# Adaptive gene loss in the common bean pan-genome during range expansion and domestication

Cortinovis, G.[1*], Vincenzi, L.[2*], Anderson, R.[3], Marturano, G.[2], Marsh, J.I.[3], Bayer, P.E.[3], Rocchetti, L.[1], Frascarelli, G.[1], Lanzavecchia, G.[1], Pieri, A.[1], Benazzo, A.[4], Bellucci, E.[1], Di Vittori, V.[1], Nanni, L.[1], Ferreira Fernández, J.J.[5], Rossato, M.[2, 6], Aguilar, O.M.[7], Morrell, P.L.[8], Rodriguez, M.[9], Gioia, T.[10], Neumann, K.[11], Alvarez Diaz, J.C.[12], Gratias-Weill, A.[12], Klopp, C.[13.], , Geffroy, V.[12], Bitocchi, E.[1], Delledonne, M. [2, 6*], Edwards, D.[3*], Papa, R [1*#].

[1] Department of Agricultural, Food and Environmental Sciences, Marche Polytechnic University, 60131 Ancona, Italy

[2] Department of Biotechnology, University of Verona, 37134 Verona, Italy

[3] Centre for Applied Bioinformatics and School of Biological Sciences, University of Western Australia, Perth, WA 6009, Australia

[4] Department of Life Sciences and Biotechnology, University of Ferrara, 44100 Ferrara, Italy

[5] Regional Service for Agrofood Research and Development (SERIDA), Ctra AS-267 PK 19, 33300 Asturias, Spain

[6] Genartis s.r.l., 37126 Verona, Italy

[7] Institute of Biotechnology and Molecular Biology, UNLP-CONICET, CCT La Plata, La Plata, Argentina

[8] Department of Agronomy and Plant Genetics, University of Minnesota, 55108-6026, St. Paul Minnesota, USA

[9] Department of Agriculture, University of Sassari, 07100 Sassare, Italy

[10] School of Agricultural, Forestry, Food and Environmental Sciences, University of Basilicata, 85100 Potenza, Italy

[11] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Seeland, Germany

[12] CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), Univ Evry, University Paris-Saclay, Orsay 91405, France

[13] INRAE, Genotoul Bioinformatics Platform, Applied Mathematics and Informatics of Toulouse, Sigenae, MIAT UR875, Castanet Tolosan, France

[*] These authors contributed equally to this work

[#] Corresponding authors

## Abstract

The common bean (*Phaseolus vulgaris* L.) is an important grain legume crop [1,2] whose life history offers an ideal evolutionary model to identify adaptive variants suitable for breeding programs [3]. Here we present the first common bean pan-genome based on five high-quality genomes and whole-genome reads representing 339 genotypes. We found ~243 Mb of additional sequences containing 7,495 protein-coding genes missing from the reference, constituting 51% of the total presence/absence variations (PAVs). There were more putatively deleterious mutations in PAVs than core genes, probably reflecting the lower effective population size of PAVs as well as fitness advantages due to the purging effect of gene loss. Our results suggest pan-genome shrinkage occurred during wild range expansion from Mexico to South America, with more PAV loss per individual in Andean vs Mesoamerican populations. Selection during wild spreading and domestication was also associated with PAV loss involved in important adaptive traits. Our findings provide evidence that partial or complete gene loss was a key adaptive trait leading to localized and genome-wide reductions. This departs from established paradigms and reveals how evolutionary forces shape gene composition within plant genomes. The common bean pan-genome offers a valuable resource for legume research and breeding, climate change mitigation, and sustainable agriculture.

## Main

Food legumes provide valuable genetic resources to address agriculture-related societal challenges, including climate change mitigation, biodiversity conservation, and the need for sustainable agriculture and healthy diets [4-7]. The common bean (*Phaseolus vulgaris* L.) is a diploid (2n=2x=22) and predominantly self-pollinating annual grain legume crop with a prominent role in agriculture and broader societal importance [1,2]. It is also a useful model of crop evolution [3] reflecting the parallel and independent life history of two geographically isolated and genetically differentiated gene pools (Mesoamerican and Andean) following the expansion of wild from Mexico to South America ~150,000–200,000 years ago, an order of magnitude earlier than its dual domestication [8-11]. Previous studies using a single reference genome have provided insights into the population structure of the common bean [12] and the genetic basis of important adaptive traits [13]. However, pan-genomic diversity must be explored to gain a comprehensive understanding [14-17]. We therefore constructed the first *P. vulgaris* pan-genome using a non-iterative approach and investigated its genetic variation in terms of PAVs within a representative panel of genetically and phenotypically well-characterized accessions. This publicly available common bean pan-genome provides a valuable starting point to identify genes and genomic mechanisms affecting adaptation and will accelerate legume improvement.

## Characterization of the common bean pan-genome

To generate the common bean pan-genome, we applied a non-iterative approach to five high-quality *de novo* genome assemblies of wild and domesticated genotypes and incorporated short-read whole genome sequencing (WGS) data from 339 representative common bean accessions, comprising 33 wild and 306 domesticated forms. This revealed ~242 Mb of additional sequence containing 7,495 genes missing from the reference genome. The new sequences account for 22% of all discovered genes, with 9% (3,040 genes) derived from the high-quality genomes and the remaining 13% (4,455 genes) from the panel of 339 WGS genotypes. The final size of the reconstructed pan-genome was ~780 Mb, with 34,928 predicted protein-coding genes (**Supplementary Table 1a**).

The new reference pan-genome was used for both variant calling and PAV calling (**Supplementary Table 1b**). We detected 23,343,365 variant sites, 19,002,047 of which were classified as single-nucleotide variants (SNVs) and 4,341,318 as insertions/deletions (InDels). Following PAV calling, the categorization of all 34,928 predicted genes by frequency unveiled that 58% of the pan-genome consists of core genes found across all lines (20,369 genes), with the remaining 42% comprising PAVs. These PAVs are either partially shared among accessions or exclusive to a single genotype, totalling 14,559 genes (**Supplementary Table 1c**). Notably, 51% of these PAVs originate from non-reference regions (NRRs), representing sequences absent in the reference genome. The growth curve related to the size calculation suggested a closed pan-genome. In agreement, the pan-genes reached the saturation point (99%, 34,579 genes) and remained constant without substantial increase when the number of accession genomes exceeded 120. In contrast, the size of the core genes decreased with each added genotype (**Fig. 1a**). This indicates that the final pan-genome includes almost all the gene content of *P. vulgaris*. Gene Ontology (GO) enrichment analysis showed that the core genes are significantly enriched for terms associated with homeostatic (GO:0042592) and catabolic (GO:0043632) processes (**Supplementary Fig. 1a; Supplementary Table 1d**) whereas the PAVs are significantly enriched for terms related to defence (GO:0006952), responses to external stimuli (GO:0009605), responses to light (GO:0019684), and reproduction (GO:0000003, GO:0022414) (**Supplementary Fig. 1b; Supplementary Table 1e**).

To investigate the evolution of the core genes and PAVs, we calculated the non-synonymous and synonymous ratio (Ka/Ks) for each gene in each accession (**Supplementary Table 1f**). This revealed a statistically significant difference ($p < 2.2 \times 10^{-16}$), with the PAVs including a greater number of harmful variants relative to benign variants when compared to the core genes (**Supplementary Fig. 1c; Supplementary Table 1g**). When we split the PAVs into three subcategories based on their frequency (*soft-core* 0.90 ≤ freq. < 1; *accessory* 0.10 ≤ freq. < 0.90; and *rare* freq. < 0.10), we observed a significant increase ($p = 0.048$) in the proportion of putative harmful variants among the rare genes compared to the soft-core genes (**Fig. 1b; Supplementary Table 1g**). These results may reflect the lower effective population size of the PAVs (reducing the efficiency of purifying selection) and/or the higher fitness gain from purging genes that have accumulated deleterious mutations (loss-of-function mutations).

## Evolutionary trajectory of the common bean

The common bean is characterized by three eco-geographic gene pools. The two major ones are the Mesoamerican (M) and Andean (A) populations, which encompass both wild and domesticated forms. The third originates from Northern Peru/Ecuador (PhI) and has a relatively narrow distribution of only wild individuals [11]. The Mesoamerican and Andean gene pools include five domesticated subgroups (M1, M2, A1, A2 and A3) corresponding to the Jalisco-Durango, Mesoamerica, Nueva Granada, Peru, and Chile races [13]. We constructed a neighbour-joining (NJ) phylogenetic tree (**Fig. 2a**) and conducted PAV-based principal component analysis (PCA) (**Fig. 2b**), both of which confirmed this well-defined population structure. Both analyses further divided the M1/Jalisco-Durango races into two clusters that we named cluster A and cluster B, respectively. The analysis of variance conducted on M1/Jalisco-Durango accessions, considering the first component for the days to flowering, revealed that cluster A is significantly later-flowering than cluster B (**Fig. 2c; Supplementary Table 2a**). The Jalisco (cluster A) and Durango (cluster B) races are therefore genetically distinguishable based on photoperiod sensitivity. This outcome also confirmed that our pan-genome enhances the characterization of genetic diversity and improves its analysis, exploitation and management. Cumulatively, the first and the second principal components of the PAV-based PCA explained 47.2% of the total variance, where PC1 mainly defined the differences between Mesoamerican and Andean gene pools while PC2 split the groups and subgroups within each gene pool (**Fig. 2b**). The NJ tree further underscored the suitability of core genes for phylogenetic reconstruction because they mitigate biases arising from the absence of genetic material among the compared accessions. In contrast to the tree based on single-nucleotide polymorphisms (SNPs) located on PAVs (**Supplementary Fig. 2a**), the NJ tree based solely on core SNPs properly grouped the wild PhI accession close to the wild Mesoamerican genotypes originating from Guatemala and Costa Rica (**Fig. 2a**), which are most closely related to the PhI gene pool [11].

