatives in Central America; the same occurred with the Northern Andean pool, as wild genotypes belong preferentially to the Southern Andean pool. More importantly, the Fst index was calculated considering wild genotypes and landraces from Mesoamerica and the Andes independently, i.e. landraces were only contrasted with their wild neighbours and no differentiation index was calculated between gene pools. The implication of such limited comparisons points to a crucial weakness of the screening: gene flow across gene pools was neglected and thus, many loci with high differentiation index between wild genotypes and landraces are not necessarily the outcome of artificial selection but rather resulting from introgressions from the opposite gene pool due to the dispersal and recent introduction of landraces in different regions. To prove this scenario, we contrasted the domestication candidates provided by Schmutz and co-workers with the PCGs encoded within introgressed blocks identified in our analyses. Indeed, one third of their domestication candidate genes have signals of introgression, indicating that these are neutral loci easily transferred by hybridizations between common bean subpopulations. In particular, 327 PCGs were found to be introgressed between domesticated accessions from the Andean and Mesoamerican gene pools or from *P. pseudovulgaris* into *P. vulgaris* domesticated subpopulations.

The lack of more wild genotypes from the Andean gene pool in our study makes it impossible to exhaustively examine the extent of the noisy signal introduced by gene flow in the published results (Schmutz *et al.* 2014). However, the high intersection of introgressed PCGs and the reported domestication candidates makes it clear that without considering the role of gene flow through introgression along common bean evolution, a large proportion of domestication genes are erroneously called. Restricting domestication haplotypes as those that are not found in any wild genotype (from Mesoamerica or the Andes) and contrasting such signals with introgression events, makes us confident that our approach is closer to the definition of a core set of domestication genes that could be enriched and validated as more accessions are sequenced.

Unbalanced inter- and intra-species introgression events were detected in Mesoamerica and across hemispheres.

Consistent with previous findings, we observed unbalanced intra- and inter-species genomic flow in Mesoamerica and across hemispheres. Interestingly, it was particularly in the inter-species triads that we observed the combined introgression signal (fd+d_{XY}) localized at the ends of the chromosomes (Figures 31,32), which is consistent with the already described punctuated distribution of recombination hotspots in common bean (Bhakta *et al.* 2015). Intra-species introgressions do not seem to be restricted to sub-tellomeric regions, however, these observations could be biased given that we conducted our introgression screenings using a synteny-based pseudoassembly and thus, actual distances between scaffolds and contigs may vary.

Even though gene flow is more frequent from domesticated genotypes towards wild subpopulations, which can be a consequence of selection against introgressants by farmers, there is still an important transfer of stress response genes from wilds into the domesticated genotypes that has probably accelerated the adaptation of cultivars to different environments along their dispersal in Mexico. At the same time, it was evidenced that while *P. vulgaris* plants growing in the Southern hemisphere can be cross-pollinated by their AHZ wild neighbours, this does not occur in the opposite direction, indicating that given the short period of time that separates the AHZ *Phaseolus* populations from their *P. vulgaris* relatives from MA and AN, reproductive barriers have not been fully established. These results agree with previous evidence of unsuccessful crosses between a genotype from Cajamarca, Peru (G21245), included in our sampling, and two tester populations from MA (G04830) and AN (G00122) (Koinange and Gepts, 1992). In the same report, other genotypes were crossed and the hybrid weakness was measured in the progeny; the accession G21245 was successfully crossed with 36% and 75% of the MA and AN genotypes in the test, respectively, showing an unbalanced reproductive barrier.

Previous studies have tried to identify the genetic source of incompatibility between the Andean and Mesoamerican gene pools, as certain crosses result in temperature dependent hybrid weakness associated with a severe root phenotype. It has been found that such phenotype is controlled by the interaction of the root- and shoot-expressed semidominant alleles dosage-dependent lethal 1 (DL1) and DL2, which communicate via long-distance signalling (Hannah *et al.* 2007). Biochemical data showed that root death likely occurs by defence-related programmed cell death, as indicated by salicylic acid accumulation. DL2-expressing cotyledons supply a potent inhibitory signal that is sufficient to cause such death in DL1-expressing roots. However, the differential introgression that we observe within *P. vulgaris* and between *P. vulgaris* and *P. pseudovulgaris*, might contribute to a more complete understanding of the genetic basis of reproductive isolation in the genus.

Differential shaping of the common bean genome by domestication and genomic introgression

Functional categories related to the emergence of domestication syndrome traits were identified, including the starch metabolism pathway, photoperiod sensitivity and vernalization related genes. Moreover, several lncRNAs also displayed haplotype clusters associated to the domestication phenotype. Recently, by means of differential expression analyses, a series of long intergenic ncRNAs, a class of transcripts known to play regulatory functions in animal systems (Wang *et al.* 2011), were suggested to play a role in pig domestication (Zhou *et al.* 2014). Nevertheless, it remains unclear at this point which are the targets of these ncRNAs with signature of selection, but it still opens the possibility that domestication traits could be also influenced by the action of transcriptional regulators rather than by alterations in protein sequences.

A remarkable observation is that domestication gene candidates seem to act as barrier loci as selective sweeps, i.e. regions that are rich in domestication-associated haplotypes, do not display signals of introgression. The functional description of introgressed loci from *P. pseudovulgaris* into *P. vulgaris* supports the concept of “genome permeability across species” discussed in section 2.3. If neutral loci move freely between populations, and this rate of introgression decreases when genomic markers are responsible for the phenotypic discrimination of the species (possibly under directional selection), the fact that genes involved in reproductive processes and organ development do not introgress from *P. pseudovulgaris* as frequently as they do between *P. vulgaris* subpopulations, suggest that there are more loci involved in the incompatibility of common bean accessions and the isolation of AHZ genotypes, beyond the described DI1/DI2 system (Hannah *et al.* 2007).

Spatio-temporal model of *Phaseolus* speciation in the Americas

Taking together the coalescent results, the metabolomic and phylogenomic data and the significant d_{XY} distance between AHZ accessions and *P. vulgaris* subpopulations discussed in previous sections, we have strong evidence pointing to an early speciation event in tropical Andes. Indeed, the Amotape-Huancabamba zone has been described as a transition zone between Northern and Central Andes, where climatic dynamics and oro-geographic conditions have produced the highest degree of plant species diversity and endemism along the Andean Mountains (Richter *et al.* 2009; Luebert and Weigend 2014; Mutke *et al.* 2014). Following this line of evidence, we propose two hypotheses that explain the divergence of *P. vulgaris* lineages in the MA, AHZ and AN regions (Figures 39, 40). First, we propose a two waved migration event in which *Phaseolus vulgaris* –or an ancestral form of the species- dispersed from Mesoamerica reaching the AH Zone in Central Andes where it remained isolated and suffered an allopatric speciation. Hundreds of thousands of years later, a small population of *P. vulgaris* with Mesoamerican genetic background invaded the Southern Andes, giving rise to the second gene pool that was latter accessible for domestication (Figure 39).

