Selection and adaptive introgression guided the complex evolutionary history of the European common bean

Bellucci E.[1]*, Benazzo A.[2]*, Xu C.[3]*, Bitocchi E.[1]*, Rodriguez M.[4,5]*, Alseekh S.[6,7]*, Di Vittori V.[1,6]*, Gioia T.[8], Neumann K.[10], Cortinovis G.[1], Frascarelli G.[1], Murube E.[1], Trucchi E.[2], Nanni L.[1], Ariani A.[9], Logozzo G.[8], Shin J.H.[3], Liu C.[11], Jiang L.[6], Ferreira J.J.[12], Campa A.[12], Attene G.[4], Morrell P.L.[11], Bertorelle G.[2], Graner A.[10]*, Gepts P.9*, Fernie A.R.[6,7]*, Jackson S.A.[3]*, Papa R.[1]*[#]

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

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

[3] Center for Applied Genetic Technologies, University of Georgia, 30602 Athens, GA, USA

[4] Department of Agriculture, University of Sassari, 07100 Sassari, Italy

[5] Centro per la Conservazione e Valorizzazione della Biodiversità Vegetale—CBV, Università degli Studi di Sassari, 07041 Alghero, Italy

[6] Max Planck Institute of Molecular Plant Physiology (MPI-MP), 14476 Potsdam-Golm, Germany

[7] Center for Plant Systems Biology, 4000 Plovdiv, Bulgaria

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

[9] Department of Plant Sciences, University of California, 95616-8780 Davis, CA, USA

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

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

[12] Regional Agrifood Research and Development Service (SERIDA), 33310, Villaviciosa, Asturias, Spain

*These authors contributed equally to this work

[#]**Corresponding author: R. Papa**

Department of Agricultural, Food and Environmental Sciences
Marche Polytechnic University
Via Brecce Bianche 60131 Ancona, Italy

Abstract

Domesticated crops have been disseminated by humans over vast geographic areas. After 1492, the common bean (*Phaseolus vulgaris* L.) was introduced in Europe. By combining whole-genome profiling, metabolic fingerprinting and phenotypic characterisation, we found that the first common bean cultigens successfully introduced into Europe were of Andean origin, after Francisco Pizarro's expedition to northern Peru in 1529. We found that hybridisation, selection and recombination have shaped the genomic diversity of the European common bean in parallel with political constraints. There is clear evidence of adaptive introgression into the Mesoamerican-derived European genotypes, with 44 Andean introgressed genomic segments shared by more than 90% of European accessions and distributed across all chromosomes except PvChr11. Genomic scans for signatures of selection highlighted the role of genes relevant to flowering and environmental adaptation, suggesting that introgression was crucial for the dissemination of this tropical crop to the temperate regions of Europe.

Introduction

Following the process of domestication, crops were spread by humans over vast geographic areas, where they adapted to new and often extreme environments (*1*). The Columbian Exchange (*2*) started in 1492 with the transatlantic journey of Christopher Columbus. This large-scale set of reciprocal biological introductions between continents provides a paradigm for the rapid adaptation of crop plants to changing environments. Changes in flowering time and photoperiod sensitivity were selected in parallel in the common bean (*Phaseolus vulgaris*), maize (*Zea mays*), potato (*Solanum tuberosum*), to name a few crops that have undergone selection for these adaptive traits (*1, 3*). Among crops originating from the Americas, the common bean was rapidly adopted and successfully disseminated across Europe (*4*) and it is now possible to identify local European varieties with Andean and Mesoamerican origins (*5, 6, 7, 8, 9, 10*). Nowadays, common bean as other food legumes is crucial for main societal challenges and to promote the transition to plant-based diets (*11*).

The introduction of the common bean to Europe from two distinct centres of origin offered an opportunity for widespread genepools hybridisation and recombination (*9*). Studies of common bean evolution in Europe can exploit the parallel domestication processes and the major genetic differences between the two American genepools. This provides an ideal model to study the role of introgression during the adaptation of common bean accessions to European environments (*12*).

