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

[°] Co-first authors

[*] These authors contributed equally to this work

[#] Corresponding author

Abstract

The common bean (*Phaseolus vulgaris* L.) is a crucial grain legume crop [1,2] whose life history offers an ideal evolutionary model to identify and study adaptive variants in wild and domestication populations [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 strong 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 signatures during wild spreading and domestication were 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 novel result has major implications for the understanding of the process of plant adaptation and claims for a paradigm shift in evolutionary genetics. Moreover, the common bean pan-genome is a valuable resource for food legume research and breeding towards 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 crucial societal importance [1,2]. It is also an ideal 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 more 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,

104 A1, A2 and A3) corresponding to the Jalisco-Durango, Mesoamerica, Nueva Granada, Peru, and Chile races
105 [13]. We constructed a neighbour-joining (NJ) phylogenetic tree (**Fig. 2a**) and conducted PAV-based
106 principal component analysis (PCA) (**Fig. 2b**), both of which confirmed this well-defined population
107 structure. Both analyses further divided the M1/Jalisco-Durango races into two clusters that we named
108 cluster A and cluster B, respectively. The analysis of variance conducted on M1/Jalisco-Durango accessions,
109 considering the first component for the days to flowering, revealed that cluster A is significantly later-
110 flowering than cluster B (**Fig. 2c; Supplementary Table 2a**). The Jalisco (cluster A) and Durango (cluster B)
111 races are therefore genetically distinguishable based on photoperiod sensitivity. This outcome also
112 confirmed that our pan-genome enhances the characterization of genetic diversity and improves its
113 analysis, exploitation and management. Cumulatively, the first and the second principal components of the
114 PAV-based PCA explained 47.2% of the total variance, where PC1 mainly defined the differences between
115 Mesoamerican and Andean gene pools while PC2 split the groups and subgroups within each gene pool
116 (**Fig. 2b**). The NJ tree further underscored the suitability of core genes for phylogenetic reconstruction
117 because they mitigate biases arising from the absence of genetic material among the compared accessions.
118 In contrast to the tree based on single-nucleotide polymorphisms (SNPs) located on PAVs (**Supplementary**
119 **Fig. 2a**), the NJ tree based solely on core SNPs properly grouped the wild PhI accession close to the wild
120 Mesoamerican genotypes originating from Guatemala and Costa Rica (**Fig. 2a**), which are most closely
121 related to the PhI gene pool [11].

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

132 To study inter-gene-pool hybridization, the PAV matrix for American domesticated accessions was
133 analysed by using Fisher's exact test to compare the Mesoamerican and Andean populations. We found
134 5,556 PAVs (65% of the total) with a statistically significant difference in frequency ($p < 0.05$) between the
135 two gene pools. These included 778 diagnostic PAVs, 91% (707) of which were fixed in the Mesoamerican
136 gene pool and 9% (71) in the Andean gene pool (**Supplementary Table 2d**). GO enrichment analysis applied
137 to the 778 diagnostic genes revealed enrichment in processes related to detoxification (GO:0098754),
138 metabolism (GO:0008152), and responses to stimuli (GO:0050896) (**Supplementary Fig. 2b**). Interestingly,

139 none of these PAVs were found to be diagnostic between gene pools in Europe (**Supplementary Table 2d**),
140 and when Fisher's exact test was applied to the subset of 114 European accessions, we did not detect any
141 diagnostic genes between the Mesoamerican and Andean gene pools (**Supplementary Table 2e**). These
142 outcomes clearly reflect the extensive inter-gene-pool hybridization in European germplasm and confirm its
143 key role in the adaptation of common bean to new agricultural environments [13].
144 To investigate the influence of PAVs on important trait (i.e., flowering time) variations and identify
145 candidate genes associated with them, we conducted a PAV-based genome-wide association study (GWAS)
146 involving 218 American and European domesticated genotypes. We identified 39 significative association
147 events correlated with day-to-flowering and photoperiod sensitivity, previously detailed in [13]. These
148 associations were linked to 35 potential candidate PAVs, highlighting their likely involvement in regulating
149 floral transition (**Supplementary Table 2f**). An interesting example is the GWAS peak associated with
150 flowering time and photoperiod sensitivity located on Phvul.003G185200 (Chr03:40,838,810-40,850,729).
151 This PAV demonstrates orthology to the *HDA5* gene in *Arabidopsis thaliana*, which displays deacetylase
152 activity. Notably, *A. thaliana* mutants with impaired HDA5 expression patterns manifest late-flowering
153 phenotypes due to the up-regulation of two floral repressor genes, namely FLOWERING LOCUS C (FLC) and
154 MADS AFFECTING FLOWERING 1 (MAF1) [19]. It is noteworthy that common bean genotypes lacking the
155 PAV Phvul.003G185200 exhibit early flowering phenotypes compared to those accessions carrying the gene
156 (**Supplementary Fig. 2c**), implying an adaptive role correlated to the gene loss. Furthermore, we found that
157 9 out of the 35 candidate PAVs for GWA analysis show signature of selection in various comparisons,
158 specifically, two PAVs between wild and domesticated Mesoamerican populations and seven PAVs between
159 wild and domesticated Andean populations. Overall, although the majority (59%) of the candidate PAVs
160 were located on the reference genome, a significant 41% were situated on the NRRs (**Supplementary Table**
161 **2f**), reaffirming the efficacy of the pan-genome in identifying functional variants associated with
162 economically or evolutionarily important traits.
163
164 **Pan-genome shrinkage during wild expansion to South America**
165 One of the most striking outcomes we observed was the difference in pan-genome size between the
166 Mesoamerican and Andean gene pools (**Fig. 3a**). We calculated the total number of PAVs per individual and
167 found that accessions from the same gene pool clustered together in separate groups, with Mesoamerican
168 accessions featuring more PAVs per accession than Andean ones (**Fig. 3b, c; Supplementary Table 3a**). One
169 possible explanation is that this reduction in pan-genome size may simply reflect genetic drift and the two
170 sequential bottlenecks that occurred solely in the Andean population [12]. To better understand the roles
171 of different evolutionary forces in shaping the PAV content of the Mesoamerican and Andean gene pools,
172 and to distinguish between the effects of adaptation, population demography and history, we first focused
173 on the analysis of wild accessions. Considering a panel of wild genotypes representing the entire