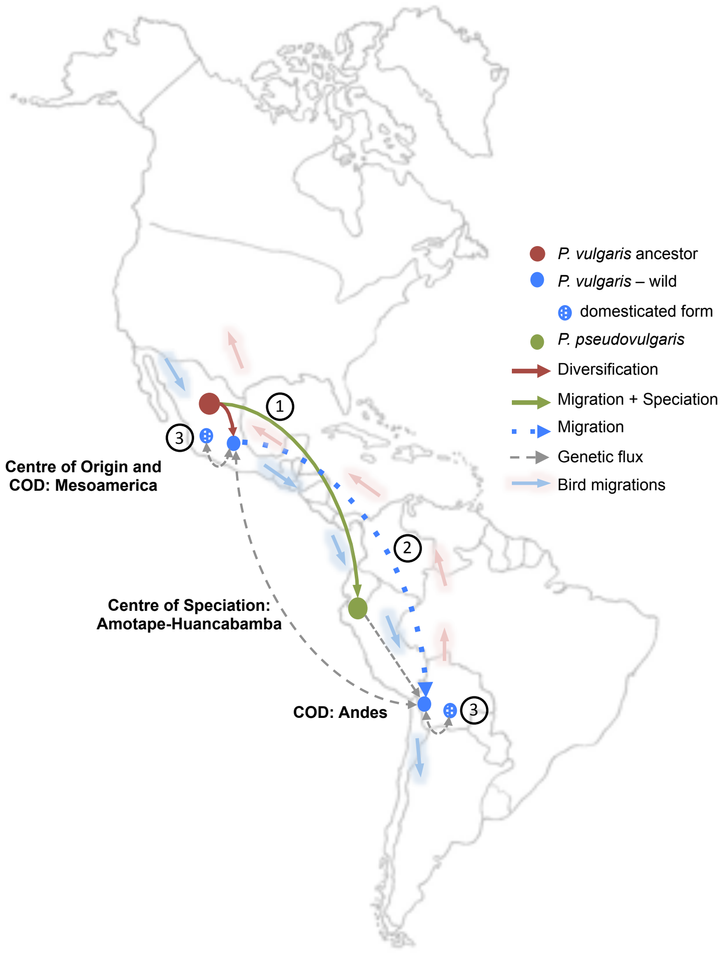


Figure 39. First spatio-temporal model of common bean migrations and lineage divergence in America. Two-waved model of dispersal mediated by bird migrations. Dispersal from MA to AH region, followed by speciation (1) predates the split of *P. vulgaris* lineages (2); domestication corresponds to the most recent evolutionary event (3).

It has been reported that birds, particularly from the genus Columba, eat common bean seeds while they are still inside the pods (Debouck *et al.* 1993). In this scenario, the convergent migration strategies of bird species moving within terrestrial regions in Central and South America could have promoted the dispersal of the seeds from Mesoamerica to the Andes, leaving the Amotape-Huancabamba deflection isolated for seed exchange. As demographic models have shown (La Sorte *et al.* 2015), terrestrial bird populations move along the narrow isthmus connecting North and South America and following the Andean corridor of mountains but do not necessarily reach the Western side of Peru and Ecuador, particularly when migrating from South to North during winter and spring, which could also contribute to the unidirectional flow from the AHZ population to its Andean neighbours.

A second spatio-temporal model, not incompatible with the first one, implicates glacial periods in Southern Andes during the Pleistocene. The fact that other *Phaseolus* species from the Vulgaris group can be found in Central America reaching Peru, suggests that seeds from this group could move after the formation of the Panama isthmus, following dispersal waves that have been dated for terrestrial organisms at 20 Mya and 6Mya (Bacon *et al.* 2015). The colonization of the Andean region by *Phaseolus* plants and its diversity could have been strongly affected by a glacial period spanning from 140Kya to 180Kya (reviewed by Hain *et al.* 2014), that matches the splitting of the MA and AN gene pools and the suggested bottleneck duration in the Andean wild population (Schmutz *et al.* 2014). After this glacial cycle, a small founder remaining *P. vulgaris* population in South America followed an exponential growth ending in the wild population that was afterwards domesticated (Figure 40).

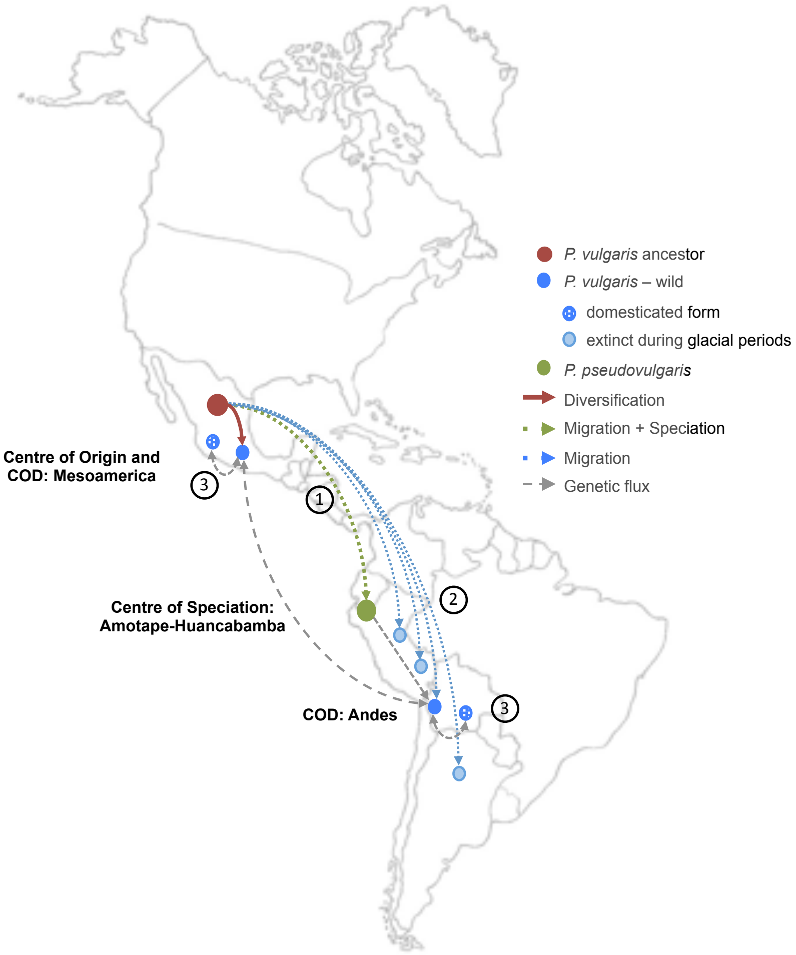


Figure 40. Second spatio-temporal model of common bean lineage divergence in America. Diversity extinction in the Southern hemisphere was caused by glacial periods. Migration from MA to AH region, followed by speciation (1) predates the split of *P. vulgaris* lineages (2); domestication corresponds to the most recent evolutionary event (3).

In both cases, fossil evidence would be needed to confirm the migration patterns, which has been impossible to obtain so far. Nevertheless, the emergence by allopatry of a cryptic species in the Amotape-Huancabamba zone should be considered as part of the evolutionary history of *P. vulgaris*.

9 Conclusions and perspectives.

Undoubtedly, genetic diversity is a necessary condition for further evolution in response to selection pressures and represents the raw material to develop improved breeds or cultivars (Gepts and Papa, 2002). Thus, systematic efforts to bring genetic diversity from wild relatives into crop plants to incorporate a wider range of useful adaptations are required in order to increase the resiliency and productivity of agriculture. The data herein produced allowed us to define introgression signals and differential haplotypes that, for the first time, can be combined to define domestication and putative adaptation loci. Our introgression screenings unveiled the capacity of a preferentially autogamous plant to outcross and fix loci from different populations even from more distant species. This capacity, as well as the transference of stress response genes as advantageous genes, should be exploited to accelerate breeding programs in the next few years by controlled hybridization strategies. In this regard, several points should be highlighted as future directions in the field.

It is crucial to preserve the diversity of common bean populations in the Mesoamerican area, particularly in Mexico, as it is the centre of origin with the highest richness of *Phaseolus* species. Recently, the National Centre for Genetic Resources (CNRG, INIFAP) received the largest collection of Mexican *Phaseolus* seeds kept at the International Centre for Tropical Agriculture (CIAT), that contains over 6000 samples of which, around 5000 are *P. vulgaris*, including 500 wild accessions. This collection is a tremendous genetic reservoir, as the passport information associated to each sample reflects the capacity of many of them to grow in a wide range of habitats and environmental conditions where cultivars are not adapted. Propagating this collection should be a priority as many of these populations have already been lost in their natural growing conditions. The generation of biological material from this reservoir should be accompanied by genotyping strategies (GBS or RAD-seq) in order to compile a complete catalogue of genetic variants associated to the passport information of the samples. Such sequencing efforts could be easily translated into target populations to be included in modern breeding programs.