When we examined the total number of PAVs per genetic group (**Supplementary Table 2b**), we found that the wild Mesoamerican and Andean populations have more PAVs compared to their domesticated counterparts (**Fig. 2d**). This supports the well-established notion that domestication is usually associated with a reduction of genetic diversity. The amplification of gene loss in domesticated common bean could reflect a classic bottleneck effect rather than natural selection [18]. We also found that the M1/Jalisco-Durango and A2/Peru races have more PAVs than the other subgroups in the same gene pool (**Fig. 2d**). This was supported by nucleotide diversity analysis applied to the 1,451,663 core SNPs (**Fig. 2e**; **Supplementary Table 2c**) and agrees with a recent hypothesis proposing that the M1/Durango-Jalisco and A2/Peru races were the first domesticated Mesoamerican and Andean populations from which the M2, A1 and A3 races arose during a secondary domestication phase [13].

To study inter-gene-pool hybridization, the PAV matrix for American domesticated accessions was analysed by using Fisher's exact test to compare the Mesoamerican and Andean populations. We found 5,556 PAVs (65% of the total) with a statistically significant difference in frequency ($p$ < 0.05) between the two gene pools. These included 778 diagnostic PAVs, 91% (707) of which were fixed in the Mesoamerican gene pool and 9% (71) in the Andean gene pool (**Supplementary Table 2d**). GO enrichment analysis applied to the 778 diagnostic genes revealed enrichment in processes related to detoxification (GO:0098754), metabolism (GO:0008152), and responses to stimuli (GO:0050896) (**Supplementary Fig. 2b**). Interestingly, none of these PAVs were found to be diagnostic between gene pools in Europe (**Supplementary Table 2d**), and when Fisher's exact test was applied to the subset of 114 European accessions, we did not detect any diagnostic genes between the Mesoamerican and Andean gene pools (**Supplementary Table 2e**). These outcomes clearly reflect the extensive inter-gene-pool hybridization in European germplasm and confirm its key role in the adaptation of common bean to new agricultural environments [13].

To investigate the influence of PAVs on important trait variations and identify candidate genes associated with them, we conducted a PAV-based genome-wide association study (GWAS) involving 218 American and European domesticated genotypes. We identified 39 significative association events correlated with day-to-flowering and photoperiod sensitivity, previously detailed in [13]. These associations were linked to 35 potential candidate PAVs, highlighting their likely involvement in regulating floral transition (**Supplementary Table 2f**). An interesting example is the GWAS peak associated with flowering time and photoperiod sensitivity located on Phvul.003G185200 (Chr03:40,838,810-40,850,729). This PAV demonstrates orthology to the *HDA5* gene in *Arabidopsis thaliana*, which displays deacetylase activity. Notably, *A. thaliana* mutants with impaired HDA5 expression patterns manifest late-flowering phenotypes due to the up-regulation of two floral repressor genes, namely FLOWERING LOCUS C (FLC) and MADS AFFECTING FLOWERING 1 (MAF1) [19]. It is noteworthy that common bean genotypes lacking the PAV Phvul.003G185200 exhibit early flowering phenotypes compared to those accessions carrying the gene (**Supplementary Fig. 2c**), implying an adaptive role correlated to the loss. Furthermore, we found that 9 out of the 35 candidate PAVs for GWA analysis show signature of selection in various comparisons, specifically, two PAVs between wild and domesticated Mesoamerican populations and seven PAVs between wild and domesticated Andean populations. Overall, although the majority (59%) of the candidate PAVs were located on the reference genome, a significant 41% were situated on the NRRs (**Supplementary Table 2f**), reaffirming the efficacy of the pan-genome in identifying functional variants associated with economically or evolutionarily important traits.

## Pan-genome shrinkage during wild expansion to South America

One of the most striking outcomes we observed was the difference in pan-genome size between the Mesoamerican and Andean gene pools (**Fig. 3a**). We calculated the total number of PAVs per individual and found that accessions from the same gene pool clustered together in separate groups, with Mesoamerican accessions featuring more PAVs per accession than Andean ones (**Fig. 3b, c; Supplementary Table 3a**). One possible explanation is that this reduction in pan-genome size may simply reflect genetic drift and the two sequential bottlenecks that occurred solely in the Andean population [12]. To better understand the roles of different evolutionary forces in shaping the PAV content of the Mesoamerican and Andean gene pools, and to distinguish between the effects of adaptation, population demography and history, we first focused on the analysis of wild accessions. Considering a panel of wild genotypes representing the entire geographical distribution in Latin America, we applied bivariate fit analysis and found a significant correlation ($p < 0.0001$) between the number of PAVs per individual and the latitude. Analysis of variance in which wild individuals were grouped by latitude followed by spatial interpolation revealed a significant and progressive loss of genes ranging from the accessions of Northern Mexico to those of Northwestern Argentina (**Fig. 4a, b; Supplementary Table 3b**). Furthermore, $F_{ST}$ analysis on PAVs comparing Mesoamerican and Andean wild populations may suggests the occurrence of selection for gene loss during wild range expansion (**Fig. 4c**). We found that 64% of all candidate PAVs in the top 5% of the $F_{ST}$ distribution (Fst≥0.85) are absent in the wild Andean gene pool. This high rate of gene loss in the Andean population significantly exceeds the gene loss rate observed in the entire variable genome (26%), demonstrating a more than twofold increase. This difference in gene loss rates was statistically validated using bootstrap resampling, strongly suggesting that gene loss during the process of wild differentiation was not a random occurrence but evident outcome of selective forces (**Supplementary Fig. 3a, b**). Moreover, functional annotation of the candidate PAVs revealed the enrichment of genes involved in pollen germination, innate immunity, abiotic stress tolerance, and root hair growth, indicating a potential adaptive role (**Supplementary Table 3c**). Overall, our findings suggest that selective pressure favouring the loss of genes involved in adaptive mechanisms, coupled with the influence of genetic drift resulting from the founder effect, may have contributed to the shrinking of the Andean pan-genome during wild differentiation.

## Footprints of selection for gene loss during domestication

The PAVs putatively shaped by selection during domestication in Mesoamerica and the Andes revealed further evidence that gene loss underpinned the successful adaptation of the American common bean. $F_{ST}$ analysis was applied to PAVs in wild and domesticated forms (separately for each gene pool) with only PAVs in the top 5% of the $F_{ST}$ distribution considered as candidates. We found 460 PAVs potentially under selection in the Mesoamerican population ($F_{ST} \geq 0.31$) and 514 in the Andean population ($F_{ST} \geq 0.27$) (**Supplementary Table 4a, b**). Functional annotation of the candidate PAVs revealed the enrichment of genes associated with domestication syndrome and adaptive traits such as dormancy, floral transition, light acclimation, defence, and symbiotic interactions (**Supplementary Table 4c, d**). Importantly, the candidate Phvul.003G265200 (Chr03: 50,365,995-50,368,501) is orthologous to 11 members of the plant Rho GTPase subfamily (ROP), including *ROP6* encoding a small Rho-like GTP binding protein. This GTPase subfamily is required for symbiotic interactions [20-22], and in the plasma membrane of *Lotus japonicus* cells it interacts directly with NOD FACTOR RECEPTOR 5, one of two nodulation factor receptors essential for nodule formation during symbiosis [23]. From our analysis, Phvul.003G265200 is a putative selected PAV ($F_{ST}$ = 0.50) for the Mesoamerican gene pool whose frequency declined by more than 60% during progression from the wild (0.94) to the domesticated (0.25) population (**Supplementary Table 4a**). Overall, no significant differences were observed in terms of absences between the wild and domesticated populations of both gene pools. However, a significant proportion of PAVs putatively under selection, specifically 63% (289 genes) in the Mesoamerican population (**Fig. 5a**) and 80% (411 genes) in the Andean one (**Fig. 5b**), were less frequent in domesticated than wild populations. When considering all PAVs, the percentage of PAVs with lower frequencies in domesticated populations fell significantly to 22% ($p < 2.2 \times 10^{-16}$) for the Mesoamerican gene pool and 29% ($p < 2.2 \times 10^{-16}$) for the Andean one (**Fig. 5a, b**). These findings suggest that selection during domestication resulted in gene loss, but unlike the range expansion of wild populations, it did not completely abolish the selected genes. This may reflect the different evolutionary timescales involved: wild differentiation occurred ~150,000 years ago whereas domestication was much more recent at ~8,000 years ago. These findings are consistent with previous observations that selection during the domestication of common bean in Mesoamerica has directly affected the transcriptome, leading to a ~20% decrease in gene expression levels attributed to loss-of-function mutations [18]. We also detected 29 PAVs with high $F_{ST}$ values in common between the Mesoamerican and the Andean gene pools, and these are mainly associated with the tryptophan metabolic pathway. Tryptophan holds significance as a precursor in secondary metabolism, contributing to the synthesis of essential molecules like auxin, serotonin, and melatonin. These compounds play diverse roles in plant physiology, influencing processes such as seed germination, root development, senescence, and flowering. Additionally, they contribute to the plant's response mechanisms against both biotic and abiotic stresses [24]. We analysed their frequencies and found that ~86% in both gene pools decreased in frequency during the progression from wild to domesticated accessions (**Supplementary Table 4e**). This may indicate a pattern of genomic convergence for some key adaptive genes between the Mesoamerican and Andean populations during their parallel domestication events.

## Conclusions

The global economic and social importance of the common bean means that pan-genomic analysis could boost the conservation and exploitation of its genetic resources to address key challenges in agriculture and wider society [1-6]. The genotypes selected for this study encompass both wild and domesticated forms, ensuring that the pan-genome comprehensively captures the extensive genetic variation within this species. This publicly accessible tool serves as a valuable resource for studies in population genetics, functional genomics, and plant breeding. PAV analysis provided insight into the evolutionary dynamics of pan-genome adaptation, including putative signals of selection for complete gene loss during wild differentiation between the Mesoamerican and Andean gene pools, contributing to the smaller pan-genome of the Andean population. We also identified putative selection footprints for partial gene loss during both domestication processes in Mesoamerica and the Andes, causing localized reductions in gene frequencies in domesticated populations compared to their wild counterparts. Candidate genes that have been completely or partially lost appear to be involved in important adaptive mechanisms, such as flowering time, symbiosis, biotic and abiotic stress tolerance, and root hair growth.