174 geographical distribution in Latin America, we applied bivariate fit analysis and found a significant
175 correlation ($p < 0.0001$) between the number of PAVs per individual and the latitude. Analysis of variance in
176 which wild individuals were grouped by latitude followed by spatial interpolation revealed a significant and
177 progressive loss of genes ranging from the accessions of Northern Mexico to those of Northwestern
178 Argentina (**Fig. 4a, b; Supplementary Table 3b**). Furthermore, F_{ST} analysis on PAVs comparing
179 Mesoamerican and Andean wild populations may suggests the occurrence of selection for gene loss during
180 wild range expansion (**Fig. 4c**). We found that 64% of all candidate PAVs in the top 5% of the F_{ST} distribution
181 (Fst≥0.85) are absent in the wild Andean gene pool. This high rate of gene loss in the Andean population
182 significantly exceeds the gene loss rate observed in the entire variable genome (26%), demonstrating a
183 more than twofold increase. This difference in gene loss rates was statistically validated using bootstrap
184 resampling, strongly suggesting that gene loss during the process of wild differentiation was not a random
185 occurrence but the evident outcome of selective forces (**Supplementary Fig. 3a, b**). Moreover, functional
186 annotation of the candidate PAVs revealed the enrichment of genes involved in pollen germination, innate
187 immunity, abiotic stress tolerance, and root hair growth, indicating a potential adaptive role during wild
188 range expansion (**Supplementary Table 3c**). Overall, our findings suggest that selective pressure favouring
189 the loss of genes involved in adaptive mechanisms, coupled with the influence of genetic drift resulting
190 from the founder effect, may have contributed to the shrinking of the Andean pan-genome during wild
191 differentiation.
192
193 **Footprints of selection for gene loss during domestication**
194 The PAVs putatively shaped by selection during domestication in Mesoamerica and the Andes revealed
195 further evidence that gene loss underpinned the successful adaptation of the American common bean. F_{ST}
196 analysis was applied to PAVs in wild and domesticated forms (separately for each gene pool) with only PAVs
197 in the top 5% of the F_{ST} distribution considered as candidates. We found 460 PAVs potentially under
198 selection in the Mesoamerican population ($F_{ST} \geq 0.31$) and 514 in the Andean population ($F_{ST} \geq 0.27$)
199 (**Supplementary Table 4a, b**). Functional annotation of the candidate PAVs revealed the enrichment of
200 genes associated with domestication syndrome and adaptive traits such as dormancy, floral transition, light
201 acclimation, defence, and symbiotic interactions (**Supplementary Table 4c, d**). Importantly, the candidate
202 Phvul.003G265200 (Chr03: 50,365,995-50,368,501) is orthologous to 11 members of the plant Rho GTPase
203 subfamily (ROP), including *ROP6* encoding a small Rho-like GTP binding protein. This GTPase subfamily is
204 required for symbiotic interactions [20-22], and in the plasma membrane of *Lotus japonicus* cells it interacts
205 directly with NOD FACTOR RECEPTOR 5, one of two nodulation factor receptors essential for nodule
206 formation during symbiosis [23]. From our analysis, Phvul.003G265200 is a putative selected PAV (F_{ST} =
207 0.50) for the Mesoamerican gene pool whose frequency declined by more than 60% during progression
208 from the wild (0.94) to the domesticated (0.25) population (**Supplementary Table 4a**). Overall, no

209 significant differences were observed in terms of absences between the wild and domesticated populations
210 of both gene pools. However, a significant proportion of PAVs putatively under selection, specifically 63%
211 (289 genes) in the Mesoamerican population (**Fig. 5a**) and 80% (411 genes) in the Andean one (**Fig. 5b**),
212 were less frequent in domesticated than wild populations. When considering all PAVs, the percentage of
213 PAVs with lower frequencies in domesticated populations fell significantly to 22% ($p < 2.2 \times 10^{-16}$) for the
214 Mesoamerican gene pool and 29% ($p < 2.2 \times 10^{-16}$) for the Andean one (**Fig. 5a, b**). These findings suggest
215 that selection during domestication resulted in gene loss, but unlike the range expansion of wild
216 populations, it did not completely abolish the selected genes. This may reflect the different evolutionary
217 timescales involved: wild differentiation occurred ~150,000 years ago whereas domestication was much
218 more recent at ~8,000 years ago. These findings are consistent with previous observations that selection
219 during the domestication of common bean in Mesoamerica has directly affected the transcriptome, leading
220 to a ~20% decrease in gene expression levels attributed to loss-of-function mutations [18]. We also
221 detected 29 PAVs with high F_{ST} values in common between the Mesoamerican and the Andean gene pools,
222 and these are mainly associated with the tryptophan metabolic pathway. Tryptophan holds significance as
223 a precursor in secondary metabolism, contributing to the synthesis of essential molecules like auxin,
224 serotonin, and melatonin. These compounds play diverse roles in plant physiology, influencing processes
225 such as seed germination, root development, senescence, and flowering. Additionally, they contribute to
226 the plant's response mechanisms against both biotic and abiotic stresses [24]. We analysed their
227 frequencies and found that ~86% in both gene pools decreased in frequency during the progression from
228 wild to domesticated accessions (**Supplementary Table 4e**). This may indicate a pattern of genomic
229 convergence for some key adaptive genes between the Mesoamerican and Andean populations during
230 their parallel domestication events.
231

232 **Conclusions**

233 The global economic and social importance of the common bean means that pan-genomic analysis could
234 boost the conservation and exploitation of its genetic resources to address key challenges in sustainable
235 agriculture and wider society [1-6]. The genotypes selected for this study encompass both wild and
236 domesticated forms, ensuring that the pan-genome comprehensively captures the extensive genetic
237 variation within this species. This publicly accessible tool serves as a valuable resource for studies in
238 population genetics, functional genomics, and plant breeding. PAV analysis provided insight into the
239 evolutionary dynamics of pan-genome adaptation, including signals of selection for complete gene loss
240 during wild differentiation between the Mesoamerican and Andean gene pools, contributing to the smaller
241 pan-genome of the Andean population. We also identified selection footprints for gene loss during both
242 domestication processes in Mesoamerica and the Andes, causing localized reductions in gene frequencies
243 in domesticated populations compared to their wild counterparts. Interestingly, 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.

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 [25-30]. Moreover, the variation in genome size between different populations of microorganisms and plants has been already described [31-33]. For instance, in contrast to their European native counterparts, invasive genotypes exhibited a reduced genome size resulting in phenotypic effects that could enhance the species' invasive potential [34]. Similarly, Díez et al. [35] 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. However, it is still an open question as to whether genome size variation is shaped by natural selection. Here, our results claim that under the influence of specific and diverse agro-ecological pressures, the relinquishment of particular genes can confer a selective advantage on a given population. 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 research presents pioneering evidence, establishing a new paradigm in which natural selection drives gene loss, favouring adaptation over mere stochastic responses. 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 could have profound implications for our understanding of crop adaptation under 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

278 MIDAS and G12873 single seed descent (SSD) genotypes were multiplied in the greenhouse. For both
279 samples, 2 g of young leaf material was collected after 48 h of dark treatment and high-molecular-weight
280 (HMW) DNA was extracted as previously described [36]. DNA quality was evaluated according to Oxford
281 Nanopore Technologies (ONT) requirements. Specifically, purity was assessed using a NanoDrop 1000
282 spectrophotometer (Thermo Fisher Scientific), the concentration was determined using a dsDNA Broad
283 Range Assay Kit with Qubit 4.0 (Thermo Fisher Scientific), and the fragment size (≤ 400 kb) was determined
284 using the CHEF Mapper electrophoresis system (Bio-Rad Laboratories). Fragments < 25 kb were removed
285 using the Short Reads Eliminator kit (Circulomics) leaving 75% of the DNA from the MIDAS samples and 95%
286 from the G12873 samples.
287 *P. vulgaris* genotypes of BAT93 and JaloEEP558 were sowed in soil and grown in a growth chamber at 23°C,
288 75% humidity, and 8 h dark and 16 h light photoperiods under fluorescent tubes (166lE). Young trifoliate
289 leaves of BAT93 and JaloEEP558 genotypes were collected and flash-frozen with liquid nitrogen. Three days
290 before sampling, plants were dark treated to optimize the high molecular weight DNA extraction. In
291 addition, 109 SSD accessions were multiplied in the greenhouse and young leaves were collected in silica
292 gel for drying and subsequent DNA extraction using the DNeasy 96 Plant kit (Qiagen) according to the
293 manufacturer's instructions. For each sample, 50–70 mg of dried leaf material was pulverized with a
294 Tissue-Lyser II (Qiagen) at 30 Hz for 6 min. The DNA quality and quantity were evaluated using a
295 NanoPhotometer NP80 (Implen), and the concentration was determined using a Qubit BR dsDNA assay kit
296 (Thermo Fisher Scientific).
297
298 **Sequencing low-coverage WGS accessions**
299 DNA libraries for all samples were prepared using a KAPA Hyper Prep kit and PCR-free protocol (Roche). For
300 each genotype, 200 ng of DNA was fragmented by sonication using a Covaris S220 device (Covaris) and
301 WGS DNA libraries were generated using a 0.7–0.8× ratio of AMPureXP beads for final size selection.
302 Libraries were quantified using the Qubit BR dsDNA assay kit and equimolar pools were quantified by real-
303 time PCR against a standard curve using the KAPA Library Quantification Kit (Kapa Biosystems). Libraries
304 were sequenced on the NovaSeq 6000 Illumina platform in 150PE mode, producing 15–30 million
305 fragments per sample.
306
307 **Sequencing and assembly of MIDAS and G12873 genomes**
308 Following quality control and priming according to ONT specifications, libraries were sequenced on a
309 MinION device with a SpotON flow cell (FLO-MIN106 R9.4.1-Rev D). Two libraries were prepared for each
310 genotype according to the SQK-LSK109 ligation sequencing protocol (ONT) with minor adjustments. Each
311 library was loaded twice, and the flow cell was washed using the Flow Cell Wash Kit (ONT). Illumina PCR-
312 free libraries were prepared starting with 1 ug of fragmented gDNA using the KAPA Hyper prep protocol.