Furthermore, the ecological description of wild *Phaseolus* species distributed in Mexico (López *et al.* 2005; Delgado-Salinas and Gama López, 2015) highlights other candidate species to be evaluated as potential genomic resources. For example, it was observed that *P. leptostachyus* (together with *P. coccineus*) grows in the widest range of climatic conditions; *P.*

leptostachyus, P. microcarpus, P. tuerckeimii and *P. pedicellatus* were observed in dry regions with high temperature; *P. tuerckheimii, P. lunatus* and *P. leptostachyus* were abundant in tropical climates with high degree of humidity (prone to fungal diseases). In terms of altitude adaptation, *P. acutifolius, P. filiformis, P. leptostachyus, P. lunatus, P. microcarpus* or *P. tuerckheimii* were collected from very low areas (sea-level), which contrast with *P. vulgaris* preference for altitude. In addition to *P. filiformis, P. maculatus* could be considered an alternative candidate to evaluate drought resistance, while *P. pauciflorus and P. pedicellatus* for cold adaptation.

The reduced number of individuals included in this study, particularly outside the Vulgaris clade, did not allow a detailed analysis of genomic introgressions outside common bean and its closest sister species. Thus, more individuals and populations from the above mentioned species should be also part of the sequencing efforts in the near future to explore both, the domestication signatures in other *Phaseolus* species and to identify the loci and pathways behind the adaptation to diverse environmental conditions. These genome wide association studies would facilitate the improvement of common bean, as well as of other bean species that are commonly consumed in the country.

The importance of regulatory elements and the remodelling of co-expression networks during adaptation and domestication is only starting to be understood in common bean. Given the availability of non-coding RNA predictions for *P. vulgaris* reference genomes, their association to domestication and adaptive traits should be carefully evaluated through transcriptional correlations between these regulators and their targets. For this purpose, extensive RNA-seq experiments need to be designed, including wild and domesticated cultivars at different developmental stages and subjected to particular stresses. Additionally, the fixation of polymorphisms as a consequence of population expansions and artificial selection could result in structural changes particularly in long non-coding RNAs. The consequences of such alterations will need to be evaluated in the context of environmental adaptation.

Ultimately, domestication gene candidates should be experimentally validated. Even though a few examples of successful transformation (Kwapata *et al.* 2012) and virus-based gene silencing (Diaz-Camino *et al.* 2011) have been reported, the recalcitrance of common bean for genetic transformation has made functional studies really challenging. The generation of mutant populations of *P. vulgaris* trough fast neutron radiation (O'Rourke *et al.* 2013) and

TILLING [targeted induced local lesions in genomes (Porch *et al.* 2009)] protocols has been used as an alternative for the identification of genes behind visual phenotypic differences. Thus, the combination of such strategies and genome re-sequencing of mutant plants could eventually facilitate the direct association of coding and non-coding loci to the emergence of domestication and adaptation traits.

From a genomic perspective, it is clear that important sequencing efforts are required in order to incorporate the impact of introgression in predictive genotype-phenotype models, which should be translated into powerful breeding strategies in the near future.

10 References

1. Abbott R, *et al.* Hybridization and speciation. J Evol Biol. 2013. 26:229-46.

2. Aguilar OM, Riva O, Peltzer E. Analysis of Rhizobium etli and of its symbiosis with wild *Phaseolus vulgaris* supports coevolution in centers of host diversification. Proc Natl Acad Sci USA. 2004. 101:13548-53.

3. Angioi SA, *et al*. Beans in Europe: origin and structure of the European landraces of *Phaseolus vulgaris* L. Theor Appl Genet. 2010. 121:829-43.

4. Acosta-Gallegos JA, Kelly JD, Gepts P. Prebreeding in common bean and use of genetic diversity from wild germplasm. Crop Science. 2007. 47: S44-S59.

5. Alexa A and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0. 2016.

6. Andueza-Noh RH, *et al*. Multiple domestications of the Mesoamerican gene pool of lima bean (*Phaseolus lunatus* L.): evidence from chloroplast DNA sequences. Genet Resour Crop Evol. 2013. 60:1069–86.

7. Arenas-Huertero C. *et al.* Conserved and novel miRNAs in the legume *Phaseolus vulgaris* in response to stress. Plant Mol. Biol*.* 2009. 70, 385-401.

8. Arumuganthan K, Earle E. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 1991. 9: 208-18.

9. Bacon CD, *et al*. Biological evidence supports an early and complex emergence of the Isthmus of Panama. Proc Natl Acad Sci U S A. 2015. 112:6110-5.

10. Bai L, Kim E, DellaPenna D, Brutnell TP. Novel lycopene epsilon cyclase activities in maize revealed through perturbation of carotenoid biosynthesis*.* Plant J*.* 2009. 59, 588-99.

11. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Curr Protoc Bioinformatics. 2007; Chapter 4:Unit 4.3.

12. Bhakta MS, Jones VA, Vallejos CE. Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris* L. PLoS One. 2015. 10:e0116822.

13. Beebe S, Toro O, González AV, Chacón MI, Debouck DG. Wild-weed-crop complexes of common bean (*Phaseolus vulgaris* L., Fabaceae) in the Andes of Peru and Colombia, and their implications for conservation and breeding. Genet Resour Crop Evol. 1997. 44: 73-91.

14. Beebe S, Rao IM, Blair M, Acosta-Gallegos JA. Phenotyping common beans for adaptation to drought. Front Physiol. 2013. 4:35.

15. Bellucci E, *et al*. Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. Plant Cell. 2014. 26:1901-12.

16. Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms. Ann. Bot. (London) 1995. 76**:** 113-16.

17. Bitocchi E, *et al*. Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Proc Natl Acad Sci USA. 2012. 109:E788-96.

18. Bitocchi E, *et al*. Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytol. 2013. 197:300-13.

19. Blair MW, *et al*. Development of a genome-wide anchored microsatellite map for common bean (*Phaseolus vulgaris* L.). Theor Appl Genet. 2002. 107:1362-74.

20. Blair MW, Iriarte G, Beebe S. QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (*Phaseolus vulgaris* L.) cross. Theor Appl Genet 2006.112:1149-63.

21. Blair MW, Soler A, Cortés AJ. Diversification and population structure in common beans (*Phaseolus vulgaris* L.). PLoS One. 2012. 7:e49488. [1]

22. Blair MW, Pantoja W, Carmenza Muñoz L. First use of microsatellite markers in a large collection of cultivated and wild accessions of tepary bean (*Phaseolus acutifolius* A. Gray). Theor Appl Genet. 2012. 125:1137-47. [2]

23. Bredeson JV, *et al.* Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat Biotechnol. 2016. 34:562-70.

24. Butare L, *et al*. New genetic sources of resistance in the genus *Phaseolus* to individual and combined aluminium toxicity and progressive soil drying stresses. Euphytica. 2011. 181:385–404.

25. Caetano-Anolles G, Crist-Estes DK, Bauer WD. Chemotaxis of *Rhizobium meliloti* to the plant flavone luteolin requires functional nodulation genes. J Bacteriol. 1988. 170: 3164-9.

26. Cannon SB, *et al.* Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proc Natl Acad Sci U S A. 2006. 103:14959–64.

27. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25:1972–3.