The significant role of reductive genome evolution in adaptation is now widely recognized [25-27]. For instance, in contrast to their European native counterparts, invasive genotypes exhibited a reduced genome size resulting in phenotypic effects that enhanced the species' invasive potential. This included an accelerated early growth rate driven by a negative correlation between genome size and the rate of stem elongation [28]. Similarly, Díez et al. [29] documented the genome size variations within the *Zea mays* species during the post-domestication process, revealing that maize landraces have significantly smaller genome sizes compared to their closest wild relatives, the teosintes. Notably, a negative correlation between maize genome size and altitude was observed [29]. Moreover, gene loss is considered functionally equivalent to other loss-of-function mutations, such as premature stop codons, providing an important source of adaptive phenotypic diversity [30-33]. A notable example is found in *A. thaliana*, where loss of function mutations in the *SCARECROW* (*SCR*) and/or SnRK2-type protein kinase (*SRK*) genes underlie the switch from obligate outcrossing (self-incompatibility) to self-fertilization [34]. This transition is widely recognized as one of the predominant evolutionary shifts among flowering plants, allowing the successful colonization of oceanic islands, an ecological principle known as Baker's rule. Accordingly, under the influence of specific and diverse agro-ecological pressures, the relinquishment of particular genes can confer a selective advantage on a given population. Overall, our findings support the "less is more" hypothesis in which non-functionalization is a common adaptive response [35]. This may be relevant when populations face selective pressure resulting from radical environmental changes, such as the expansion of wild common bean from the warm and humid climate of Central Mexico to higher and cooler altitudes in the Andes. Our results therefore establish a novel scenario in which evolutionary forces drive partial or complete gene loss due to selective pressure favouring adaptation rather than responses to stochastic events only. Mutations are more likely to cause a loss rather than a gain of function, so adaptive gene loss provides a rapid evolutionary response to environmental changes. This has profound implications for strategies to mitigate climate change. The common bean pan-genome is a valuable starting point that will lead to a deeper understanding of the genetic variations and genome dynamics responsible for key adaptive traits in food legumes.

## Methods

### Sources of genetic diversity

The pan-genome was constructed from five high-quality genomes representing the Mesoamerican and Andean gene pools. The *P. vulgaris* reference genome Pvulgaris_442_v2.0 (PV442) was downloaded from Phytozome (https://phytozome-next.jgi.doe.gov/), the genomes of BAT93 and JaloEPP558 were provided by the INRAE Institute, and the genomes of MIDAS and G12873 were sequenced and assembled *de novo* specifically for this study (**Supplementary Table 5a**). We also integrated 339 representative low-coverage WGS common bean genotypes, including 220 domesticated and 10 wild accessions from previous studies

[11, 13]. The remaining 109 accessions were multiplied in the greenhouse, and DNA extracted from young leaves was used for sequencing (**Supplementary Table 5b**).

### Plant growth and DNA extraction

MIDAS and G12873 single seed descent (SSD) genotypes were multiplied in the greenhouse. For both samples, 2 g of young leaf material was collected after 48 h of dark treatment and high-molecular-weight (HMW) DNA was extracted as previously described [36]. DNA quality was evaluated according to Oxford Nanopore Technologies (ONT) requirements. Specifically, purity was assessed using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific), the concentration was determined using a dsDNA Broad Range Assay Kit with Qubit 4.0 (Thermo Fisher Scientific), and the fragment size (≤ 400 kb) was determined using the CHEF Mapper electrophoresis system (Bio-Rad Laboratories). Fragments < 25 kb were removed using the Short Reads Eliminator kit (Circulomics) leaving 75% of the DNA from the MIDAS samples and 95% from the G12873 samples.

*P. vulgaris* genotypes of BAT93 and JaloEEP558 were sowed in soil and grown in a growth chamber at 23°C, 75% humidity, and 8 h dark and 16 h light photoperiods under fluorescent tubes (166lE). Young trifoliate leaves of BAT93 and JaloEEP558 genotypes were collected and flash-frozen with liquid nitrogen. Three days before sampling, plants were dark treated to optimize the high molecular weight DNA extraction. In addition, 109 SSD accessions were multiplied in the greenhouse and young leaves were collected in silica gel for drying and subsequent DNA extraction using the DNeasy 96 Plant kit (Qiagen) according to the manufacturer's instructions. For each sample, 50–70 mg of dried leaf material was pulverized with a Tissue-Lyser II (Qiagen) at 30 Hz for 6 min. The DNA quality and quantity were evaluated using a NanoPhotometer NP80 (Implen), and the concentration was determined using a Qubit BR dsDNA assay kit (Thermo Fisher Scientific).

### Sequencing low-coverage WGS accessions

DNA libraries for all samples were prepared using a KAPA Hyper Prep kit and PCR-free protocol (Roche). For each genotype, 200 ng of DNA was fragmented by sonication using a Covaris S220 device (Covaris) and WGS DNA libraries were generated using a 0.7–0.8× ratio of AMPureXP beads for final size selection. Libraries were quantified using the Qubit BR dsDNA assay kit and equimolar pools were quantified by real-time PCR against a standard curve using the KAPA Library Quantification Kit (Kapa Biosystems). Libraries were sequenced on the NovaSeq 6000 Illumina platform in 150PE mode, producing 15–30 million fragments per sample.

### Sequencing and assembly of MIDAS and G12873 genomes

Following quality control and priming according to ONT specifications, libraries were sequenced on a MinION device with a SpotON flow cell (FLO-MIN106 R9.4.1-Rev D). Two libraries were prepared for each genotype according to the SQK-LSK109 ligation sequencing protocol (ONT) with minor adjustments. Each library was loaded twice, and the flow cell was washed using the Flow Cell Wash Kit (ONT). Illumina PCR-free libraries were prepared starting with 1 ug of fragmented gDNA using the KAPA Hyper prep protocol. This process involved extending the adapter ligation time up to 30 minutes and conducting post-clean-up size selection using 0.7X AMPure XP beads. The library's concentration and size distribution were assessed using a Bioanalyzer 2100 in combination with high-sensitivity DNA reagents and chips. Sequencing was performed on a NovaSeq6000 instrument to generate 150-bp paired-end reads. MIDAS and G12873 whole-genome assemblies were generated using the nanopore-based approach based on 26 Gb (50-fold coverage) and 36 Gb (69-fold coverage), respectively. Raw nanopore reads were corrected using Canu v2.1 [37] and the resulting corrected reads were assembled *de novo* using wtdbg2 v2.5 [38]. Draft assemblies were refined by iterative polishing using long reads (Racon v1.4.3 and Medaka v1.0.3) [39] and short reads (three rounds of Pilon v1.23) [40]. Completeness was assessed using BUSCO v4.1.2 [41] and the Fabales_odb10 dataset (**Supplementary Table 5c**).

### Sequencing and assembly of BAT93 and JaloEEP558 genomes

High molecular weight DNA of BAT93 and JaloEEP558 genotypes was sequenced with a PacBio Sequel II system by GENTYANE platform (INRAE Clermont-Ferrand, France). A total of 21.09 and 29.35 Gb of PacBio HiFi reads were generated from BAT93 and JaloEEPP558, respectively. PacBio HiFi reads were *de novo* assembled into contigs using HiFiasm (v 0.9.0) with default parameters [42].

### Orthologous/paralogous analysis and clustering threshold settings

To incorporate two distinct populations, namely the Andean and the Mesoamerican gene pools, into the pangenome, precise differentiation between orthologous and paralogous genes is imperative. Consequently, a meticulous strategy was devised to ensure the preservation of solitary orthologous gene copies along with all paralogous counterparts. The relationship between orthologous genes was calculated with minimap2 v2.17 [43] to align the MIDAS and G12873 genome assemblies using the open reading frames (ORFs) of 2,330 complete single-copy BUSCO (Benchmarking Universal Single-Copy Orthologs) genes selected from the *P. vulgaris* reference genome PV442 (**Supplementary Table 5d**). The percentage identity was calculated for each ORF based on the number of matches in the alignments as a proportion of ORF length. The relationship between paralogous genes was calculated using the three most abundant gene families (OG0000273, OG0000328, and OG0000085) in the *P. vulgaris* PV442 reference genome, composed of 26, 37, and 42 genes, respectively. An all-versus-all comparison between the members of the same family was implemented using minimap2. The percentage of  identity was calculated for each gene family by dividing the number of matches in the alignments by the reference gene ORF length and then averaging the identity percentages for each family. Finally, the results of both tests were used to establish a clustering threshold of 90% to retain only one orthologous and all paralogous genes in the pan-genome (**Supplementary Table 5e**).

**Pan-genome construction**

We used a paired genome alignment strategy for pan-genome construction [44]. The PV442 reference genome was independently mapped onto MIDAS, G12873, BAT93 and JaloEPP558 with minimap2 v2.17 using the alignment pre-set *-x asm5*, which considers regions with an average divergence < 5%. The bam files of the four alignments were converted to delta files and structural variants were called using Assemblytics v 1.2.1 [45]. Only deletions were selected as NRRs [44]. Uncovered contigs private to the four analysed genomes were identified by applying samtools depth v1.1 [46] to the bam files and were extracted as NRRs. Deletions and uncovered contigs were independently filtered for a minimum length of 1 kb and clustered using CD-HIT-EST v4.8.1 [47] with a sequence identity of 90%, as described above for the orthologous and paralogous genes. To ensure that the NRRs identified through this method didn't encompass orthologous genes already existing in the PV442 reference genome, we specifically employed highly conserved BUSCO genes. We conducted a comparative analysis between the full complement of 4,947 MIDAS and 4,812 G12873 BUSCO genes present in PV442 and the NRRs derived from MIDAS and G12873 using BLASTp. Illumina data representing the 339 low-coverage WGS common bean accessions were trimmed with fastp v0.21.0 [48] and aligned to the preliminary pan-genome using bowtie2 v2.3.5.1 [49] with default parameters. The unmapped reads from these alignments were extracted using samtools v1.11, assembled using MaSuRCA v3.4.2 [50] with default parameters, and clustered using CD-HIT-EST v4.8.1 with a sequence identity of 90%. Finally, the NRRs derived from the panel of 339 common bean accessions were added to the preliminary pan-genome to generate the final pan-genome. To exclude putative contaminants and/or organelle sequences, NRRs were compared to the NCBI non-redundant nucleotide database using BLASTn, considering a minimum 80% identity and 25% coverage, leading to the removal of 1,194 sequences. Overall, we identified 64,174 added sequences, 86% of which reflected the mapping of the 339 low-coverage WGS accessions. The remaining 14% was identified by comparing the reference genome independently with the other four high-quality genomes (**Supplementary Table 1a**).