313 This process involved extending the adapter ligation time up to 30 minutes and conducting post-clean-up
314 size selection using 0.7X AMPure XP beads. The library's concentration and size distribution were assessed
315 using a Bioanalyzer 2100 in combination with high-sensitivity DNA reagents and chips. Sequencing was
316 performed on a NovaSeq6000 instrument to generate 150-bp paired-end reads. MIDAS and G12873 whole-
317 genome assemblies were generated using the nanopore-based approach based on 26 Gb (50-fold coverage)
318 and 36 Gb (69-fold coverage), respectively. Raw nanopore reads were corrected using Canu v2.1 [37] and
319 the resulting corrected reads were assembled *de novo* using wtdbg2 v2.5 [38]. Draft assemblies were
320 refined by iterative polishing using long reads (Racon v1.4.3 and Medaka v1.0.3) [39] and short reads (three
321 rounds of Pilon v1.23) [40]. Completeness was assessed using BUSCO v4.1.2 [41] and the Fabales_odb10
322 dataset (**Supplementary Table 5c**).
323
324 **Sequencing and assembly of BAT93 and JaloEEP558 genomes**
325 High molecular weight DNA of BAT93 and JaloEEP558 genotypes was sequenced with a PacBio Sequel II
326 system by GENTYANE platform (INRAE Clermont-Ferrand, France). A total of 21.09 and 29.35 Gb of PacBio
327 HiFi reads were generated from BAT93 and JaloEEPP558, respectively. PacBio HiFi reads were *de novo*
328 assembled into contigs using HiFiasm (v 0.9.0) with default parameters [42].
329
330 **Orthologous/paralogous analysis and clustering threshold settings**
331 To incorporate two distinct populations, namely the Andean and the Mesoamerican gene pools, into the
332 pangenome, precise differentiation between orthologous and paralogous genes is imperative.
333 Consequently, a meticulous strategy was devised to ensure the preservation of solitary orthologous gene
334 copies along with all paralogous counterparts. The relationship between orthologous genes was calculated
335 with minimap2 v2.17 [43] to align the MIDAS and G12873 genome assemblies using the open reading
336 frames (ORFs) of 2,330 complete single-copy BUSCO (Benchmarking Universal Single-Copy Orthologs) genes
337 selected from the *P. vulgaris* reference genome PV442 (**Supplementary Table 5d**). The percentage identity
338 was calculated for each ORF based on the number of matches in the alignments as a proportion of ORF
339 length. The relationship between paralogous genes was calculated using the three most abundant gene
340 families (OG0000273, OG0000328, and OG0000085) in the *P. vulgaris* PV442 reference genome, composed
341 of 26, 37, and 42 genes, respectively. An all-versus-all comparison between the members of the same
342 family was implemented using minimap2. The percentage of identity was calculated for each gene family
343 by dividing the number of matches in the alignments by the reference gene ORF length and then averaging
344 the identity percentages for each family. Finally, the results of both tests were used to establish a clustering
345 threshold of 90% to retain only one orthologous and all paralogous genes in the pan-genome
346 (**Supplementary Table 5e**).
347

348 **Pan-genome construction**

349 We used a paired genome alignment strategy for pan-genome construction [44]. The PV442 reference
350 genome was independently mapped onto MIDAS, G12873, BAT93 and JaloEPP558 with minimap2 v2.17
351 using the alignment pre-set *-x asm5*, which considers regions with an average divergence < 5%. The bam
352 files of the four alignments were converted to delta files and structural variants were called using
353 Assemblytics v 1.2.1 [45]. Only deletions were selected as NRRs [44]. Uncovered contigs private to the four
354 analysed genomes were identified by applying samtools depth v1.1 [46] to the bam files and were
355 extracted as NRRs. Deletions and uncovered contigs were independently filtered for a minimum length of 1
356 kb and clustered using CD-HIT-EST v4.8.1 [47] with a sequence identity of 90%, as described above for the
357 orthologous and paralogous genes. To ensure that the NRRs identified through this method didn't
358 encompass orthologous genes already existing in the PV442 reference genome, we specifically employed
359 highly conserved BUSCO genes. We conducted a comparative analysis between the full complement of
360 4,947 MIDAS and 4,812 G12873 BUSCO genes present in PV442 and the NRRs derived from MIDAS and
361 G12873 using BLASTp. Illumina data representing the 339 low-coverage WGS common bean accessions
362 were trimmed with fastp v0.21.0 [48] and aligned to the preliminary pan-genome using bowtie2 v2.3.5.1
363 [49] with default parameters. The unmapped reads from these alignments were extracted using samtools
364 v1.11, assembled using MaSuRCA v3.4.2 [50] with default parameters, and clustered using CD-HIT-EST
365 v4.8.1 with a sequence identity of 90%. Finally, the NRRs derived from the panel of 339 common bean
366 accessions were added to the preliminary pan-genome to generate the final pan-genome. To exclude
367 putative contaminants and/or organelle sequences, NRRs were compared to the NCBI non-redundant
368 nucleotide database using BLASTn, considering a minimum 80% identity and 25% coverage, leading to the
369 removal of 1,194 sequences. Overall, we identified 64,174 added sequences, 86% of which reflected the
370 mapping of the 339 low-coverage WGS accessions. The remaining 14% was identified by comparing the
371 reference genome independently with the other four high-quality genomes (**Supplementary Table 1a**).