28. Cavanagh CR, *et al.* Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci USA. 2013. 110:8057-62.

29. Chacón MI, Pickersgill B, Debouck DG. Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. Theor Appl Genet. 2005.110:432-44.

30. Chambers MC, *et al.* A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012. 30: 918–20.

31. Cingolani P, *et al*. 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 (Austin). 2012. 6:80-92.

32. Córdoba JM, Chavarro C, Schlueter JA, Jackson SA, Blair MW. Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. BMC Genomics. 2010. 11:436.

33. Cortés AJ, Chavarro MC, Blair MW. SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet. 2011. 123:827-45.

34. Cortés AJ, Monserrate FA, Ramírez-Villegas J, Madriñán S, Blair MW. Drought tolerance in wild plant populations: the case of common beans (*Phaseolus vulgaris* L.). PLoS One. 2013. 8:e62898.

35. Counterman BA, Noor MA. Multilocus test for introgression between the cactophilic species Drosophila mojavensis and Drosophila arizonae. Am Nat. 2006. 168:682-96.

36. Croft D, *et al.* The Reactome pathway knowledgebase. Nucleic Acids Res. 2014. 42:D472–7.

37. Debouck DG, Toro O, Paredes OM, Johnson WC, Gepts P. Genetic diversity and ecological distribution of *Phaseolus vulgaris* in northwestern South America. Econ Bot 1993. 47:408-23

38. Delgado-Salinas A, Bibler R, Lavin M. Phylogeny of the genus *Phaseolus* (Leguminosae): a recent diversification in an ancient landscape. Syst Bot 2006. 31: 779-91.

39. Delgado-Salinas A, Gama López S. Diversidad y distribución de los frijoles silvestres en México. Revista Digital Universitaria. 2015. 16.

40. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature 2002. 418: 700-07.

41. Díaz-Camino C, Annamalai P, Sanchez F, Kachroo A, Ghabrial SA. An effective virus-based gene silencing method for functional genomics studies in common bean. Plant Methods. 2011. 7:16.

42. Ding Y, Tao Y, Zhu C. Emerging roles of microRNAs in the mediation of drought stress response in plants. J Exp Bot. 2013. 64, 3077-86.

43. Dowling T, Secor CL. The role of hybridization and introgression in the diversification of animals. Annu Rev Ecol Systemat 1997. 28: 593-619.

44. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7 Mol Biol Evol. 2012. 29: 1969-73

45. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011. 28:2239-52

46. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004. 5:113.

47. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008. 9:18.

48. Ellstrand N, Prentice H, Hancock J, Gene flow and introgression from domesticated plants into their wild relatives. Annu Rev Ecol Syst 1999. 30:539–63.

49. Escalante AM, Coello G, Eguiarte LE, Piñero D. Genetic structure and mating systems in wild and cultivated populations of *Phaseolus coccineus* and *P. vulgaris*. Am J Bot 1994. 81:1096-103.

50. Félix DT, Coello-Coello J, Martínez-Castillo J. Wild to crop introgression and genetic diversity in Lima bean (*Phaseolus* lunatus L.) in traditional Mayan milpas from Mexico. Conserv Genet. 2014. 15:1315-28.

51. Fernández F, Gepts P, López M. Stages of development of the common bean plant ed. Cali, Colombia: Centro Internacional De Agricultura Tropical (CIAT); 1986.

52. Ferreira J, *et al*. Gene flow in common bean (*Phaseolus vulgaris* L.). Euphytica. 2007. 153:165-70.

53. Flint-Garcia SA, Thornsberry JM, Buckler ES IV. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 2003. 54:357-74.

54. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011. 6:e16526.

55. Fonsêca A, *et al*. Cytogenetic map of common bean (*Phaseolus vulgaris* L.). Chromosome Res. 2010. 18:487-502.

56. García Mendoza EA. Guía técnica para el cultivo del fríjol. Iica-Red. 2009.

57. Garzon LN, Blair M. Development and mapping of SSR markers linked to resistance-gene homologue clusters in common bean. Crop J. 2014. 2:183–94.

58. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997. 14:685–95.

59. Geffroy V, *et al*. Identification of an ancestral resistance gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen *Colletotrichum lindemuthianum*. Mol Plant-Microbe Inter. 1999. 12: 774-84.

60. Geffroy V, *et al*. Inheritance ofpartial resistance against *Colletotrichum lindemuthianum* in *Phaseolus vulgaris* and co-localization of quantitative trait loci with genesinvolved in specific resistance. Mol Plant-Microbe Inter. 2000. 13: 287-96.

61. Gepts P. Origin and evolution of common bean: past events and recent trends. Hort Sci 1998. 33: 1124-30.

62. Gepts P, Papa, R. Evolution during domestication. In: Encyclopedia of Life Sciences. London: Nature Publishing Group 2002.

63. Gepts P. Crop Domestication as a Long-term Selection Experiment. Plant Breed Rev 2004. 24.

64. Gibb S, Strimmer K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. Bioinformatics 2012. 28: 2270–71.

65. Gioia T, *et al*. Evidence for introduction bottleneck and extensive inter-gene pool (Mesoamerica x Andes) hybridization in the European common bean (*Phaseolus vulgaris* L.) germplasm. PLoS One. 2013. 8:e75974.

66. Götz S, *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008. 36:3420–35.

67. Green RE, *et al.* A draft sequence of the Neandertal genome. Science. 2010. 328:710-22.

68. Griebel T, Marco-Sola S. GEM-Tools. Available at: https://github.com/ gemtools/gemtools. Accessed 5 Feb 2016.

69. Grisi MCM, *et al*. Genetic mapping of a new set of microsatellite markers in a reference common bean (*Phaseolus* vulgaris) population BAT93 x Jalo EEP558. Genet Mol Res. 2007. 6:691–706.

70. Guindon S, *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010. 59:307-21.

71. Guo X, *et al.* Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts. BMC Genomics. 2007. 8:228.

72. Guzmán P, *et al*. Characterization of Phaeoisariopsis griseola isolates by random amplified polymorphic DNA markers suggests pathogen coevolution with *Phaseolus vulgaris*. Phytopathology 1995. 85: 600-7.

73. Haghighi KR, Ascher PD. Fertile, intermediate hybrids between *Phaseolus vulgaris* and *P. acutifolius* from congruity backcrossing. Sex Plant Reprod. 1988. 1:51-8.

74. Hain MP, Sigman DM, Haug GH. Chap. 8.18: The biological pump in the past. In, Treatise on Geochemistry (Second Edition). Volume 8: The Oceans and Marine Geochemistry. Amsterdam, NL, Elsevier, 2014. 485-517.

75. Hajjar R, Hodgkin T. The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica. 2007. 156:1-13.

76. Hancock JF. Contributions of domesticated plant studies to our understanding of plant evolution. Ann Bot. 2005. 96:953–63.

77. Hannah MA, *et al.* Hybrid weakness controlled by the dosage-dependent lethal (DL) gene system in common bean (*Phaseolus vulgaris*) is caused by a shoot-derived inhibitory signal leading to salicylic acid-associated root death. New Phytol. 2007. 176:537-49.

78. Harrison RG, Larson EL. Hybridization, introgression, and the nature of species boundaries. J Hered. 2014. 105:795-809.

79. He Z, *et al*. Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genet. 2011. 7:e1002100.

80. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python environment for tree exploration. BMC Bioinformatics. 2010. 11:24.

81. Huerta-Cepas J, Gabaldón T. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics. 2011. 27:38-45.

82. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014. 42:D897-902.

83. Hufford MB, *et al*. The genomic signature of crop-wild introgression in maize. PLoS Genet. 2013. 9:e1003477.

84. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, *et al.* InterPro in 2011: New developments in the family and domain prediction database. Nucleic Acids Res. 2012. 40:D306-12.