**Pan-genome annotation**

Repetitive sequences were identified and soft-masked using RepeatModeler v2.0.2 [51] and RepeatMasker v4.1.2-p1 [52]. Protein-coding genes were identified using a hybrid-approach prediction with Augustus v3.3.3 [53]. Proteins from *P. vulgaris* and correlated species (*Medicago truncatula* and *Glycine soja*) plus RNA-Seq data (unpublished data from [18]) were aligned to the genome and used as extrinsic evidence.

Protein sequences were aligned with Hisat2 v2.2.1 [54] and RNA-Seq data were aligned using Genome Threader v1.7.1 [55]. BUSCO genes in the Fabales_odb10 database were used to train the model for the Augustus predictor. Predicted genes were scanned with InterProScan v5.46-81.0 [56] for the presence of protein domains. Using a custom script, genes with transposon-related domains were filtered out and retained in the annotation if they contained at least one known protein domain. The filtered proteins were compared to the pan-genome with BLASTp v2.12.0 [57] and filtered by the best hits. The predicted genes were clustered with the proteins of all species considered in the annotation using OrthoFinder v2.5.4 [58]. Finally, functional annotation was achieved by integrating information about orthologous genes and by identifying functional domains using a custom script.

### PAV calling

Illumina data representing the 339 low-coverage WGS accessions were aligned to the pan-genome using bowtie2 v2.3.5.1 and the coverage of each predicted gene was calculated for each accession using samtools v1.11 (**Supplementary Table 5f**). PAV calling thresholds were defined for each accession according to the minimum coverage value of 1000 randomly selected BUSCO genes (ORFs). The BUSCO genes are orthologous genes that should be present in all considered accessions, but a few underrepresented genes in a given accession could constitute a bias. To avoid this, values below 1% (the 10 least covered genes) were discarded. The identified genes were classified based on their frequency as core genes if present in all the accessions or PAVs if partially shared or private to a single genotype (**Supplementary Table 1b, c**).

### Pan-genes and core genes size calculation

The curves describing the pan-genome and core genome sizes were evaluated by considering 1,000 random orders of the 339 genotypes. The orders were chosen randomly among all possible permutations (n! where n=[1,339]). For each ordering, the gene sets of the accessions were progressively added to the total genome size without considering the genes already present in the total set. The same procedure was applied for the core genome size, but the gene sets were intersected when each genome was added, thus keeping only the genes in common for each iteration (**Supplementary Table 5g, h**).

### Variant calling

SNVs and InDels were called with bcftools v1.10.2 [59] based on the alignment of 339 accessions with the pan-genome using bowtie2 v2.3.5.1. We used bcftools mpileup v1.10.2 to generate a genotype likelihood table, and variants were identified using bcftools call v1.10.2 and the pileup table, producing the final vcf file.

### Non-synonymous and synonymous mutations

The Ka/Ks ratio was computed for each gene in each accession using KaKs calculator v2.0 [60]. For each gene, the consensus sequence of each accession was extracted using bcftools consensus v1.10.2. The calculator compares the pan-genome gene sequence with the gene sequence of each accession to identify non-synonymous and synonymous variants and then computes the ratio. The calculator reported *NA* when there were no variants in a specific accession or when the denominator of the Ka/Ks ratio was zero. It was possible to compute the analysis for 30,850 of 34,928 genes. Sometimes the length of one of the two compared sequences was not divisible by three so the sequence could not be read in triplets (**Supplementary Table 1f**). The average Ka/Ks value per gene was used to assess the significance of the sample median (**Supplementary Table 1g**).

**Data analysis**

Pan-genome analysis focused on a representative subset of 99 well-characterized accessions among the original 339, including wild and American domesticated forms. In some cases, we also analysed the subset of 114 European domesticated accessions (**Supplementary Table 5b**).

For GO enrichment, the annotated core genes and PAVs in the pan-genome were analysed using the *buildGOmap* R function to infer indirect annotations and generate data suitable for *clusterProfiler* [61, 62]. Diagnostic genes were analysed using Metascape [63]. *A. thaliana* orthologs were identified using OrthoFinder [58] and by comparing all protein sequences from *P. vulgaris* (v2.1) and *A. thaliana* (TAIR10).

For PCA, the PAV matrix (1/0) was analysed using the *logisticPCA* package in R [64].

ANOVA within subgroup M1 was carried out using the first principal component related to days-to-flowering and photoperiod sensitivity [13] as a representative phenotypic trait.

$F_{ST}$ analysis involved the separate testing of PAVs in the Mesoamerican and Andean gene pools by comparing the frequency of each PAV between wild and domesticated forms. Each PAV was considered as a single locus (0/1) and $F_{ST}$ was calculated by applying the formula $F_{ST}$ = (H total – H within) / H total, where H is the heterozygosity [65]. The same procedure was applied to wild accessions when comparing the Mesoamerican and Andean gene pools. Only PAVs in the top 5% of the $F_{ST}$ distribution were considered as candidates.

The functions of interesting PAVs and those associated with *A. thaliana* orthologs detected by OrthoFinder [58] were investigated manually in the NCBI database (https://www.ncbi.nlm.nih.gov/).

Phylogenetic analysis involved the extraction and filtering of SNVs located in core genes and PAVs using bcftools [59], resulting in two final datasets: 1,451,663 SNPs for the core genes and 338,475 SNPs for the PAVs. The datasets were used to calculate the genetic distance between individuals and compute maximum composite likelihood values with 1000 bootstraps for the NJ tree in MEGA11 [66]. The final trees were visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

The filtered dataset of SNPs in core genes was also used to quantify the genetic diversity within groups of accessions by estimating $\pi$. The --*window-pi* vcftools flag was used to obtain measures of nucleotide diversity in 250-kb windows. The windowed-*pi* estimates were then divided by the total number of SNPs to calculate a global estimate for each genetic group.

Fisher's exact test with the false discovery rate corrected for multiple comparisons was applied in R to identify PAVs that differed significantly in frequency between the Mesoamerican and Andean gene pools for the American and European accessions.

The phenotypic data used for PAV-based GWAS encompassed the flowering time and photoperiod sensitivity data previously analysed by Bellucci et al. [13]. GWA analysis was run by using both the Mixed Linear Model (MLM) [67] and the Fixed and random model Circulating Probability Unification (FarmCPU) [68] implemented in the R package GAPIT v3 [69]. The threshold for each Genome Wide Association (GWA) scan was determined by the Bonferroni corrected *p* value at α = 0.05. The kinship matrix (IBS method) was calculated with Tassel 5 [70] and the population structure (at K2 obtained from Bellucci et al. [13]) were included into the models as fixed factors. Quantile-quantile (Q-Q) plots were obtained by plotting the observed -log10(*p* values) against the expected -log10(*p* values) under the null hypothesis of no association.

## Data Availability

The 109 raw sequencing reads generated and analyzed in this study have been deposited in the Sequence Read Archive (SRA) of the National Center of Biotechnology Information (NCBI) under BioProject number PRJNA1042929. Additional data comprising 230 raw sequencing reads have been sourced from Frascarelli et al. [11] and Bellucci et al. [13]. The pan-genome assembly and its annotation are publicly accessible via this link: https://doi.org/10.6084/m9.figshare.24573874.

## References

1. Broughton, W.J., Hernandez, G., and Blair, M.W. (2003) Beans (*Phaseolus* spp.) - Model food legumes. *Plant and Soil*, 252(1):55-128. DOI: https://doi.org/10.1023/A:1024146710611

2. Myers, J.R. and Kmiecik, K. (2017). Common bean: economic importance and relevance to biological science research. In: Pérez de la Vega, M., Santalla, M., and Marsolais, F. (eds) *The Common Bean Genome*. Compendium of Plant Genomes. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-63526-2_1

3. Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M.L., Gioia, T., Santo, D., Nanni, L., Attene, G., Papa, R. (2017) Beans (*Phaseolus ssp*.) as a model for understanding crop evolution. *Frontiers in Plant Science,* 8:722. DOI: https://doi.org/10.3389/fpls.2017.00722

4. Intergovernmental Panel on Climate Change (IPCC) (2019). Climate Change and Land. Retrieved from https://www.ipcc.ch/srccl/

5. Gerten, D., Heck, V., Jägermeyr, J., Bodirsky, B.L., Fetzer, I., Jalava, M., Kummu, M., Lucht, W., Rockström, J., Schaphoff, S. et al. (2020) Feeding ten billion people is possible within four terrestrial planetary boundaries. *Nature Sustainability*, 3(3):200-208. DOI: https://doi.org/10.1038/s41893-019-0465-1

6. Bellucci, E., Mario Aguilar, O., Alseekh, S., Bett, K., Brezeanu, C., Cook, D., De la Rosa, L., Delledonne, M., Dostatny, D.F., Ferreira, J.J., et al. (2021) The INCREASE project: Intelligent Collections of food-legume genetic resources for European agrofood systems. *Plant Jour*nal, 108:646-660. DOI: https://doi.org/10.1111/tpj.15472

7. Cortinovis, G., Oppermann, M., Neumann, K., Graner, A., Gioia, T., Marsella, M., Alseekh, S., Fernie, A. R., Papa, R., Bellucci, E. et al.(2021). Towards the development, maintenance, and standardized phenotypic characterization of single-seed-descent genetic resources for common bean. *Current Protocols*, 1, e133. DOI: https://doi.org/10.1002/cpz1.133

8. Schmutz, J., McClean, P.E., Mamidi, S., Wu, G.A., Cannon, S.B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., et al. (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics, 46:707-713. DOI: https://doi.org/10.1038/ng.3008