372

373 **Pan-genome annotation**

374 Repetitive sequences were identified and soft-masked using RepeatModeler v2.0.2 [51] and RepeatMasker
375 v4.1.2-p1 [52]. Protein-coding genes were identified using a hybrid-approach prediction with Augustus
376 v3.3.3 [53]. Proteins from *P. vulgaris* and correlated species (*Medicago truncatula* and *Glycine soja*) plus
377 RNA-Seq data (unpublished data from [18]) were aligned to the genome and used as extrinsic evidence.
378 Protein sequences were aligned with Hisat2 v2.2.1 [54] and RNA-Seq data were aligned using Genome
379 Threader v1.7.1 [55]. BUSCO genes in the Fabales_odb10 database were used to train the model for the
380 Augustus predictor. Predicted genes were scanned with InterProScan v5.46-81.0 [56] for the presence of
381 protein domains. Using a custom script, genes with transposon-related domains were filtered out and
382 retained in the annotation if they contained at least one known protein domain. The filtered proteins were

383 compared to the pan-genome with BLASTp v2.12.0 [57] and filtered by the best hits. The predicted genes
384 were clustered with the proteins of all species considered in the annotation using OrthoFinder v2.5.4 [58].
385 Finally, functional annotation was achieved by integrating information about orthologous genes and by
386 identifying functional domains using a custom script.
387
388 **PAV calling**
389 Illumina data representing the 339 low-coverage WGS accessions were aligned to the pan-genome using
390 bowtie2 v2.3.5.1 and the coverage of each predicted gene was calculated for each accession using samtools
391 v1.11 (**Supplementary Table 5f**). PAV calling thresholds were defined for each accession according to the
392 minimum coverage value of 1000 randomly selected BUSCO genes (ORFs). The BUSCO genes are
393 orthologous genes that should be present in all considered accessions, but a few underrepresented genes
394 in a given accession could constitute a bias. To avoid this, values below 1% (the 10 least covered genes)
395 were discarded. The identified genes were classified based on their frequency as core genes if present in all
396 the accessions or PAVs if partially shared or private to a single genotype (**Supplementary Table 1b, c**).
397
398 **Pan-genes and core genes size calculation**
399 The curves describing the pan-genome and core genome sizes were evaluated by considering 1,000 random
400 orders of the 339 genotypes. The orders were chosen randomly among all possible permutations (n! where
401 n=[1,339]). For each ordering, the gene sets of the accessions were progressively added to the total
402 genome size without considering the genes already present in the total set. The same procedure was
403 applied for the core genome size, but the gene sets were intersected when each genome was added, thus
404 keeping only the genes in common for each iteration (**Supplementary Table 5g, h**).
405
406 **Variant calling**
407 SNVs and InDels were called with bcftools v1.10.2 [59] based on the alignment of 339 accessions with the
408 pan-genome using bowtie2 v2.3.5.1. We used bcftools mpileup v1.10.2 to generate a genotype likelihood
409 table, and variants were identified using bcftools call v1.10.2 and the pileup table, producing the final vcf
410 file.
411
412 **Non-synonymous and synonymous mutations**
413 The Ka/Ks ratio was computed for each gene in each accession using KaKs calculator v2.0 [60]. For each
414 gene, the consensus sequence of each accession was extracted using bcftools consensus v1.10.2. The
415 calculator compares the pan-genome gene sequence with the gene sequence of each accession to identify
416 non-synonymous and synonymous variants and then computes the ratio. The calculator reported *NA* when
417 there were no variants in a specific accession or when the denominator of the Ka/Ks ratio was zero. It was

418 possible to compute the analysis for 30,850 of 34,928 genes. Sometimes the length of one of the two
419 compared sequences was not divisible by three so the sequence could not be read in triplets
420 (**Supplementary Table 1f**). The average Ka/Ks value per gene was used to assess the significance of the
421 sample median (**Supplementary Table 1g**).
422
423 **Data analysis**
424 Pan-genome analysis focused on a representative subset of 99 well-characterized accessions among the
425 original 339, including wild and American domesticated forms. In some cases, we also analysed the subset
426 of 114 European domesticated accessions (**Supplementary Table 5b**).
427 For GO enrichment, the annotated core genes and PAVs in the pan-genome were analysed using the
428 *buildGOmap* R function to infer indirect annotations and generate data suitable for *clusterProfiler* [61, 62].
429 Diagnostic genes were analysed using Metascape [63]. *A. thaliana* orthologs were identified using
430 OrthoFinder [58] and by comparing all protein sequences from *P. vulgaris* (v2.1) and *A. thaliana* (TAIR10).
431 For PCA, the PAV matrix (1/0) was analysed using the *logisticPCA* package in R [64].
432 ANOVA within subgroup M1 was carried out using the first principal component related to days-to-
433 flowering and photoperiod sensitivity [13] as a representative phenotypic trait.
434 F_{ST} analysis involved the separate testing of PAVs in the Mesoamerican and Andean gene pools by
435 comparing the frequency of each PAV between wild and domesticated forms. Each PAV was considered as
436 a single locus (0/1) and F_{ST} was calculated by applying the formula F_{ST} = (H total – H within) / H total, where
437 H is the heterozygosity [65]. The same procedure was applied to wild accessions when comparing the
438 Mesoamerican and Andean gene pools. Only PAVs in the top 5% of the F_{ST} distribution were considered as
439 candidates.
440 The functions of interesting PAVs and those associated with *A. thaliana* orthologs detected by OrthoFinder
441 [58] were investigated manually in the NCBI database (https://www.ncbi.nlm.nih.gov/).
442 Phylogenetic analysis involved the extraction and filtering of SNVs located in core genes and PAVs using
443 bcftools [59], resulting in two final datasets: 1,451,663 SNPs for the core genes and 338,475 SNPs for the
444 PAVs. The datasets were used to calculate the genetic distance between individuals and compute maximum
445 composite likelihood values with 1000 bootstraps for the NJ tree in MEGA11 [66]. The final trees were
446 visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).
447 The filtered dataset of SNPs in core genes was also used to quantify the genetic diversity within groups of
448 accessions by estimating π. The *--window-pi* vcftools flag was used to obtain measures of nucleotide
449 diversity in 250-kb windows. The windowed-*pi* estimates were then divided by the total number of SNPs to
450 calculate a global estimate for each genetic group.

451 Fisher's exact test with the false discovery rate corrected for multiple comparisons was applied in R to
452 identify PAVs that differed significantly in frequency between the Mesoamerican and Andean gene pools
453 for the American and European accessions.
454 The phenotypic data used for PAV-based GWAS encompassed the flowering time and photoperiod
455 sensitivity data previously analysed by Bellucci et al. [13]. GWA analysis was run by using both the Mixed
456 Linear Model (MLM) [67] and the Fixed and random model Circulating Probability Unification (FarmCPU)
457 [68] implemented in the R package GAPIT v3 [69]. The threshold for each Genome Wide Association (GWA)
458 scan was determined by the Bonferroni corrected *p* value at $\alpha = 0.05$. The kinship matrix (IBS method) was
459 calculated with Tassel 5 [70] and the population structure (at K2 obtained from Bellucci et al. [13]) were
460 included into the models as fixed factors. Quantile-quantile (Q-Q) plots were obtained by plotting the
461 observed -log10(*p* values) against the expected -log10(*p* values) under the null hypothesis of no association.
462

463 **Data Availability**

464 The 109 raw sequencing reads generated and analyzed in this study have been deposited in the Sequence
465 Read Archive (SRA) of the National Center of Biotechnology Information (NCBI) under BioProject number
466 PRJNA1042929. Additional data comprising 230 raw sequencing reads have been sourced from Frascarelli
467 et al. [11] and Bellucci et al. [13]. The pan-genome assembly and its annotation are publicly accessible via
468 this link: https://doi.org/10.6084/m9.figshare.24573874.
469