85. Hurst LD. Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet. 2009. 10:83-93.

86. Jiao Y, *et al.* Ancestral polyploidy in seed plants and angiosperms. Nature. 2011. 473:97-100.

87. Johannsen W. The Genotype Conception of Heredity. Am Nat. 1911. 45: 129-59.

88. Jurka J, *et al*. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005. 110:462-7.

89. Kami J, Velásquez VB, Debouck DG, Gepts P. Identification of presumed ancestral DNA sequences of phaseolin in *Phaseolus vulgaris*. Proc Natl Acad Sci USA. 1995. 92:1101-04.

90. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012. 40:D109-14.

91. Kaplan L, Lynch TF, Smith CE Jr. Early Cultivated Beans (*Phaseolus vulgaris*) from an Intermontane Peruvian Valley. Science. 1973. 179:76-7.

92. Kaplan L, Lynch TF. *Phaseolus* (Fabaceae) in Archaeology: AMS Radiocarbon Dates and Their Significance for Pre-Colombian Agriculture. Econ Bot. 1999. 53:261-72.

93. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 2008. 9:286-98.

94. Kelly JD, Gepts P, Miklas PN, Coyne DP. Tagging and mapping of genes and QTL and molecular marker-assisted selection for traits of economic importance in bean and cowpea. Field Crops Res. 2003. 82:135-54.

95. Koenig D, *et al.* Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Natl Acad Sci USA. 2013. 110:E2655-62.

96. Koinange EMK, Gepts P. Hybrid weakness in wild *Phaseolus vulgaris* L. J Hered 1992. 83:135-9.

97. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014.15:356.

98. Kwak M, Paul Gepts P. Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet. 2009. 118:979-92.

99. Kwapata K, Nguyen T, Sticklen M. Genetic transformation of common bean (*Phaseolus vulgaris* L.) with the GUS color marker, the BAR herbicide resistance, and the barley (*Hordeum vulgare*) HVA1 drought tolerance genes. Int. J. Agron. 2012. 198960–67

100. La Sorte FA, Fink D, Hochachka WM, Kelling S. Convergence of broad-scale migration strategies in terrestrial birds. Proc Biol Sci. 2016. 283:1823.

101. Lassmann T, Frings O, Sonnhammer ELL. Kalign2: High-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res. 2009. 37:858–65.

102. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009. 25:1754-60.

103. Li H, *et al.* The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. Bioinformatics. 2009. 25:1–2.

104. Li X, Gruber MY, Hegedus DD, Lydiate DJ, Gao MJ. Effects of a coumarin derivative, 4-methylumbelliferone, on seed germination and seedling establishment in Arabidopsis. J Chem Ecol. 2011. 37:880-90.

105. Liang M, Nielsen R. Q & A: who is *H. sapiens* really, and how do we know? BMC Biol. 2011. 9:20.

106. Llaca V, Delgado-Salinas A, Gepts P. Chloroplast DNA as an evolutionary marker in the *Phaseolus vulgaris* complex. Theor Appl Genet. 1994. 88:646-52.

107. López Soto JL, Ruiz Corral JA, Sánchez González J, Lépiz Ildefonso R. Climatic adaptation of 25 wild bean species (*Phaseolus* spp) in México. Rev. Fitotec. Mex. 2005. 28: 221-30.

108. Lohmueller KE, *et al.* Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 2011. 7:e1002326.

109. Luebert F, Weigend M. Phylogenetic insights into Andean plant diversification. Front Ecol Evol. 2014. 2:27.

110. Lyons E, *et al*. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol. 2008. 148:1772-81.

111. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004. 20:2878-9.

112. Mallet J. Hybrid speciation. Nature. 2007. 446:279-83.

113. Mamidi S, *et al*. Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. Funct Plant Biol. 2011. 38:953–67.

114. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol. 2015 32:244-57.

115. Martínez-Castillo J, Zizumbo-Villarreal J, Gepts P, Delgado-Valerio P, Colunga-GarcíaMarín P. Structure and genetic diversity of wild populations of Lima bean (*Phaseolus lunatus* L.) from the Yucatan peninsula, Mexico. Crop Sci. 2006. 46:1071-80.

116. Martínez-Castillo J, Zizumbo-Villarreal J, Gepts P, Colunga-GarcíaMarín P. Gene Flow and Genetic Structure in the Wild–Weedy–Domesticated Complex of *Phaseolus lunatus* L. in its Mesoamerican Center of Domestication and Diversity. Crop Sci. 2007. 47:58–66.

117. Martínez-Castillo J, Camacho-Pérez L, Villanueva-Viramontes S, Andueza-Noh RH, Chacón-Sánchez MI. Genetic structure within the Mesoamerican gene pool of wild *Phaseolus lunatus* (Fabaceae) from Mexico as revealed by microsatellite markers: Implications for conservation and the domestication of the species. Am J Bot. 2014. 101:851-64.

118. McClean P, Gepts P, Kami J. Genomics and genetic diversity in common bean. In: Wilson RF, Stalker HT, Brummer EC (eds.), Legume crop genomics. AOCS Press, Champaign, IL. 2004. pp. 60-82.

119. Mensack MM, *et al*. Evaluation of diversity among common beans (*Phaseolus vulgaris* L.) from two centers of domestication using 'omics' technologies. BMC Genomics. 2010. 11:686.

120. Mercado-Ruaro P, Delgado-Salinas A. Karyotypic studies on species of *Phaseolus* (Fabaceae: Phaseolinae). Am J Bot. 1998. 85:1.

121. Micheletto S, *et al*. Comparative transcript profiling in roots of *Phaseolus acutifolius* and *P. vulgaris* under water deficit stress. Plant Science. 2007. 173:510–20.

122. Miedes E, Vanholme R, Boerjan W, Molina A. The role of the secondary cell wall in plant resistance to pathogens. Front Plant Sci. 2014. 5:358.

123. Montero-Vargas JM, *et al*. Metabolic phenotyping for the classification of coffee trees and the exploration of selection markers. Molecular BioSystems 2013. 9: 693–99.

124. Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nat Rev Genet 2011. 13:85-96.
125. Muñoz LC, Blair MW, Duque MC, Tohme J, Roca W. Introgression in common bean × tepary bean interspecific congruity-backcross lines as measured by aflp markers. Crop Sci. 2004. 44:637–45.

126. Mutke J, Jacobs R, Meyers K, Henning T, Weigend M. Diversity patterns of selected Andean plant groups correspond to topography and habitat dynamics, not orogeny. Front Genet. 2014. 5:351.

127. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: Inference of RNA alignments. Bioinformatics. 2009. 25:1335-7.

128. Newbler assembler. Available at: http://454.com/products/analysis-software/index.asp. Accessed 5 Feb 2016.

129. Ng JL, *et al*. Flavonoids and auxin transport inhibitors rescue symbiotic nodulation in the *Medicago truncatula* cytokinin perception mutant cre1. Plant Cell. 2015. 27:2210-26.

130. Nolte V, Schlötterer C. African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. Genetics 2008. 178, 405-12.

131. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302:205–17.

132. O'Rourke JA, *et al*. A re-sequencing based assessment of genomic heterogeneity and fast neutron-induced deletions in a common bean cultivar. Front Plant Sci. 2013. 4:210.

133. Ørom UA, *et al.* Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010. 143:46–58.

134. Papa R, Gepts P. Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. Theor Appl Genet. 2003. 106:239–50

135. Papa R, Acosta-Gallegos JA, Delgado-Salinas A, Gepts P. A genome-wide analysis of differentiation between wild and domesticated *Phaseolus vulgaris* from Mesoamerica. Theor Appl Genet. 2005. 111:1147-58.

136. Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007. 23:1061–7.