9. Bitocchi, E., Nanni, L., Bellucci, E., Rossi, M., Giardini, A., Spagnoletti Zeuli, P., Logozzo, G., Stougaard, J., McClean, P., Attene, G., et al. (2012) Mesoamerican origin of the common bean (*Phaseolus vulgaris L.*) is revealed by sequence data. *Proceedings of the National Academy of Sciences,* 109(14):E788-E796. DOI: https://doi.org/10.1073/pnas.1108973109

10. Bitocchi, E., Bellucci, E., Giardini, A., Rau, D., Rodriguez, M., Biagetti, E., Santilocchi, R., Spagnoletti Zeuli, P., Gioia, T., Logozzo, G., et al. (2013) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytologist*, 197:300-313. DOI: https://doi.org/10.1111/j.1469-8137.2012.04377.x

11. Frascarelli, G., Galise, T.R., D'Agostino, N., Cafasso, D., Cozzolino, S., Cortinovis, G., Sparvoli, F., Bellucci, E., Di Vittori, V., Nanni, L., et al. (2023). The evolutionary history of the common bean (*Phaseolus vulgaris*) revealed by chloroplast and nuclear genomes. Preprint at *bioRxiv*, DOI: https://doi.org/10.1101/2023.06.09.544374

12. Cortinovis, G.; Frascarelli, G.; Di Vittori, V.; and Papa, R. (2020) Current state and perspectives in population genomics of the common bean. Plants, 9, 330. DOI: https://doi.org/10.3390/plants9030330

13. Bellucci, E., Benazzo, A., Xu, C., Bitocchi, E., Rodriguez, M., Alseekh, S., Di Vittori, V., Gioia, T., Neumann, K., Cortinovis, G., et al. (2023) Selection and adaptive introgression guided the complex evolutionary history of the European common bean. *Nature Communications*, 14, 1908. DOI: https://doi.org/10.1038/s41467-023-37332-z

14. Golicz, A.A., Batley, J., and Edwards, D. (2016) Towards plant pangenomics. *Plant Biotechnology Journal*, 14:1099-1105. DOI: https://doi.org/10.1111/pbi.12499

15. Tranchant-Dubreuil, C., Rouard, M., and Sabot, F. (2019) Plant pangenome: impacts on phenotypes and evolution. *Annual Plant Reviews*, DOI: 10.1002/9781119312994.apr0664

16. Furaste Danilevicz, M., Tay Fernandez, C.G., Marsh, J.I., Bayer, P.E., and Edwards, D. (2020) Plant pangenomics: approaches, applications and advancements. *Current Opinion in Plant Biology*, 54:18-25. DOI: https://doi.org/10.1016/j.pbi.2019.12.005

17. Khan, A.W., Garg, V., Roorkiwal, M., Golicz, A.A., Edwards, D., and Varshney, R.K. (2019) Super-Pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends in Plant Science,* 25(2):148-158. DOI: https://doi.org/10.1016/j.tplants.2019.10.01p

18. Bellucci, E., Bitocchi, E., Ferrarini, A., Benazzo, A., Biagetti, E., Klie, S., Minio, A., Rau, D., Rodriguez, M., Panziera, A., et al. (2014) Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. *The Plant Cell*, 26(5):1901-1912. DOI: https://doi.org/10.1105/tpc.114.124040

19. Luo, M., Tai, R., Yu, C.W., Yang, S., Chen, C.Y., Lin, W.D., Schmidt, W., and Wu K. (2015) Regulation of flowering time by the histone deacetylase HDA5 in *Arabidopsis. The Plant Journal,* 82,925-936. DOI: https://doi.org/10.1111/tpj.12868

20. Blanco, F.A., Peltzer Meschini, E., Zanetti, M.E., and Aguilar, O.M. (2009) A small GTPase of the Rab family is required for root hair formation and preinfection stages of the Common bean–rhizobium symbiotic association. *The Plant Cell,* 21(9):2797-2810. DOI: https://doi.org/10.1105/tpc.108.063420

21. Dalla Via, V., Traubenik, S., Rivero, C., et al. (2017) The monomeric GTPase RabA2 is required for progression and maintenance of membrane integrity of infection threads during root nodule symbiosis. *Plant Molecular Biology*, 93:549-562. DOI: https://doi.org/10.1007/s11103-016-0581-5

22. Oladzad, A., González, A., Macchiavelli, R., de Jensen, C.E., Beaver, J., Porch, T., and McClean, P. (2020) Genetic factors associated with nodulation and nitrogen derived from atmosphere in a middle American common bean panel. *Frontiers in Plant Science*, 11:576078. DOI: https://doi.org/10.3389/fpls.2020.576078

23. Ke, D., Fang, Q., Chen, C., Zhu, H., Chen, T., Chang, X., Yuan, S., Kang, H., Ma, L., Hong, Z., et al. (2012) The small GTPase ROP6 interacts with NFR5 and is involved in nodule formation in Lotus japonicus. *Plant Physiology*, 159(1):131-143. DOI: https://doi.org/10.1104/pp.112.197269

24. Corpas, F.J., Gupta, D.K., and Palma, J.M. (2021). Tryptophan: a precursor of signaling molecules in higher plants. In: *Gupta, D.K., Corpas, F.J.* (eds) Hormones and Plant Response. Plant in Challenging Environments, vol 2. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-77477-6_11

25. Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio*, 3(2):e00036-12. DOI: https://doi.org/10.1128/mbio.00036-12

26. Wolf, Y.I. and Koonin, E.V. (2013) Genome reduction as the dominant mode of evolution. *Bioessays*, 35:829-837. DOI: https://doi.org/10.1002/bies.201300037

27. Suda, J., Meyerson, L.A., Leitch, I.J., and Pyšek, P. (2014). The hidden side of plant invasions: the role of genome size. *New Phytologist,* 205:994-1007. DOI: https://doi.org/10.1111/nph.13107

28. Lavergne, S., Muenke, N.J., and Molofsky, J. (2010) Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Annals of Botany,* 105(1):109-16. DOI: https://doi.org/10.1093/aob/mcp271

29. Díez, C.M., Gaut, B.S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L.E., and Tenaillon, M.I. (2013) Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytologist,* 199:264-276. DOI: https://doi.org/10.1111/nph.12247

30. Hottes, A.K., Freddolino, P.R., Khare, A., Donnell, Z.N., Liu, J.C., and Tavazoie, S. (2013) Bacterial adaptation through loss of function. *PLoS Genetics*, 9(7):e1003617. DOI: https://doi.org/10.1371/journal.pgen.1003617

31. Albalat, R. and Cañestro, C. (2016) Evolution by gene loss. *Nature Reviews Genetics*, 17:379-391. DOI: https://doi.org/10.1038/nrg.2016.39

32. Murray, A.W. (2020) Can gene-inactivating mutations lead to evolutionary novelty? *Current Biology*, 30(10):R465-R471. DOI: https://doi.org/10.1016/j.cub.2020.03.072

33. Monroe, J.G., McKay, J.K., Weigel, D., and Flood, P.J. (2021) The population genomics of adaptive loss of function. *Heredity*, 126:383-395, DOI: https://doi.org/10.1038/s41437-021-00403-2

34. Shimizu, K.K., Shimizu,-Inatsugi, R., Tsuchimatsu, T. and Purugganan, M.D. (2008), Independent origins of self-compatibility in *Arabidopsis thaliana*. *Molecular Ecology*, 17:704-714. DOI: https://doi.org/10.1111/j.1365-294X.2007.03605.x

35. Olson, M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics,* 64:18-23. DOI: https://doi.org/10.1086/302219

36. Lutz, K.A., Wang, W., Zdepski, A, and Todd, P.M. (2011) Isolation and analysis of high-quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnology*, 11,54. DOI: https://doi.org/10.1186/1472-6750-11-54

37. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, DOI: http://www.genome.org/cgi/doi/10.1101/gr.215087.116

38. Ruan, J. and Li, H. (2019) Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17:155-158. DOI: https://doi.org/10.1038/s41592-019-0669-3

39. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737-746. DOI: 10.1101/gr.214270.116

40. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, 9(11):e112963. DOI: https://doi.org/10.1371/journal.pone.0112963

41. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31:3210-3212. DOI: https://doi.org/10.1093/bioinformatics/btv351

42. Cheng, H., Concepcion, G.T., Feng, X. Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods,* 18,170–175. DOI: https://doi.org/10.1038/s41592-020-01056-5

43. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094-3100. DOI: https://doi.org/10.1093/bioinformatics/bty191

44. Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T.M., Kamal, N., Lang, D., Himmelbach, A., et al. (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 588:284-289. DOI: https://doi.org/10.1038/s41586-020-2947-8

45. Nattestad, M. and Schatz, M.C. (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics,* 32:3021-3023. DOI: https://doi.org/10.1093/bioinformatics/btw369

46. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin R; and 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMTools. *Bioinformatics*, 25(16):2078-2079. DOI: https://doi.org/10.1093/bioinformatics/btp352

47. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22:1658-1659. DOI: https://doi.org/10.1093/bioinformatics/btl158

48. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17): i884–i890. DOI: https://doi.org/10.1093/bioinformatics/bty560

49. Langmead, B. and Salzberg, S. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357-359. DOI: https://doi.org/10.1038/nmeth.1923

50. Zimin, A.V., Marçais, G., Puiu,D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013) The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669-2677. DOI: https://doi.org/10.1093/bioinformatics/btt476

51. Flynn, J., Hubley, R., Goubert, C., Rosen, J., Clark, A., Feschotte, C., and Smit, A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences,* 117(17):9451-7. DOI: https://doi.org/10.1073/pnas.1921046117

52. Tarailo-Graovac, M., and Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, 4:4.10.1-4.10.14. DOI: https://doi.org/10.1002/0471250953.bi0410s25

53. Hoff, K.J. and Stanke, M. (2019) Predicting Genes in Single Genomes with AUGUSTUS. *Current Protocols in Bioinformatics*, ;65(1):e57. DOI: https://doi.org/10.1002/cpbi.57

54. Kim, D., Paggi, J., Park, C., Bennett, C., and Salzberg, S. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907-15. DOI: https://doi.org/10.1038/s41587-019-0201-4