Here, we combine whole-genome analysis and metabolic fingerprinting in 218 common bean landraces, integrated with genome-wide association (GWA) to characterise the genetic basis of multiple traits, including flowering time and growth habit in different environments with contrasting photoperiod conditions. We used the combined results to characterise the effects of selection and inter-genepool introgression, and to test the occurrence of adaptive introgression associated with the development and adaptation of common bean accessions in Europe.

The common bean population structure reveals pervasive admixture in Europe

Using ADMIXTURE (*13*), we reconstructed the ancestry of 218 single-seed-descent (SSD) purified accessions from the Americas (104 accessions collected as reported in SN4_Fig. 13, including 66 pure American accessions showing low admixture between gene pools, qi > 99%) and from Europe

(n=114), based on nuclear and chloroplast genetic variants (Fig. 1a–d). The accessions were spatially interpolated to investigate their geographic distribution in Europe (Fig. 1e) and we proposed a model for their introduction in the Old Continent (Fig. 1f). For common beans originating from the Americas, subdivision into the highly differentiated Andean and Mesoamerican populations was consistent with previous studies (*11*, *13*, *14*) (Fig. 1a). Next, we followed the nested procedure previously used by Rossi et al. (*16)* to investigate the population structure within each American gene pool (Supplementary Note 4.2). Again, using ADMIXTURE (*13*), we identified two main Mesoamerican groups (**M1** and **M2**) and three main Andean groups (**A1**, **A2** and **A3**) (Fig. 1b, c). In the centres of domestication, there was little evidence of admixture between genepools.

Among the European accessions, nuclear variants allowed us to identify several admixed genotypes (35 EU accessions with more than 10% and 18 with more than 20% of the genome attributed to the other gene pool based on the admixture; SN4_Supplemental Dataset 4) (Fig. 1a). In a few European accessions (n = 14), the nuclear and chloroplast assignments were inconsistent (Fig. 1a): considering only accessions with introgressed genome >70% (n=11), the Andean chloroplast genome was combined with a Mesoamerican nuclear genome (n =6) or *vice versa* (n=5). This suggests, at least in some cases, the occurrence of chloroplast capture (*17*) as a result of inter-genepool hybridisation and subsequent backcrossing. Moreover, when we explored the molecular phenotypic diversity of the American and European accessions (Supplementary Note 2.3), the metabolomic fingerprint (the molecular phenotypic space expressed as principal component 1 from 1493 putative secondary metabolites with a high hereditability of $H^2 > 0.65$) confirmed the admixture scenario. Several intermediate phenotypes between Mesoamerican and Andean accessions were observed in European landraces, but these were absent in accessions collected in the Americas (Fig. 2a-c). Notably, there was a significant correlation between the admixture coefficients and principal component 1 for both the American and the European accessions, indicating a tight relationship between the phenotypic and genotypic differences due to the genepool structure. This included a reduced difference in Europe due to admixture, particularly in the accessions of Mesoamerican origin.

Finally, to provide insight into the population structure of the American pure accessions, we considered their passport data (geographic distribution and country of origin) and phenotypic data on

growth habits and photoperiod sensitivities. We identified a clear correspondence between the genetic groups in our sample from the Americas and the well-known common bean eco-geographic races (*18*): **M1** corresponded to the higher-altitude Durango and Jalisco races, which originated primarily in northern and southern Mexico, respectively; **M2** corresponded to the lower-altitude Mesoamerican race, which is mostly photoperiod-insensitive and is distributed in lowland Mexico, in Central America and in the Caribbeans ; **A1** corresponded to the generally photoperiod-insensitive Nueva Granada race; **A2** corresponded to the Peru race, which includes entries with vigorous climbing growth habits and photoperiod sensitivity; and **A3** corresponded to the Chile race, which has also been identified in archaeological samples from northern Argentina that date from 2500 to 600 years ago (*19*). The identification of these well-defined ancestral genetic groups in the Americas offers a robust basis to study the inter-genepool and inter-race introgression that may have facilitated adaptation to European environments.