137. Pastorini J, Zaramody A, Curtis DJ, Nievergelt CM, Mundy NI. Genetic analysis of hybridization and introgression between wild mongoose and brown lemurs. BMC Evol Biol. 2009. 9:32.

138. Payró de la Cruz E, Gepts P, Colunga García-Marín P, Zizumbo Villareal D. Spatial distribution of genetic diversity in wild populations of *Phaseolus vulgaris* L. from Guanajuato and Michoacán, México. Genet Res Crop Evol. 2005. 52:589–99.

139. Peck MC, Fisher RF, Long SR. Diverse flavonoids stimulate NodD1 binding to nod gene promoters in *Sinorhizobium meliloti*. J Bacteriol. 2006. 188: 5417-27.

140. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011. 8:785–6.

141. Porch TG, *et al*. Generation fo a mutant population for TILLING common vean genotype BAT93. J Am HortSoc. 2009. 134:348-55.

142. Porch TG, *et al*. Use of wild relatives and closely related species to adapt common bean to climate change. Agronomy. 2013. 3:433-61.

143. Purcell S, *et al*. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007. 81: 559-75

144. Purugganan MD, Fuller DQ. The nature of selection during plant domestication. Nature 2009. 457:843-8.

145. Reddy P, Rendón-Anaya M, Soto del Rio MD, Khandual S. Flavonoids as Signaling Molecules and Regulators of Root Nodule Development. Dyn Soil Dyn Plant. 2007. 1:83-94.

146. Ribeiro RA, *et al*. Novel Rhizobium lineages isolated from root nodules of the common bean (*Phaseolus vulgaris* L.) in Andean and Mesoamerican areas. Res Microbiol. 2013. 164:740-8.

147. Richter M, Diertl KH, Emck P, Peters T, Beck E. Reasons for an outstanding plant diversity in the tropical Andes of Southern Ecuador. Landscape Online 2009. 12:1-35.

148. Rhoades MW *et al.* Prediction of plant microRNA targets. Cell. 2002. 110:513-20.

149. Rossi M, *et al*. Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. Evol Appl. 2009. 2:504-22.

150. SAGARPA. Estudio de gran vision y factibilidad económica y financier para el desarrollo de infraestructura de almacenamiento y distribyción de granos y oleaginosas para el mediano y largo plazo a nivel nacional. 2014. (http://www.sagarpa.gob.mx/agronegocios/documents/estudios_promercado/granos.pdf)

151. Sammeth M. Flux Capacitor. Available at: http://sammeth.net/confluence/display/FLUX/Home. Accessed 5 Feb 2016.

152. Sanseverino W, *et al.* PRGdb 2.0: Towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Res. 2013. 41:D1167–71.

153. Schmutz J, *et al.* Genome sequence of the palaeopolyploid soybean. Nature. 2010. 463:178–83.

154. Schmutz J, *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014. 46: 707-13.

155. Schneider JV, Schulte K, Aguilar JF, Huertas ML. Molecular evidence for hybridization and introgression in the neotropical coastal desert-endemic Palaua (Malveae, Malvaceae). Mol Phylogenet Evol. 2011. 60:373-84.

156. Servín-Garcidueñas LE, *et al*. Symbiont shift towards Rhizobium nodulation in a group of phylogenetically related *Phaseolus* species. Mol Phylogenet Evol. 2014. 79:1-11.

157. Sieber P, Wellmer F, Gheyselinck J, Riechmann JL, Meyerowitz EM. Redundancy and specialization among plant microRNAs: role of the MIR164 family in developmental robustness. Development. 2007. 134:1051-60.

158. Smit A, Hubley R, Green P. RepeatMasker Open-3.0. 1996. Available at: http://www.repeatmasker.org/. Accessed 5 Feb 2016.

159. Smith J, Kronforst MR. Do *Heliconius* butterfly species exchange mimicry alleles? Biol Lett. 2013. 9:20130503.

160. Smýkal P, Vernoud V, Blair MW, Soukup A, Thompson RD. The role of the testa during development and in establishment of dormancy of the legume seed. Front Plant Sci. 2014. 5:351.

161. Sotelo-Silveira, M, Chauvin AL, Marsch-Martínez N, Winkler R, De Folter S. Metabolic fingerprinting of Arabidopsis thaliana accessions. Front Plant Sci. 2015. 6:1-13.

162. Spataro G, *et al*. Genetic diversity and structure of a worldwide collection of *Phaseolus coccineus* L. Theor Appl Genet. 2011. 122:1281-91.

163. Stanke M, *et al*. AUGUSTUS: Ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006. 34:W435–9.

164. Stasolla C, Katahira R, Thorpe TA, Ashihara H. Purine and pyrimidine nucleotide metabolism in higher plants. J Plant Physiol. 2003. 160:1271-95.

165. Steinbiss S, Willhoeft U, Gremme G, Kurtz S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 2009. 37:7002–13.

166. Swanson-Wagner R, *et al*. Reshaping of the maize transcriptome by domestication. Proc Natl Acad Sci USA. 2012. 109:11878-83.

167. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012. 485:635-41.

168. Tsugawa H, *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015. 12:523–26.

169. Vlasova A, *et al.* Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biol. 2016. 17:32.

170. Vijayan P, *et al*. Capturing cold-stress-related sequence diversity from a wild relative of common bean (*Phaseolus angustissimus*). Genome. 2011. 54:620-8.
171. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 2006. 34:1692–9.

172. Wang KC, *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011. 472:120-4.

173. Weissmann S, Feldman M, Gressel J. Sequence evidence for sporadic intergeneric DNA introgression from wheat into a wild Aegilops species. Mol Biol Evol. 2005. 22:2055-62.

174. Whitney KD, Randell RA, Rieseberg LH. Adaptive introgression of herbivore resistance traits in the weedy sunflower Helianthus annuus. Am Nat. 2006. 167:794-807.

175. Williams GJ. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!). 2011 aed. Springer. 2011.
http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1441998896.

176. Winkler R. An evolving computational platform for biological mass spectrometry: workflows, statistics and data mining with MASSyPup64. Peer J. 2015. 3 (e1401): 1–34.

177. Worthington M, Soleri D, Gepts P. Genetic composition and spatial distribution of farmer-managed *Phaseolus* bean plantings: an example from a village in Oaxaca, Mexico. Crop Sci. 2012. 52:1721-35

178. Xu H, Guan Y. Detecting local haplotype sharing and haplotype association. Genetics. 2014. 197:823-38.

179. Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol Biol Evol. 2015. 32:2001–14.

180. Yoo MJ, Wendel JF. Comparative evolutionary and developmental dynamics of the cotton (Gossypium hirsutum) fiber transcriptome. PLoS Genet. 2014. 10:e1004073.

181. Yu Y. *et al.* Independent losses of function in a polyphenol oxidase in rice: differentiation in grain discoloration between subspecies and the role of positive selection under domestication. Plant Cell. 2008. 20:2946-56.

182. Yu N. *et al.* Temporal control of trichome distribution by microRNA156-targeted SPL genes in *Arabidopsis thaliana*. Plant Cell. 2010. 22:2322-35.

183. Yuste-Lisbona FJ, *et al*. Marker-based linkage map of Andean common bean (*Phaseolus vulgaris* L.) and mapping of QTLs underlying popping ability traits. BMC Plant Biol. 2012. 12:136.

184. Zhou ZY, *et al.* Genome-wide identification of long intergenic noncoding RNA genes and their potential association with domestication in pigs. Genome Biol. Evol. 2014. 6:1387-92.

185. Zizumbo-Villarreal D, *et al* P. Population structure and evolutionary dynamics of wild-weedy-domesticated. Crop Sci. 2005. 45:1073–83.

Apendix A.

Supplementary Table 1. Plant material.