55. Gremme, G., Brendel, V.P., Sparks, M.E., and Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, 47(15):965-978. DOI: https://doi.org/10.1016/j.infsof.2005.09.005

56. Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics,* 30(9):1236-40. DOI: https://doi.org/10.1093/bioinformatics/btu031

57. Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* 215:403-10. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2

58. Emms, D.M., and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238. DOI: https://doi.org/10.1186/s13059-019-1832-y

59. Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987-2993. DOI: https://doi.org/10.1093/bioinformatics/btr509

60. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics & Bioinformatics*, 8:77-80. DOI: https://doi.org/10.1016/S1672-0229(10)60008-3

61. Yu, G., Wang, L., Han, Y., and He, Q. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters *OMICS: A Journal of Integrative Biology*, 16(5):284-287. DOI: https://doi.org/10.1089/omi.2011.0118

62. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141. DOI: https://doi.org/10.1016/j.xinn.2021.100141

63. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanasichuk, O., Benner, C., Chanda, S.K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communication*, 10,1523. DOI: https://doi.org/10.1038/s41467-019-09234-6

64. Landgraf, A.J. and Lee, Y. (2020) Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis,* 180:104668. DOI: https://doi.org/10.1016/j.jmva.2020.104668

65. Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics*, 15:323-354. DOI: https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

66. Tamura, K., Stecher, G., and Kumar, S. (2021) MEGA11: Molecular evolutionary genetics analysis. *Molecular Biology and Evolution*, 38(7):3022-3027. DOI: https://doi.org/10.1093/molbev/msab120

67. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics,* 38(2):203-8. DOI: https://doi.org/10.1038/ng1702

68. Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genetics,* 12(2):e1005767. DOI: https://doi.org/10.1371/journal.pgen.1005767

69. Wang, J. and Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics Proteomics Bioinformatics*, 19(4):629–40. DOI: https://doi.org/10.1016/j.gpb.2021.08.005

70. Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics,* 23(19):2633-5. DOI: https://doi.org/10.1093/bioinformatics/btm308

## Figures

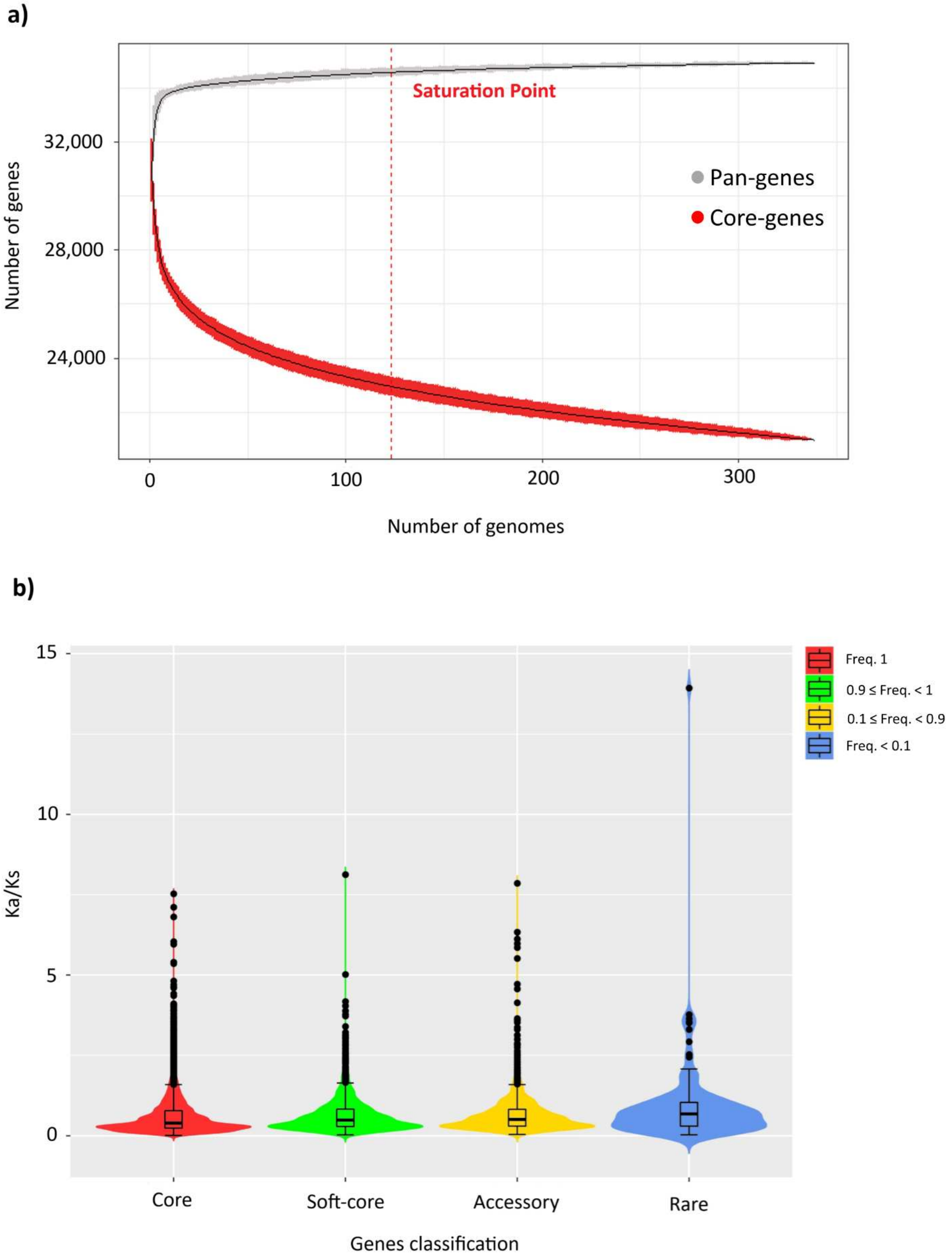
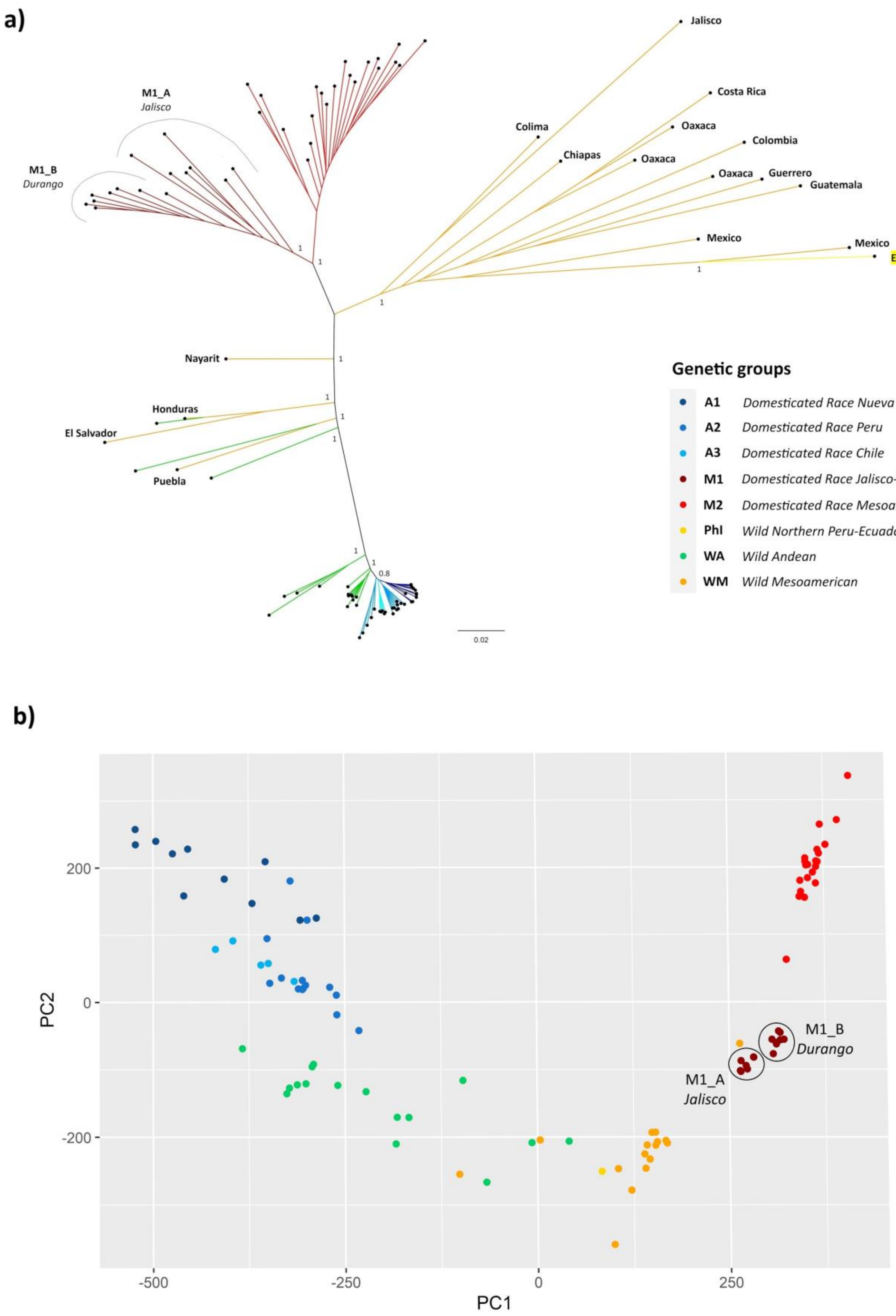


**Fig. 1: Characterization of the common bean pan-genome. a,** Pan-gene and core gene size calculation. The growth curve of pan-genes (grey) reached saturation point (99%, 34,579 genes) when 120 individuals were

included, as indicated by the dashed red line. In contrast, the growth curve of core genes (red) diminished with the addition of each genotype. **b,** Violin plots showing analysis of variance (ANOVA) related to the ratio of non-synonymous to synonymous mutations in the core genes and PAs. The PAVs are split into three subcategories based on their frequency: soft core, accessory, and rare. **Supplementary Table 1g** contains detailed statistics.