470 **References**

471 1. Bellucci, E., Mario Aguilar, O., Alseekh, S., Bett, K., Brezeanu, C., Cook, D., De la Rosa, L.,
472 Delledonne, M., Dostatny, D.F., Ferreira, J.J., et al. (2021) The INCREASE project: Intelligent
473 Collections of food-legume genetic resources for European agrofood systems. *Plant Jour*nal,
474 108:646-660. DOI: https://doi.org/10.1111/tpj.15472

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

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

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

484 5. Gerten, D., Heck, V., Jägermeyr, J., Bodirsky, B.L., Fetzer, I., Jalava, M., Kummu, M., Lucht, W.,
485 Rockström, J., Schaphoff, S. et al. (2020) Feeding ten billion people is possible within four terrestrial
486 planetary boundaries. *Nature Sustainability*, 3(3):200-208. DOI: https://doi.org/10.1038/s41893-
487 019-0465-1

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

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

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

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

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

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

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

513 13. Bellucci, E., Benazzo, A., Xu, C., Bitocchi, E., Rodriguez, M., Alseekh, S., Di Vittori, V., Gioia, T.,
514 Neumann, K., Cortinovis, G., et al. (2023) Selection and adaptive introgression guided the complex

515 evolutionary history of the European common bean. *Nature Communications*, 14, 1908. DOI:
516 https://doi.org/10.1038/s41467-023-37332-z

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

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

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

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

527 18. Bellucci, E., Bitocchi, E., Ferrarini, A., Benazzo, A., Biagetti, E., Klie, S., Minio, A., Rau, D., Rodriguez,
528 M., Panziera, A., et al. (2014) Decreased nucleotide and expression diversity and modified
529 coexpression patterns characterize domestication in the common bean. *The Plant Cell*, 26(5):1901-
530 1912. DOI: https://doi.org/10.1105/tpc.114.124040

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

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

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

541 22. Oladzad, A., González, A., Macchiavelli, R., de Jensen, C.E., Beaver, J., Porch, T., and McClean, P.
542 (2020) Genetic factors associated with nodulation and nitrogen derived from atmosphere in a
543 middle American common bean panel. *Frontiers in Plant Science*, 11:576078. DOI:
544 https://doi.org/10.3389/fpls.2020.576078

545 23. Ke, D., Fang, Q., Chen, C., Zhu, H., Chen, T., Chang, X., Yuan, S., Kang, H., Ma, L., Hong, Z., et al.
546 (2012) The small GTPase ROP6 interacts with NFR5 and is involved in nodule formation in Lotus
547 japonicus. *Plant Physiology*, 159(1):131-143. DOI: https://doi.org/10.1104/pp.112.197269

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

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

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

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

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

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

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

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

568 32. Wolf, Y.I. and Koonin, E.V. (2013) Genome reduction as the dominant mode of evolution. *Bioessays*,
569 35:829-837. DOI: https://doi.org/10.1002/bies.201300037

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

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

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

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

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

584 38. Ruan, J. and Li, H. (2019) Fast and accurate long-read assembly with wtdbg2. *Nature Methods*,
585 17:155-158. DOI: https://doi.org/10.1038/s41592-019-0669-3

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

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

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

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

598 43. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094-
599 3100. DOI: https://doi.org/10.1093/bioinformatics/bty191

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

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

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

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

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

615 49. Langmead, B. and Salzberg, S. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*,
616 9:357-359. DOI: https://doi.org/10.1038/nmeth.1923

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

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

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

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

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

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

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

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

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

642 59. Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and
643 population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987-
644 2993. DOI: https://doi.org/10.1093/bioinformatics/btr509

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

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

651 62. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021).
652 clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*,
653 2(3):100141. DOI: https://doi.org/10.1016/j.xinn.2021.100141

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

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

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

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

664 67. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S.,
665 Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping

666 that accounts for multiple levels of relatedness. *Nature Genetics,* 38(2):203-8. DOI:
667 https://doi.org/10.1038/ng1702

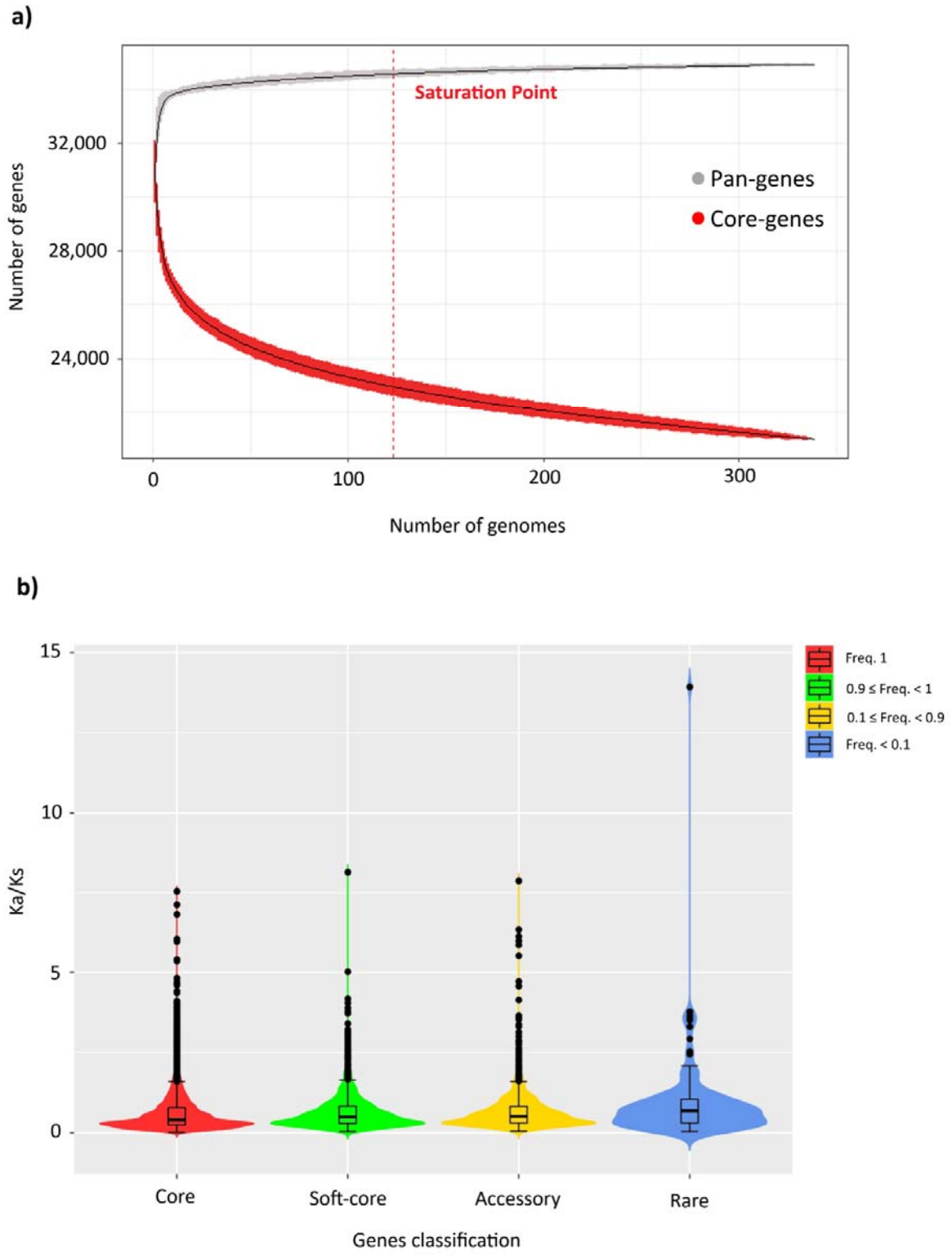
668 68. Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative usage of fixed and random
669 effect models for powerful and efficient genome-wide association studies. *PLoS Genetics,*
670 12(2):e1005767. DOI: https://doi.org/10.1371/journal.pgen.1005767