Asymmetric introgression and recombination between genepools underlie European common bean adaptation

Given the presence of admixed individuals, we studied the inter-genepool hybridisation and introgression pattern associated with the evolutionary history of the common bean in Europe by genetic assignment at the chromosome level in ChromoPainter v2.0 (*20*) (Supplementary Note 4). When compared to admixture analysis, this can provide information on recombination between markers and the size of regions that can be attributed to different ancestries. The 66 pure American accessions, from the five genetic groups identified using admixture, due to their low levels of admixture, were used as donor (reference/founder) populations for the chromosome painting of the European genotypes. On this basis, and for each European genotype, we attributed all the single-nucleotide polymorphisms (SNPs) and chromosomal regions to particular ancestries, taking into account within-accession recombination breakpoints (Supplementary Note 4). Using this approach, we were also able to detect recombination events between genepools at the whole-genome level, even in accessions that showed < 1% introgression. Overall, 71 European accessions were attributed to the Andean gene pool (EU_AND) and 43 were assigned to the Mesoamerican gene pool (EU_MES), in agreement with the admixture

analysis (Pearson correlation: r = 0.99, p < 0.01) and confirming previous knowledge about the prevalence of Andean genotypes in Europe (*5*, *21*). Globally, the inferred amount of per-accession introgressed material differed between the EU_MES and EU_AND accessions (two-sided K-S test, p = 3.3×10^{-3}), showing median proportions of 4.7% and 9.2%, respectively. In the EU_MES accessions, 0.01–44.9% of the genome had introgressed from the other genepool, with only one EU_MES accession showing < 1% genome introgression (Fig. 3a; Supplementary Note 4). These proportions were similar in the EU_AND samples, ranging from 0% to 42.2% (Fig. 3a), although two EU_AND accessions showed no introgression and 10 EU_AND accessions showed < 1% introgression from the other genepool (Fig. 3a). The pervasive effect of admixture in European individuals was confirmed by the presence of several accessions with > 20% of their genome acquired by introgression in both the EU_MES (8 of 43 accessions, 18.6%) and EU_AND (11 of 71 accessions, 15.5%) groups (Fig. 3a).

The median length of the introgressed genomic segments was higher for the EU_AND accessions (EU_AND = 217 kb, EU_MES = 70 kb; Mann-Whitney test, $P=7.22\times10^{-10}$, Fig. 3b,c), with more extended regions introgressed into EU_AND particularly on chromosomes PvChr02, PvChr05, PvChr06 and PvChr09 (Fig. 3b,c). We obtained very similar results when we repeated the analysis by excluding six accessions showing an admixture proportion >=40% (SN4_Fig.29). The EU_AND accessions carried longer Mesoamerican introgressed haplotypes, reflecting the more recent introgression of Mesoamerican genome fragments into the Andean genotypes compared to the opposite direction. When we estimated the timing of the introgression among the two gene pools in Europe, we confirm the more recent introgression from the Mesoamerican to the Andean gene pool, while the Mesoamerican gene pool was introgressed earlier (SN4_Fig.30). This estimated introgression time is compatible with historical data and with an earlier successful introduction of the Andean gene pool in Europe. Several genomic regions that carry haplotypes with a specific Andean ancestry are near fixation in the European accessions. Here, when seeking regions that may have been subject to natural selection, playing a role in the adaptation, we identified regions putatively under selection in Europe (Supplementary Note 6). Signatures of selection were detected among the regions with Andean ancestry nearly fixed in the European accessions (e.g., position 46 Mb on chromosome Pv01, which carries the *OTU5* locus, that may be involved in the phosphate starvation response; Supplementary Note 4,

SN4_Fig. 26; and position 37.9 Mb on chromosome Pv9 which carries the *LHY* locus; SN6_Supplemental Dataset 9).