a)
M1_A
Jalisco
M1_B
Durango
Jalisco
Costa Rica
Colima
Oaxaca
Colombia
Chiapas
Oaxaca
Oaxaca
Guerrero
Guatemala
Mexico
Mexico
Ecuador
Nayarit
Honduras
El Salvador
Puebla
1
0.8
0.02
Genetic groups
● A1 Domesticated Race Nueva Granada
● A2 Domesticated Race Peru
● A3 Domesticated Race Chile
● M1 Domesticated Race Jalisco-Durango
● M2 Domesticated Race Mesoamerica
● PhI Wild Northern Peru-Ecuador
● WA Wild Andean
● WM Wild Mesoamerican
b)
PC2
200
0
-200
-500
-250
0
250
PC1
M1_B
Durango
M1_A
Jalisco

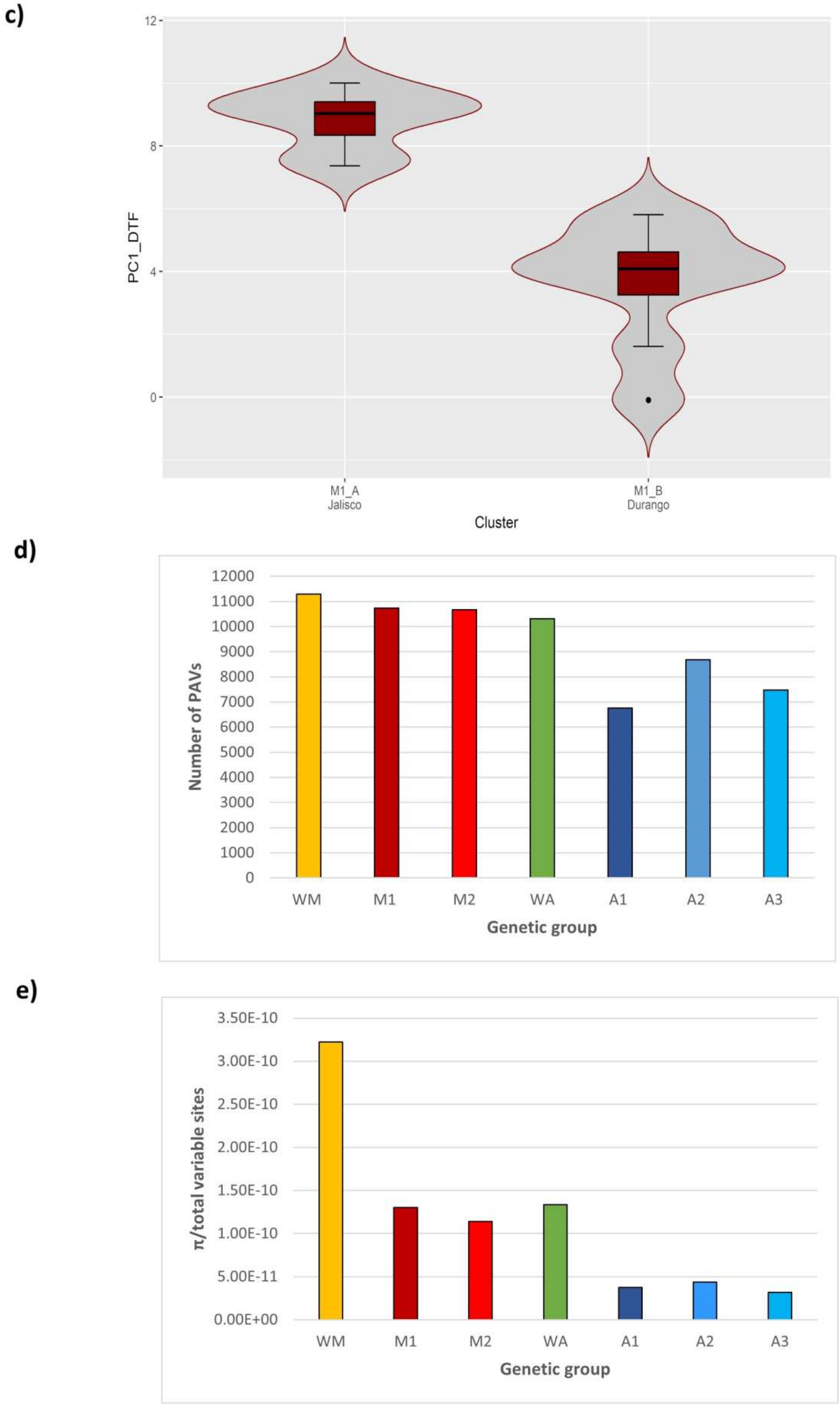

c)
PC1_DTF
12
8
4
0
M1_A
Jalisco
M1_B
Durango
Cluster
d)
Number of PAVs
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
WM
M1
M2
WA
A1
A2
A3
Genetic group
e)
π/total variable sites
3.50E-10
3.00E-10
2.50E-10
2.00E-10
1.50E-10
1.00E-10
5.00E-11
0.00E+00
WM
M1
M2
WA
A1
A2
A3
Genetic group

**Fig. 2: Population structure of *P. vulgaris*. a,** Neighbour-joining (NJ) phylogenetic tree constructed using only SNPs located in core genes (bootstrap = 1000). **b,** PAV-based principal component analysis (PCA). **c,** Violin plots showing the analysis of variance (ANOVA) for PC1 representing days to flowering and photoperiod sensitivity in the M1/Jalisco-Durango races by splitting the accessions into two clusters based on PCA and the NJ tree. **d,** Bar chart showing the number of PAVs per genetic group. **e,** Bar chart showing nucleotide diversity calculated by estimating $\pi$ in 250-kb windows. All procedures were applied to a representative subset of 99 genetically and phenotypically well-characterized *P. vulgaris* accessions.

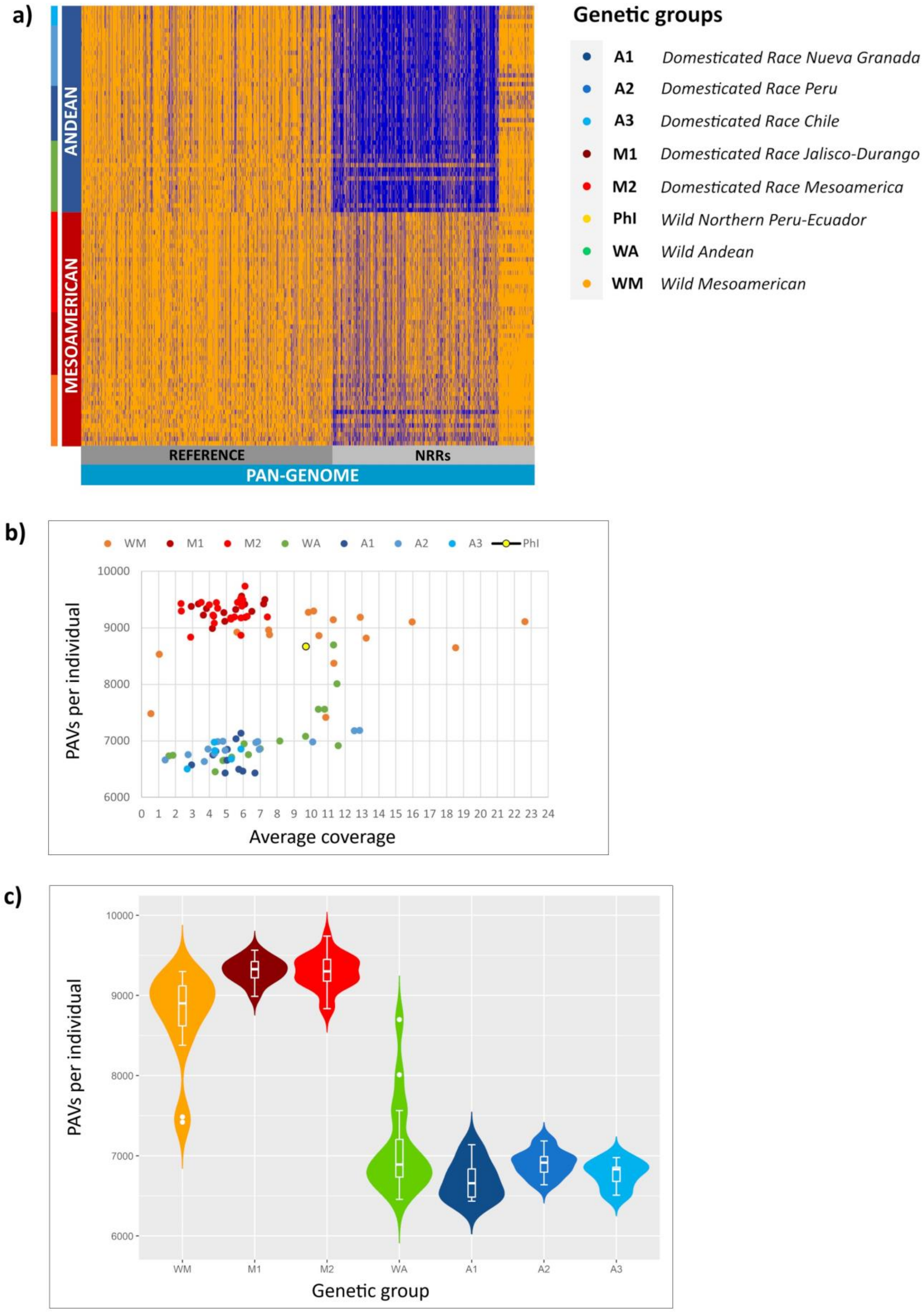

a)
ANDEAN
MESOAMERICAN
REFERENCE
NRRs
PAN-GENOME
Genetic groups
A1 Domesticated Race Nueva Granada
A2 Domesticated Race Peru
A3 Domesticated Race Chile
M1 Domesticated Race Jalisco-Durango
M2 Domesticated Race Mesoamerica
Phl Wild Northern Peru-Ecuador
WA Wild Andean
WM Wild Mesoamerican
b)
WM M1 M2 WA A1 A2 A3 Phl
PAVs per individual
10000
9000
8000
7000
6000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Average coverage
c)
PAVs per individual
10000
9000
8000
7000
6000
WM M1 M2 WA A1 A2 A3
Genetic group

**Fig. 3: Evolution of the common bean pan-genome. a,** Heat map illustrating the number of PAVs per individual in the final pan-genome. Orange indicates presence while blue indicates absence. **b,** Scatterplot showing the number of PAVs per individual (y-axis) in relation to the coverage (x-axis) of each genotype. **c,** Violin plots representing the analysis of variance (ANOVA) for the number of PAVs per individual by genetic group. All procedures were applied to a representative subset of 99 genetically and phenotypically well-characterized *P. vulgaris* accessions. **Supplementary Table 3a** contains detailed statistics.