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

674 70. Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007)
675 TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics,*
676 23(19):2633-5. DOI: https://doi.org/10.1093/bioinformatics/btm308

677

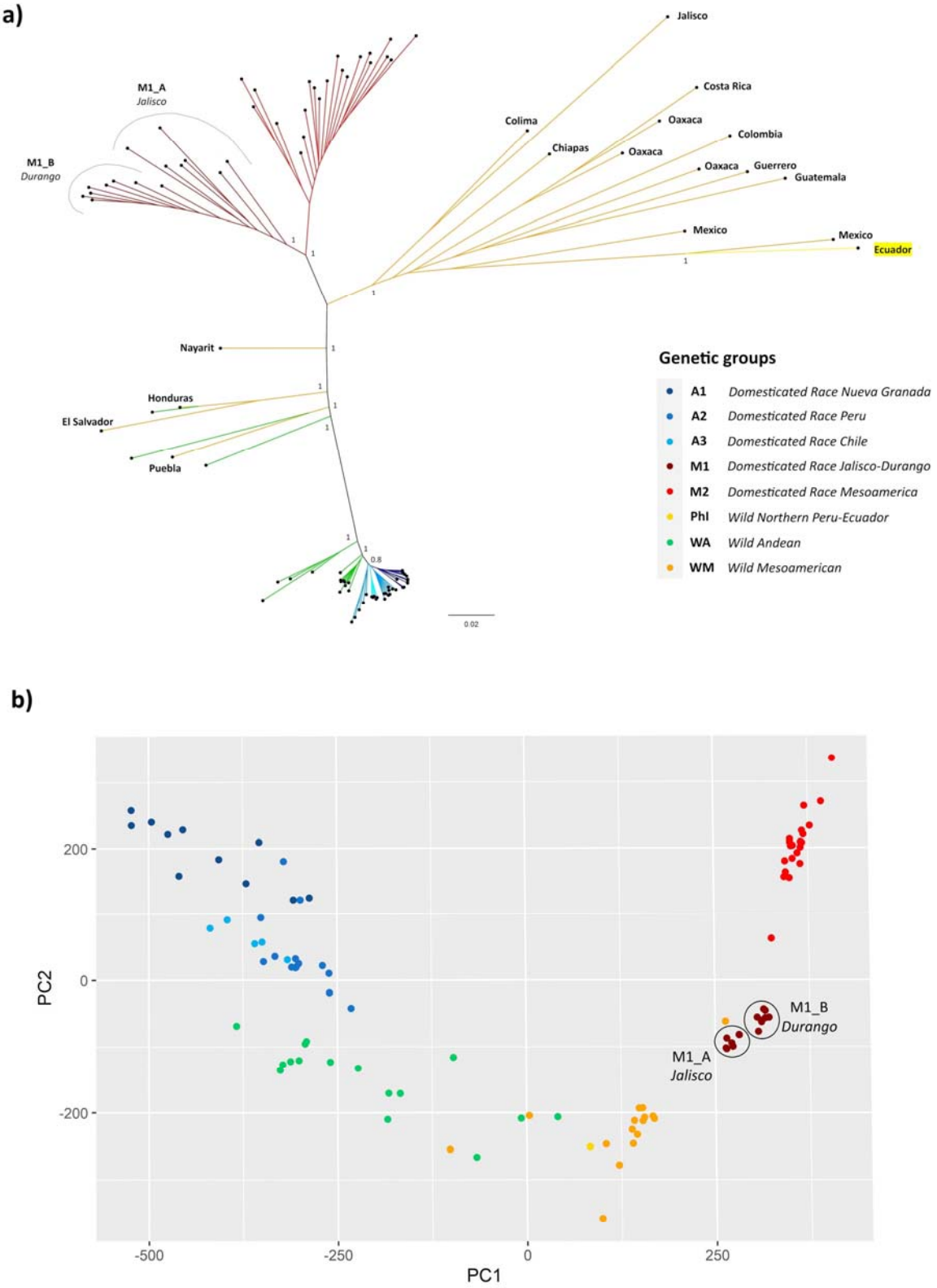
678 **Figures**



679

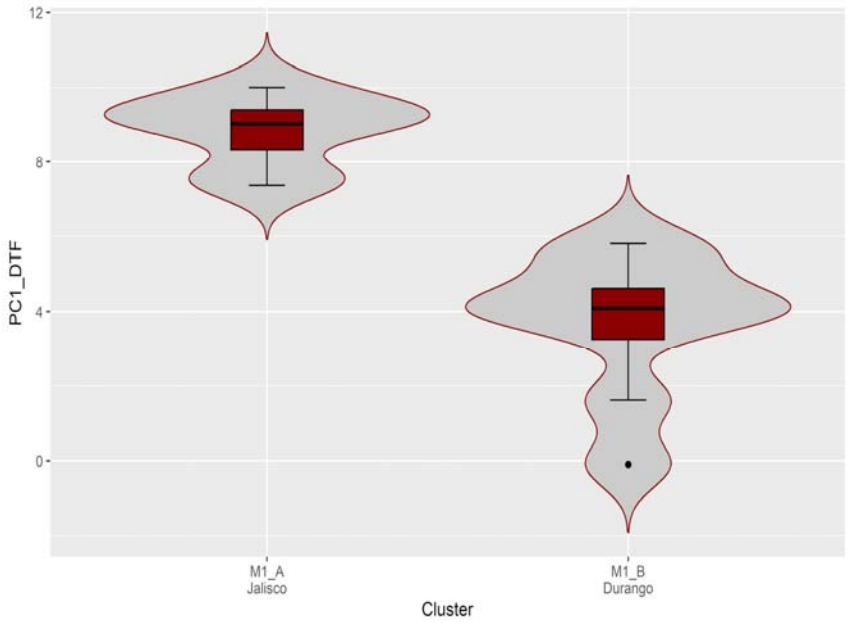
680 **Fig. 1: Characterization of the common bean pan-genome. a,** Pan-gene and core gene size calculation. The
681 growth curve of pan-genes (grey) reached saturation point (99%, 34,579 genes) when 120 individuals were

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

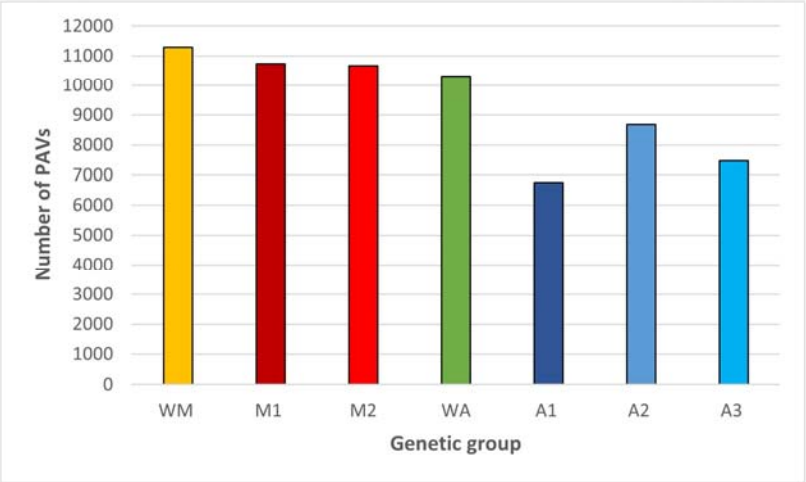

a)
Jalisco
Costa Rica
Oaxaca
Colombia
Oaxaca
Guerrero
Guatemala
Mexico
Mexico
Ecuador
Colima
Chiapas
Oaxaca
M1_A
Jalisco
M1_B
Durango
Nayarit
Honduras
El Salvador
Puebla
0.02
Genetic groups
A1 Domesticated Race Nueva Granada
A2 Domesticated Race Peru
A3 Domesticated Race Chile
M1 Domesticated Race Jalisco-Durango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