Our results indicate that the first cultigen successfully disseminated across Europe was composed of Andean types. This is shown by the smaller introgression segments of Andean origin and the higher frequencies of Andean-derived common bean accessions in Europe. Our data are also consistent with available historical records. Indeed, the first unambiguous evidence for the introduction of common bean in Europe points to Andean cultivars (*22*) probably introduced into Spain by Francisco Pizarro in 1529 following the exploration of Peru. Piero Valeriano Bolzanio received common bean seeds from Giuliano de Medici (Pope Clement VII, 1523–1534), which had been donated to the same pope by Emperor Charles V's Spanish emissaries from Sicily (where the bean seeds were harvested). The very detailed writing of Piero Valeriano Bolzanio refers to common bean seeds, describing in depth several phenotypic traits supporting their Andean origin, as also recently proposed (*23*). Valeriano documented his efforts, along with a network of collaborators in the north-east of Italy, Slovenia and Dalmatia, to grow and reproduce beans starting in 1532 (*22*), with the first report of a putative Mesoamerican genotype in Europe dated 1542 (*23*). Historical information and timelines, together with our data showing asymmetric introgression, suggest an earlier successful introduction and spread of the Andean genepool into Europe. This may also explain the high frequency of A1/Nueva Granada Andean ancestries (Fig. 1e) in Sicily, the south and north-east of Italy, Slovenia and Croatia, because they could have been among the first European areas to cultivate common bean with the earliest introduced Andean genotypes probably from the Nueva Granada race.

Adaptive differences among common beans in the New World may also have influenced the distributions in Europe. For example, M1/Durango-Jalisco genotypes can be extremely photoperiod sensitive, and may therefore have failed to adapt well to many European environments, thus, limiting their dissemination, particularly in central and northern Europe (Fig. 1e). In contrast, southern Spain, southern Italy, Sicily, North Africa, Madeira Island, and the Canary Islands are characterised by mild winters. In these environments, photoperiod-sensitive and late-flowering genotypes, or those adapted to warmer conditions, may have easily completed the crop cycle. As also reported by others (*5*, *24*), we found that Mesoamerican genotypes are more frequent in specific European regions, particularly in

south-eastern Europe (Fig. 1e), which also suggests that the history of their introduction may have contributed to their current distribution. As with the role of Charles V and Pope Clement VII in the early dissemination of the Andean beans, the political subdivision of Europe and the Mediterranean basin in the 16th century may have influenced the dissemination of the Mesoamerican genepool. The Ottoman Empire dominated the southern shores of the Mediterranean, the Nile Basin, the Red Sea into eastern Africa, and south-eastern Europe, spanning the area from modern-day Greece to Austria. The prevalence of Mesoamerican genotypes in eastern Africa and China (*24*, *25*) may reflect their initial introduction into Africa from Spain during the Ottoman Empire, which extended its rule in north-eastern Africa and controlled the exchange of goods with China through the Silk Road. Although additional comparative studies between European and Chinese centres are required, our hypothesis is compatible with our results from a *de-novo* admixture analysis applied to Chinese landraces (*25*), shown in Supplementary Note 4 (SN4_Fig. 20). The importance of political/cultural factors associated with the dissemination of common bean genotypes in Europe is compatible with the lack of significant spatial and ecological patterns between genetic, geographic and ecological distances. Indeed, the routes of dissemination based on cultural and political factors are often independent of geographic and environmental distances, making the occurrence of correlations between genetic distances and geographical or environmental differences less likely (*26*, *27*, *28*, *29*, *30*) (Supplementary Note 4).