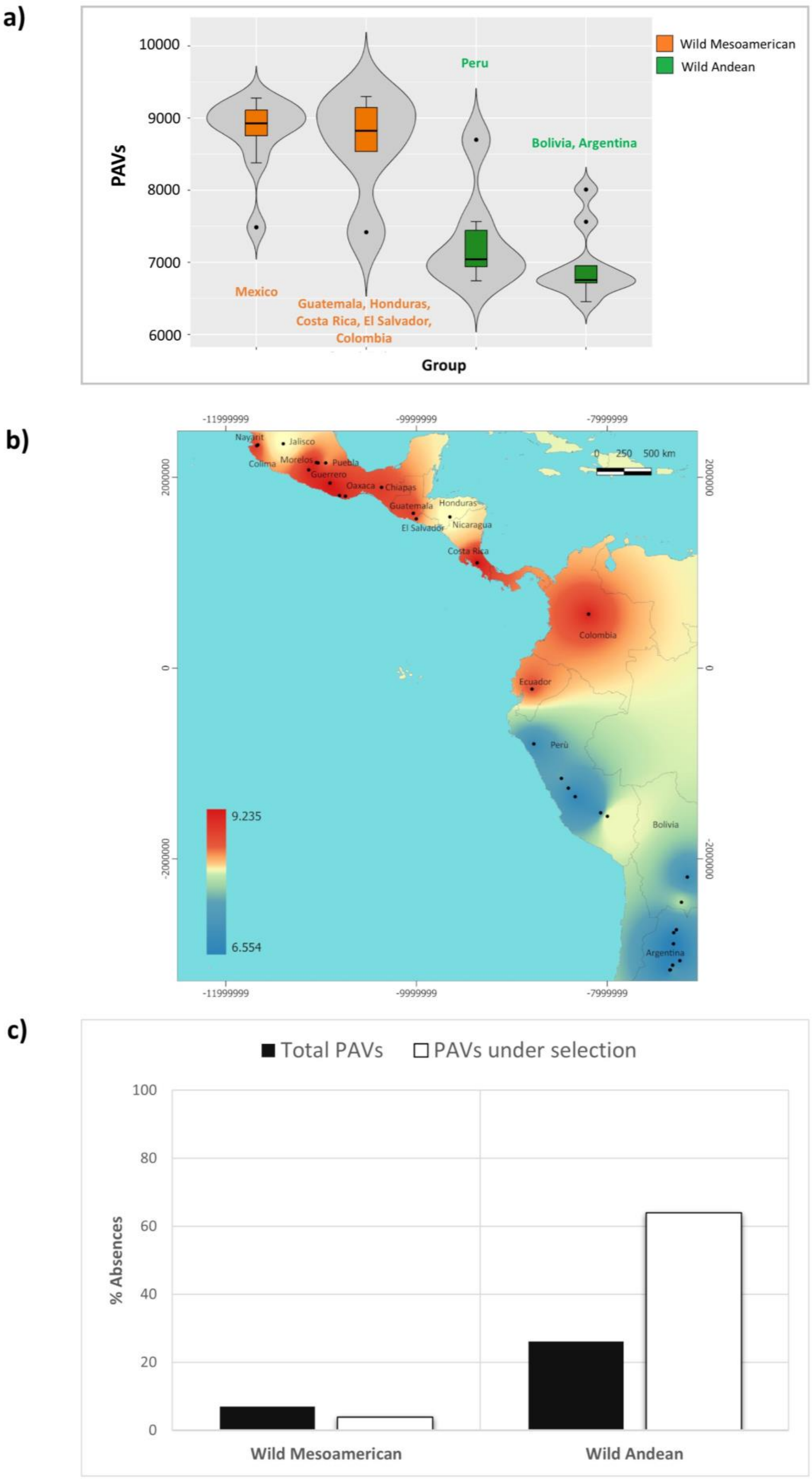

a)
10000
9000
8000
7000
6000
PAVs
Group
🟧 Wild Mesoamerican
🟩 Wild Andean
Mexico
Guatemala, Honduras, Costa Rica, El Salvador, Colombia
Peru
Bolivia, Argentina
b)
-11999999
-9999999
-7999999
2000000
0
-2000000
0 250 500 km
Nayarit
Jalisco
Colima
Morelos
Puebla
Guerrero
Oaxaca
Chiapas
Guatemala
Honduras
El Salvador
Nicaragua
Costa Rica
Colombia
Ecuador
Perù
Bolivia
Argentina
9.235
6.554
c)
■ Total PAVs
□ PAVs under selection
100
80
60
40
20
0
% Absences
Wild Mesoamerican
Wild Andean

**Fig. 4: Selection for adaptive gene loss during the expansion of wild common bean. a,** Violin plots showing the analysis of variance (ANOVA) for the number of PAVs per individual based on grouping the wild Mesoamerican and Andean accessions according to latitude coordinates. **Supplementary Table 3b** contains detailed statistics. **b,** Interpolation of the geographic distributions of the wild accessions based on the number of PAVs per individual. Dark red regions indicate a higher number of PAVs and blue regions a lower number of PAVs. **c,** Bar charts showing the proportions of absences found for the subset of PAVs putatively under selection during the wild expansion (white) and for the entire variable genome (black).

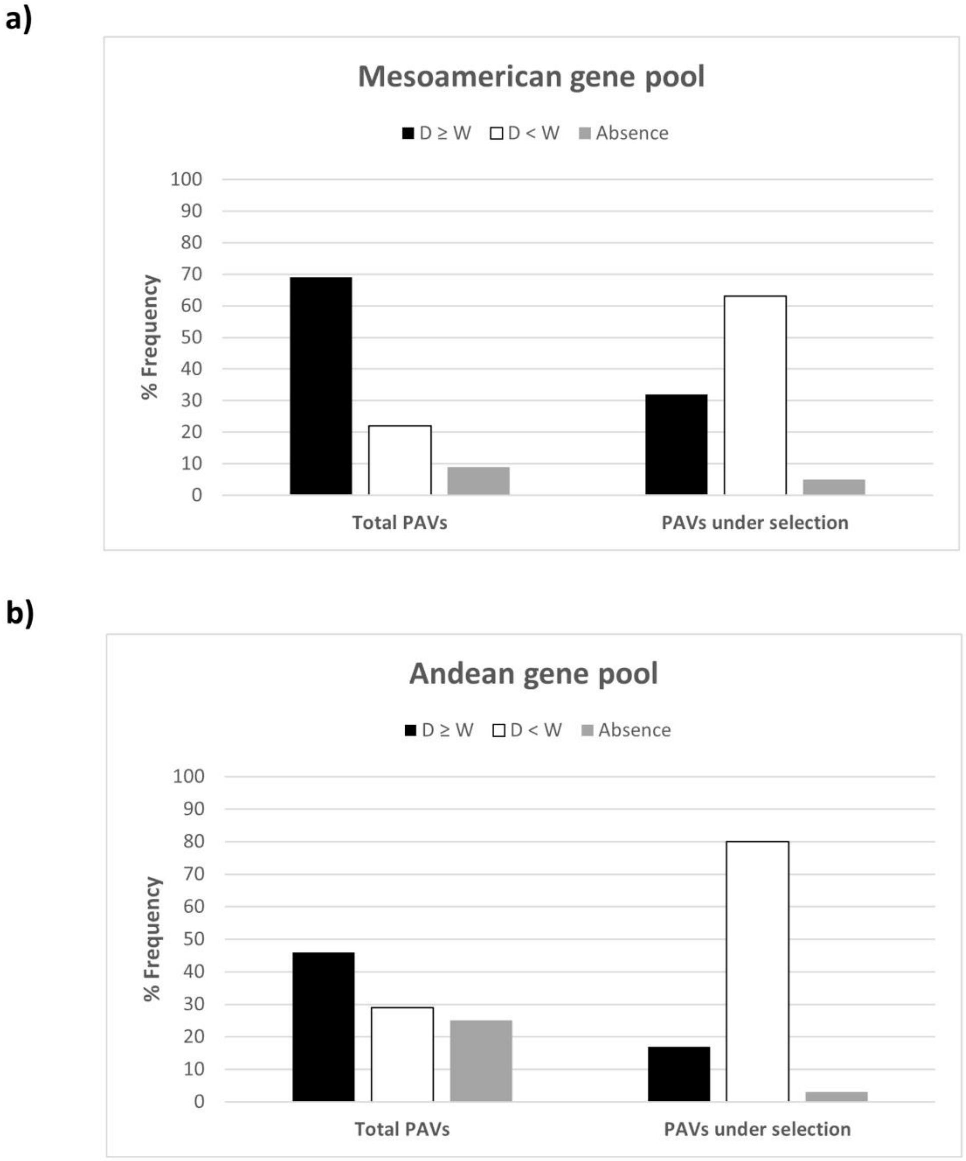


**Fig. 5: Localized adaptive reduction effects during the domestication of the common bean. a,** Bar chart showing the proportions of presence/absence in the Mesoamerican gene pool for the entire variable

genome (left) and for the subset of PAVs putatively under selection between wild and domesticated populations (right). **b,** Bar chart showing the proportions of presence/absence in the Andean gene pool for the entire variable genome (left) and for the subset of PAVs putatively under selection between wild and domesticated populations (right). In both charts, the presence values are divided based on frequency (≥/<) in the comparison between wild and domesticated forms.