Species	Clade	Group	ID	Country	State/ Province	Gene pool	Altitude (m)	100 seed weight (g)
P. vulgaris	B	Vulgaris	G24392	Mexico	Jalisco/ Arandas	MA	2020	5.0
P. vulgaris	B	Vulgaris	G24377	Mexico	Michoacan/ Cojumatlan	MA	1700	4.0
P. vulgaris	B	Vulgaris	G50368	Mexico	Oaxaca/ Tlacolula De Matamoros	MA	1480	6.8
P. vulgaris	B	Vulgaris	G23551	Mexico	Sinaloa/ Concordia	MA	710	4.2
P. vulgaris	B	Vulgaris	G12967	Mexico	Jalisco/ Ayutla	MA	-	5.0
P. vulgaris	B	Vulgaris	G23463	Mexico	Chihuahua/ Yepachic	MA	1530	9.4
P. vulgaris	B	Vulgaris	G23550	Mexico	Zacatecas/ Moyahua De E.	MA	1700	6.3
P. vulgaris	B	Vulgaris	G24594	Mexico	Chiapas/ Ixtapa	MA	1400	5.4
P. vulgaris	B	Vulgaris	G23556	Mexico	Durango/Durango	MA	1860	8.5
P. vulgaris	B	Vulgaris	Negro San Luis	Mexico	-	MA		> 20.0
P. vulgaris	B	Vulgaris	Faba	Spain	-	AND		100.0
P. vulgaris	B	Vulgaris	Jalo G09603	Brazil	Minas Gerais/ Patos De Minas	AND	-	41.0
P. vulgaris	B	Vulgaris	G19901	Argentina	Tucuman/ El Mollar	AND	1900	9.0
P. vulgaris	B	Vulgaris	G21244	Peru (1)	Cajamarca/ San Pablo	AH	2020	9.0
P. vulgaris	B	Vulgaris	G21245	Peru (2)	Cajamarca/ San Miguel	AH	1790	10.5
P. vulgaris	B	Vulgaris	G23587	Peru (3)	Cajamarca/ Chota	AH	1250	9.0
P. vulgaris	B	Vulgaris	G23724	Ecuador (1)	Loja/ Macará	AH	960	11.0
P. vulgaris	B	Vulgaris	G23582	Ecuador (2)	Chimborazo/ Alausi	AH	1710	8.4

P. costaricensis	B	Vulgaris	G40811A	Costa Rica	Cartago/ Cartago		1510	18.8
P. dumosus	B	Vulgaris	G36043	Mexico	Chiapas/ Ixtapa		1680	70.3
P coccineus	B	Vulgaris	-	Mexico	Querétaro/ La Joya		-	-
P. maculatus	B	Polystachios	PL-8841	Mexico	-	-	-	-
P. polystachios	B	Polystachios	G40782	Mexico	-	-		8.4
P. leptostachyus	B	Leptostachyus	PL-8829	Mexico	-	-	-	-
P. filiformis	B	Filiformis	-	Mexico	Baja California Sur	-	-	-
P. lunatus	B	Lunatus	PL-8834	Mexico	-	-	-	-
P. acutifolius	B	Vulgaris	-	Mexico	Chiapas	-	-	-
P. hintonii	A	Tuerckheimii	-	Mexico	Edo. Mexico	-	-	-
P. microcarpus	A	-	PL-8844	Mexico	-	-	-	-

Supplementary Table 2. Genome coverage.

Group	Species	Reads (e^6)	Theoretical Coverage[a]	Depth of Coverage[b]	Breath of Coverage[c]
Vulgaris	*P. dumosus*	359	55.2x	27.5x	85%
	P. costarricencis	317	48.8x	29.4x	87%
	P. acutifolius	242	37.2x	19.6x	60%
	P. coccineus	285	43.8x	31.5x	88%
	P. vulgaris (G24392)	90	13.8x	6.4x	91%
	P. vulgaris (G24377)	142	21.8x	9.4x	94%
	P. vulgaris (G50368)	91	14.0x	7.5x	92%
	P. vulgaris (G23556)	170	26.2x	22.4x	94%
	P. vulgaris (NegroSanLuis)	129	19.8x	18.1x	96%
	P. vulgaris (G23551)	173	26.6x	20.2x	94%
	P. vulgaris (G12967)	155	23.8x	18.3x	94%
	P. vulgaris (G23463)	198	30.5x	17.9x	95%
	P. vulgaris (G23550)	121	18.6x	12.4x	90%
	P. vulgaris (G24594)	342	52.6x	25.0x	95%
	P. vulgaris (Faba Andecha)	226	34.8x	28.6x	93%
	P. vulgaris (G19901)	127	19.5x	14.1x	93%
	P. vulgaris (G09603)	274	42.2x	36.6x	98%
	P. vulgaris (G21244)	338	52.0x	48.6x	92%
	P. vulgaris (G21245)	249	38.3x	34.3x	92%
	P. vulgaris (G23587)	299	46.1x	37.4x	93%
	P. vulgaris (G23724)	236	36.3x	15.7x	67%
	P. vulgaris (G23582)	189	29.1x	17.9x	92%
Leptostachyus	*P. leptostachyus*	258	39.7x	22.0x	77%
Lunatus	*P. lunatus*	200	30.8x	14.8x	68%
Polystachios	*P. polystachios*	293	45.1x	19.7x	67%
	P. maculatus	192	29.6x	15.4x	66%
Filiformis	*P. filiformis*	131	20.1x	15.7x	67%
Tuerckhemii	*P. hintonii*	182	28.0x	11.8x	59%
ND	*P. microcarpus*	190	29.3x	12.8x	56%

[a] Calculated over 650 Mb of theoretical genome length.

[b] Calculated over 556.4 Mb of BAT93 genome assembly.

[c] Calculated over the evelen synteny-based pseudo-assembled linkage groups after 'N' removal.

Supplementary Table 3. Detected metabolites.