687

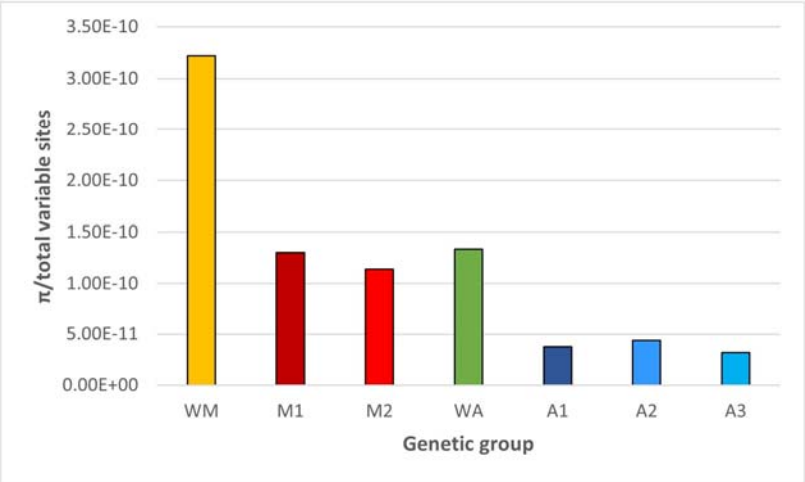
c)


12
8
4
0
PC1_DTF
M1_A
Jalisco
M1_B
Durango
Cluster

d)

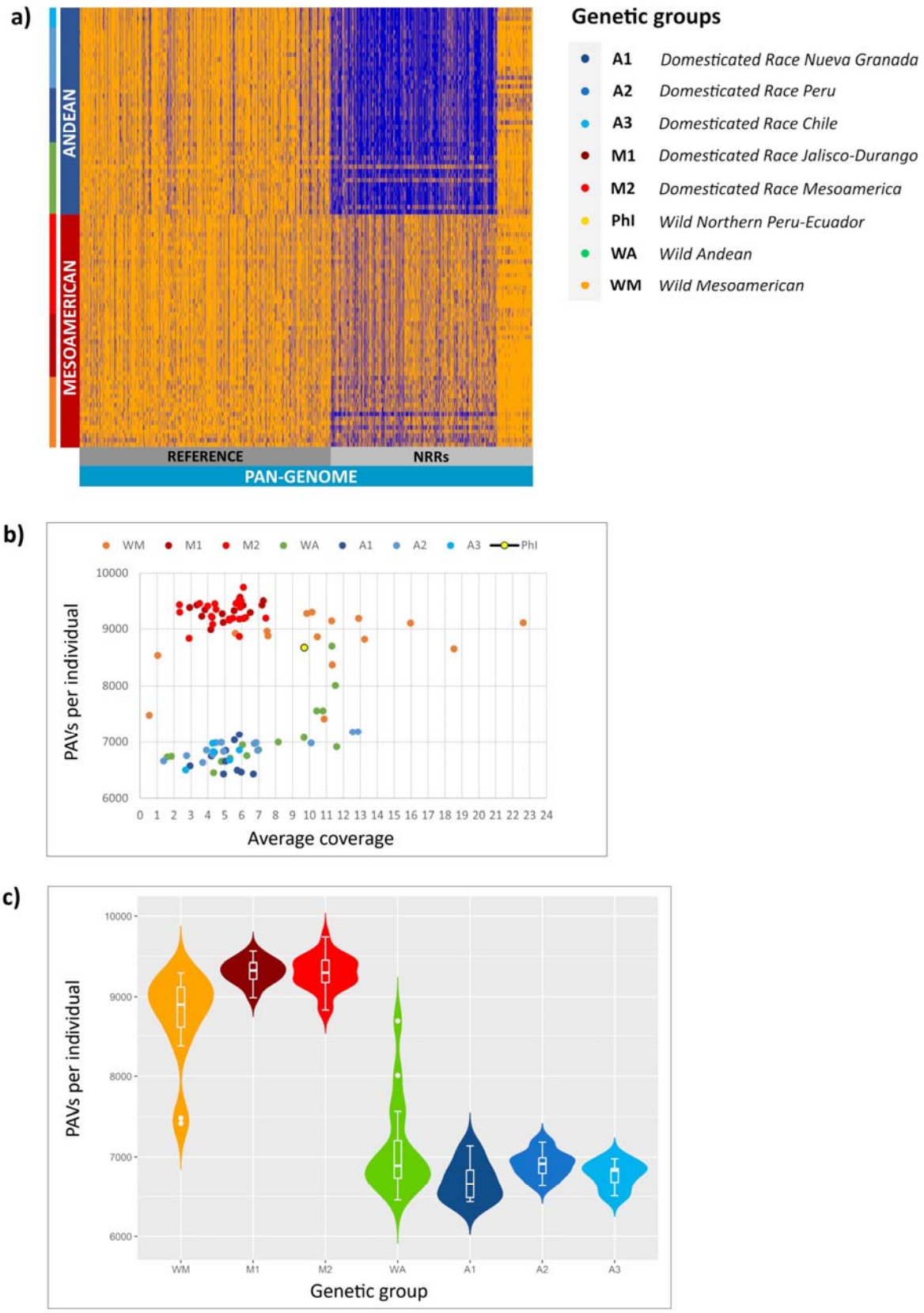

12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Number of PAVs
WM
M1
M2
WA
A1
A2
A3
Genetic group

e)


3.50E-10
3.00E-10
2.50E-10
2.00E-10
1.50E-10
1.00E-10
5.00E-11
0.00E+00
π/total variable sites
WM
M1
M2
WA
A1
A2
A3
Genetic group

688

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

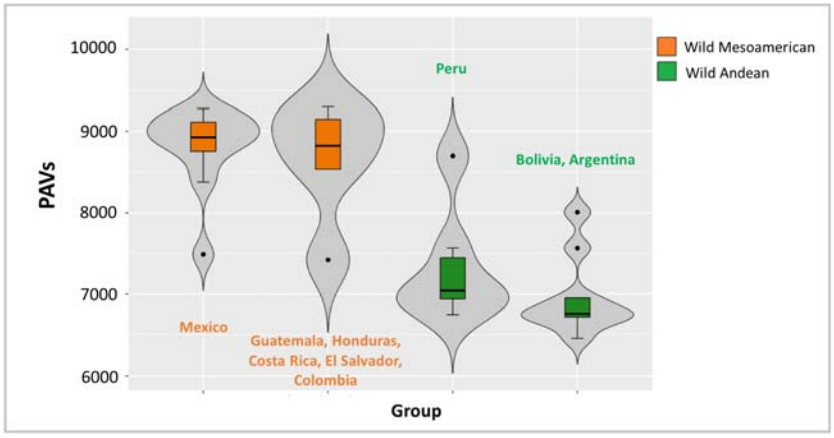

a)
ANDEAN
MESOAMERICAN
REFERENCE
NRRs
PAN-GENOME
Genetic groups
A1 Domesticated Race Nueva Granada
A2 Domesticated Race Peru
A3 Domesticated Race Chile
M1 Domesticated Race Jalisco-Durango
M2 Domesticated Race Mesoamerica
PhI Wild Northern Peru-Ecuador
WA Wild Andean
WM Wild Mesoamerican
b)
WM M1 M2 WA A1 A2 A3 PhI
PAVs per individual
Average coverage
c)
PAVs per individual
Genetic group
WM M1 M2 WA A1 A2 A3