Analysis of the Environmental associations

We used the geographic distribution of the five ancestral components inferred by ChromoPainter in an association analysis with biogeographical variables (Supplementary Note 4). Ancestral components of A3/Chile negatively correlated with latitude (Supplementary Note 4, SN4_Tab. 8; $r = -0.35$, $p = 0.0001$) and were never observed above the 47th parallel (Fig. 1e). Moreover, A3/Chile component was associated with warmer climates, particularly the maximum temperature in September (Supplementary Note 4; SN4_Tab.8; $r = 0.29$, $p < 0.002$). Although A3/Chile did not appear any more photoperiod-sensitive than A1/Nueva Granada (Supplementary Note 5; SN5_Tab. 12), some American A3/Chile individuals tend to flower later (Fig. 2d) at higher latitudes when grown in Europe (Fig. 2e). Here, we suggest that although A3/Chile was successfully introduced in Europe, a residual sensitivity to the

photoperiod might be still preserved in some European genotypes mainly belonging to this ancestry, that show delayed flowering at certain latitudes (Fig 2e), that may have also influenced the dissemination of this common bean ancestry in Europe (Fig. 1e,f). However, compared to the Mesoamerican genotypes, A3/Chile was more uniformly distributed in Europe across different longitudes (Fig. 1e), which also supports the earlier introduction of Andean genotypes. Only a few weak associations with environmental variables were detected for the other genetic groups (Supplementary Note 4).

Analysis of genetic diversity in the European common bean

To disentangle how inter-genepool hybridisation have shaped the genetic diversity of the European common bean, and given the evidence of widespread admixture in Europe, we developed a masked dataset of European accessions by filtering out all introgressed alleles or those with an ambiguous assignment (Supplementary Note 4). This allowed us to consider nucleotide diversity using the frequencies of two reconstructed non-admixed populations of Andean and Mesoamerican origin. From each European genotype, all the Andean SNPs were separated from the Mesoamerican SNPs and included in the two masked datasets. Based on the unmasked and masked datasets, American common bean accessions showed moderately higher nucleotide diversity than European accessions (Supplementary Note 4.3), apparently due to the introduction bottleneck in Europe (Fig. 4a,b). Moreover, when compared to the Mesoamerican gene pool, the Andean gene pool showed an overall lower diversity in the primary centres of domestication (Americas) using both the masked and unmasked datasets (Fig 4a, b). This confirms that the diversity of the Andean germplasm at the centre of origin might still reflect the bottleneck that occurred in the Andean wild populations during the expansion into South America before domestication (*31*), as reflected in the domesticated pool (*32*). Indeed, we detected ~70% lower diversity (θ_π/bp) in the Andean compared to the Mesoamerican accessions. Very similar results were obtained when we repeated the analysis by excluding six accessions showing an admixture proportion >=40% (Supplementary Note 4).

When American and European genetic diversities were compared within each genepool using the unmasked dataset (AM_AND *vs* EU_AND and AM_MES *vs* EU_MES), due to the admixture,