Name	m/z	Ionization	Exact Mass	Formula	RT	Ions
Trigonelline	138.05	[M+H]+	138.0555	C7H7NO2	0.33005	100 most abundant
Threonine	120.07	[M+H]+	120.0655	C4H9NO3	1.00735	100 most important
4-Methylumbelliferone	177.05	[M+H]+	177.0551	C10H8O3	2.236917	100 most important
D-Sorbitol 6-phosphate	263.05	[M+H]+	263.0532	C6H15O9P	0.3301	100 most important
Luteolin	287.04	[M+H]+	287.0555	C15H10O6	3.486033	100 most important
Kaempferol	287.04	[M+H]+	287.055	C15H10O6	3.295167	100 most important
Luteolin	287.04	[M+H]+	287.0555	C15H10O6	2.168317	100 most important
Kaempferol	287.04	[M+H]+	287.055	C15H10O6	5.202183	100 most important
Luteolin	287.04	[M+H]+	287.0555	C15H10O6	2.549283	100 most important
all-trans-Retinoic acid	301.20	[M+H]+	301.2167	C20H28O2	13.29615	100 most important
L-beta-Homotryptophan	219.09	[M+H]+	219.1133	C12H14N2O2	2.63535	100 most intense
Biotin	245.08	[M+H]+	245.0954	C10H16N2O3S	2.5326	100 most intense
Quercetin	303.03	[M+H]+	303.0504	C15H10O7	2.956717	100 most intense
Myricetin	319.03	[M+H]+	319.0449	C15H10O8	2.236917	100 most intense
alpha-Tocotrienol	425.37	[M+H]+	425.3419	C29H44O2	10.24525	100 most intense
Daidzin	439.12	[M+Na]+	439.0999	C21H20O9	0.3301	100 most intense
Kaempferol-7-neohesperidoside; LC-ESI-QTOF; MS2; [M+H]+; CE	595.15	[M+H]+	595.1658	C27H30O15	3.5719	100 most intense
Pelargonin	595.16	[M]+	595.1663	C27H31O15	2.168317	100 most intense
Isorhamnetin	317.06	[M+H]+	317.0656	C16H12O7	3.676733	30 most important
Coumarin	147.04	[M+H]+	147.0441	C9H6O2	2.549283	whole profile
D-Ala-D-ala	161.09	[M+H]+	161.0926	C6H12N2O3	2.635258	whole profile
Umbelliferone	163.03	[M+H]+	163.039	C9H6O3	14.83854	whole profile
DL-5-Hydroxylysine	163.10	[M+H]+	163.1082	C6H14N2O3	4.542092	whole profile
Nicotine	163.14	[M+H]+	163.1235	C10H14N2	11.59702	whole profile
Chalcone	209.11	[M+H]+	209.0966	C15H12O	2.253883	whole profile
gamma-Glutamylleucine	261.21	[M+H]+	261.2	C11H20N2O5	13.88493	whole profile
Genistein	271.05	[M+H]+	271.0601	C15H10O5	4.7327	whole profile
Cinchonine	295.18	[M+H]+	295.1805	C19H22N2O	14.45722	whole profile
Sinapoyl malate	341.09	[M+H]+	341.0872	C15H16O9	3.016733	whole profile
gamma-Tocotrienol	411.35	[M+H]+	411.3263	C28H42O2	17.04082	whole profile
Ononin	431.12	[M+H]+	431.1337	C22H22O9	3.016733	whole profile
Ideain	449.09	[M]+	449.1078	C21H21O11	2.549283	whole profile
Cyanidin-3-glucoside	449.10	[M]+	449.1084	C21H21O11	2.168317	whole profile
Kaempferol-3-Glucuronide	463.08	[M+H]+	463.0876	C21H18O12	3.31235	whole profile
Isoquercitrin	465.09	[M+H]+	465.1028	C21H20O12	2.046483	whole profile
Hyperoside	465.09	[M+H]+	465.1028	C21H20O12	3.016733	whole profile
Quercetin-3-Glucuronide	479.07	[M+H]+	479.0826	C21H18O13	2.91385	whole profile
Procyanidin B2	579.16	[M+H]+	579.1502	C30H26O12	2.618217	whole profile

Cyanidin-3-O-(2''-O-beta-xylopyranosyl-beta-glucopyranoside)	581.17	[M]+	581.1506	C26H29O15	3.016733	whole profile
Reserpine	609.26	[M+H]+	609.2806	C33H40N2O9	13.4009	whole profile
Cyanidin-3,5-di-O-glucoside	611.15	[M]+	611.1612	C27H31O16	2.168317	whole profile
Rutin	611.15	[M+H]+	611.1606	C27H30O16	3.016733	whole profile
Kaempferol-3-O-rutinoside	617.13	[M+Na]+	617.1477	C27H30O15	3.5719	whole profile
Quercetin-3-O-alpha-L-rhamnopyranosyl(1-2)-beta-D-glucopyranoside-7-O-alpha-L-rhamnopyranoside	757.21	[M+H]+	757.2191	C33H40O20	2.168317	whole profile

Apendix B.

Contributions

The project herein described, particularly at its initial phase where we produced the reference genome of *P. vugaris* BAT93, was enclosed in a large multidisciplinary consortium, PhasIbeAm (http://www.cyted.org/es/node/4569) that involved several research groups from Mexico, Spain, Brazil and Argentina. The contribution of each group is described below.

Section 6.2.1: Reference genome

Dr. Heinz Himmelbauer and colaborators (Centre for Genomic Regulation, Barcelona, Spain) carried out the assembly of all versions of the reference genome that was ultimately released in its version v.10.

Dra. Rosana P. Vianello-Brondani (EMBRAPA Rice and Beans, Biotechnology Laboratory, Santo Antônio de Goiás, Brazil) provided biological material of a F5 generation of an inbred line BAT93xJalo EEP558 for chromosome anchoring.

Section 6.2.2: BAT93 transcritptional atlas

Dr. Roderic Guigó and his research group (Centre for Genomic Regulation, Barcelona, Spain) were in charge of the differential expression analyses during sequential developmental stages, as well as the quantification of gene isoforms per organ.

Dr. Miguel Angel Hernández Oñate (LANGEBIO) contributed to the functional analysis of genes with preferential and specific expression.

Dra. Marta Santalla (Mision Biológica de Galicia (MBG)-National Spanish Research Council, CSIC. Pontevedra, Spain) provided all biological samples for RNA-seq experiments.

Section 6.2.3: Functional annotation and repeat detection

Dr. Tyler Alioto (Centro Nacional de Análisis Genómico, Parc Científic de Barcelona, Barcelona, Spain) contributed to the gene model prediction and functional annotation of BAT93.

Section 6.2.4: Non-coding RNA analysis

Dr. Cedric Notredame (Centre for Genomic Regulation, Barcelona, Spain) identified the set of lncRNA in BAT93.

Luca Cozzuto (Centre for Genomic Regulation, Barcelona, Spain) defined the set of small noncoding RNAs encoded in the reference BAT93 reference genome.

Section 6.3.1: Phylome

Dr. Toni Gabaldón and his team (Institució Catalana de Recerca i Estudis Avançats, ICREA) developed the Phylome database for common bean that can be accessed trough: http://phylomedb.org/

Section 6.5: Metabolomic profile

Dr. Robert Winker and Josaphat Montero (Departamento de Biotecnología y Bioquímica, Cinvestav Unidad Irapuato) produced the metabolomic profile of *P. vulgaris* accessions.

List of publications

Articles derived from this work

1. Vlasova A*, Capella-Gutiérrez S*, **Rendón-Anaya M***, Hernández-Oñate M, Minoche AE, Erb I, Câmara F, Prieto-Barja P, Corvelo A, Sanseverino W, Westergaard G, Dohm JC, Pappas GJ Jr, Saburido-Alvarez S, Kedra D, Gonzalez I, Cozzuto L, Gómez-Garrido J, Aguilar-Morón MA, Andreu N, Aguilar OM, Garcia-Mas J, Zehnsdorf M, Vázquez MP, Delgado-Salinas A, Delaye L, Lowy E, Mentaberry A, Vianello-Brondani RP, García JL, Alioto T, Sánchez F, Himmelbauer H, Santalla M, Notredame C, Gabaldón T, Herrera-Estrella A, Guigó R. Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biology. 2016 Feb 25;17:32. doi: 10.1186/s13059-016-0883-6. **Equal contributors*.

2. **Rendón-Anaya M**, Montero-Vargas JM, Saburido-Álvarez S, Vlasova A, Capella-Gutierrez S, Ordaz-Ortiz JJ, Aguilar OM, Vianello-Brondani RP, Santalla M, Delaye-Arredondo L, Gabaldón T, Gepts P, Winkler R, Guigó R, Delgado-Salinas A, Herrera-Estrella A. Genomic history of the origin and domestication of common bean unveils its closest sister species. 2016. *Submitted*.

Additional articles published during my PhD training

1. Nakamura Y, Paetz C, Brandt W, David A, **Rendón-Anaya M**, Herrera-Estrella A, Mithöfer A, Boland W. Synthesis of 6-substituted 1-oxoindanoyl isoleucine conjugates and modeling studies with the COI1-JAZ co-receptor complex of lima bean. J Chem Ecol. 2014 Jul;40(7):687-99. doi: 10.1007/s10886-014-0469-2.

2. Ortiz E, **Rendón-Anaya M**, Rego SC, Schwartz EF, Possani LD. Antarease-like Zn-metalloproteases are ubiquitous in the venom of different scorpion genera. Biochim Biophys Acta. 2014 Jun;1840(6):1738-46. doi: 10.1016/j.bbagen.2013.12.012.