696

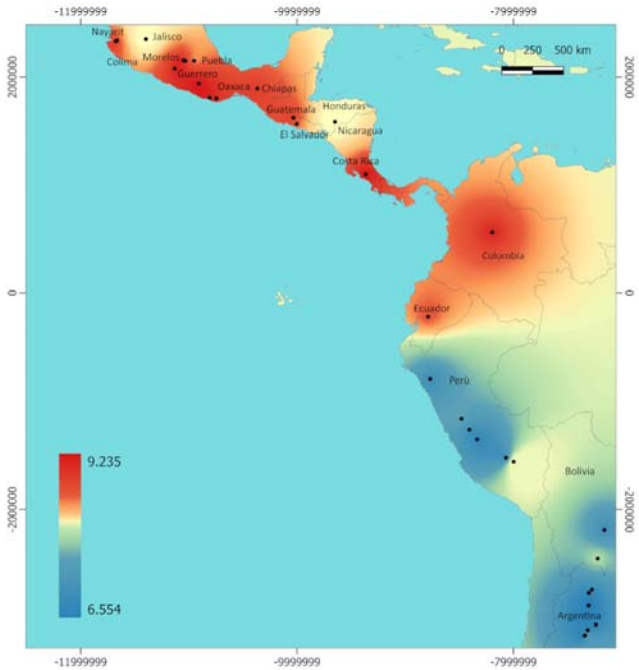
697 **Fig. 3: Evolution of the common bean pan-genome. a,** Heat map illustrating the number of PAVs per
698 individual in the final pan-genome. Orange indicates presence while blue indicates absence. **b,** Scatterplot
699 showing the number of PAVs per individual (y-axis) in relation to the coverage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

703

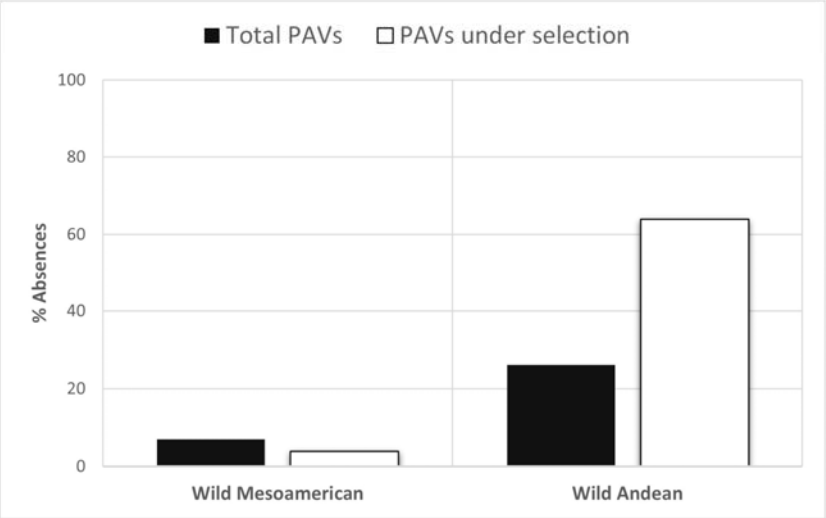
a)


Wild Mesoamerican
Wild Andean
PAVs
10000
9000
8000
7000
6000
Mexico
Guatemala, Honduras, Costa Rica, El Salvador, Colombia
Peru
Bolivia, Argentina
Group

b)


-11999999
-9999999
-7999999
2000000
0
-2000000
Nayarit
Jalisco
Colima
Morelos
Puebla
Guerrero
Oaxaca
Chiapas
Guatemala
Honduras
El Salvador
Nicaragua
Costa Rica
Colombia
Ecuador
Perù
Bolivia
Argentina
0 250 500 km
9.235
6.554

c)

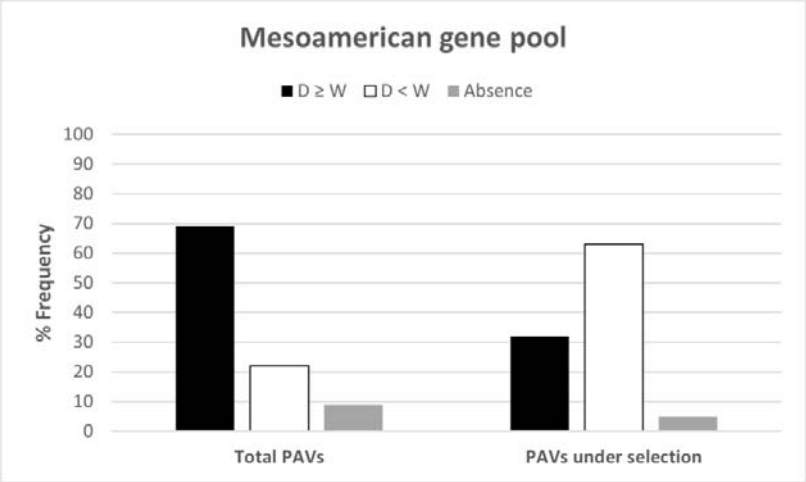

Total PAVs
PAVs under selection
% Absences
100
80
60
40
20
0
Wild Mesoamerican
Wild Andean

704

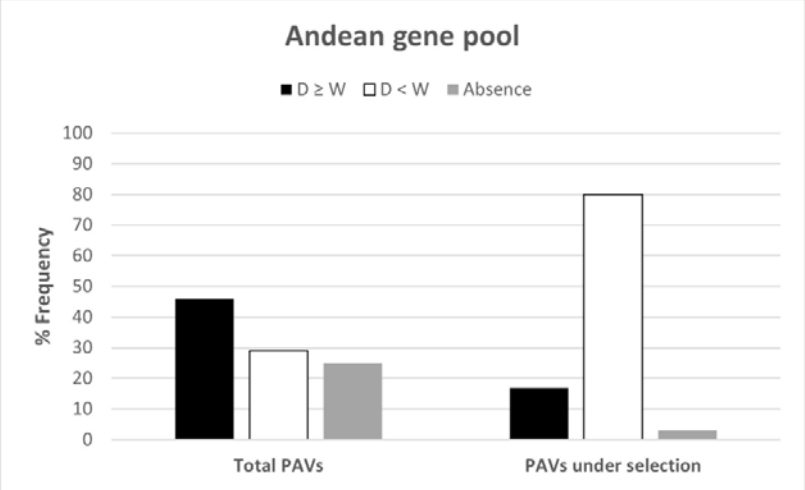
705 **Fig. 4: Selection for adaptive gene loss during the expansion of wild common bean. a,** Violin plots showing
706 the analysis of variance (ANOVA) for the number of PAVs per individual based on grouping the wild
707 Mesoamerican and Andean accessions according to latitude coordinates. **Supplementary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

712

a)



b)



713

714 **Fig. 5: Localized adaptive reduction effects during the domestication of the common bean. a,** Bar chart
715 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.