European diversity was always higher than American diversity, but the opposite was found when using the masked dataset (Fig. 4b). In other words, we show how the Andean common beans from Europe are more diverse than those from America because of admixed ancestry with the Mesoamerican gene pool, as seen by comparing the genetic diversities of the unmasked and masked datasets (EU_AND). This comparison of the estimated levels of genetic diversity in Europe reflects the key role of inter-genepool hybridisation and recombination in shaping the diversity of the European common bean. The genetic diversity was higher in M1/Durango-Jalisco than M2 Mesoamerica accessions in Americas, and also in A2/Peru than A1/Nueva Granada accessions in Americas, whereas the amount of diversity in A3/Chile accessions was very low. Combined with the neighbourhood-joining tree shown in Fig. 1d, this indicates that the A2/Peru and M1/Durango-Jalisco races were probably the first domesticated Andean and Mesoamerican populations, from which the other races arose by secondary domestication associated with the loss of photoperiod sensitivity (Fig. 1f). Indeed, earliness and loss/reduction of photoperiod sensitivity were important traits under selection during the expansion of the common bean in Europe. This is also suggested by our test for the occurrence of selection for flowering during the introduction of common bean in Europe. The genetic differentiation between American and European accessions for flowering time (PC1 based on flowering data across different European field and greenhouse trials) was measured using Q_{ST} (*33*). The Q_{ST} for flowering was compared with the distribution of the Q_{ST} for highly hereditable metabolites and with the F_{ST} distribution of the SNPs (Supplementary Note 5). We show that the Q_{ST} for the flowering is in the top 97.5% of distribution of the Q_{ST} for highly inheritable metabolites, being also an outlier (99.5%) compared to our F_{ST} distribution, suggesting that flowering is likely a candidate trait that underwent a selection process (*33*, *34*). Considering the Andean genepool, the successful introduction in Europe was connected to the domestication pattern at the centre of origin. The earlier, photoperiod-sensitive domesticated genotypes were less successfully disseminated in Europe. Indeed, the relationship between the American and European genetic groups of Andean origin (as defined by ChromoPainter; Supplementary Note 4), coupled with the phenotypic data for flowering (Fig. 2d), shows that the A2/Peru race was more photoperiod-sensitive and was not introduced into Europe successfully, due to the lack of adaptation (other than a single highly admixed accession, qA2 = 43.6%). In contrast, the remaining Andean genetic

groups (A1/Nueva Granada and partially A3/Chile) became widespread in Europe. A different scenario was seen for the Mesoamerican genotypes, especially M1/Durango-Jalisco, where introgression appears to have been an important element in the dissemination of the common bean in Europe (Fig. 1f, 2d,e). M1/Durango-Jalisco showed very high levels of admixture in the European material due to introgression from the M2/Mesoamerica and the A1/Nueva Granada and A3/Chile (Fig. 2e), which likely contributed to reduced photoperiod sensitivity compared to the American Durango-Jalisco counterpart (AM_M1) (Fig. 2d), supporting its dissemination throughout Europe (Fig. 1e,f).

For the Andean genotypes, both the diversity pattern and photoperiod sensitivity (Fig. 2d) suggest at least two domestication steps occurred: primary domestication of photoperiod-sensitive populations (A2/Peru) and secondary domestication characterised by reduced photoperiod sensitivity (A3/Chile and particularly A1/Nueva Granada). This indicates that secondary domestication (*35*) was necessary for the successful dissemination of the Andean common bean in Europe (Fig 1f). For the Mesoamerican genotypes, an open question is where and when the introgression from the Andean genepool occurred. We suggest this is likely to have happened in southern Europe and along the southern Mediterranean shore, where the warmer climate in winter may have favoured the Mesoamerican genotypes.

The average linkage disequilibrium (LD) decay in accessions from Europe and the Americas (Fig. 4c) is consistent with the historical differences between the genepools and the effects of high inter-genepool hybridisation and introgression at the whole-genome scale in Europe. Admixture in Europe increased the molecular diversity (i.e., effective population size). It also generated new genome-wide admixture LD due to new combinations of alternative alleles in each genepool. Accordingly, inter-genepool hybridisation followed by recombination reduced LD at a long distance but, as expected, had a limited effect on LD decay at short distances because regions are directly inherited from the source populations (36). When we compared the American and European accessions, LD decay was much faster over short distances (< 1.5–2.0 Mb) in American genotypes. In contrast, there was faster LD decay over greater distances (> 3 Mb) in European populations (Fig. 4c). This reflects higher historical rates of recombination in the American genotypes over short distances and the effect of recombination due to inter-genepool introgression in Europe over long distances. A similar pattern was seen when the

Mesoamerican and Andean genepools were analysed separately (Supplementary Note 7, SN7_Fig.48). However, the Andean accessions were characterised by higher baseline LD levels. Indeed, the AM_MES and AM_AND populations reached r^2 = 0.2 at ~500 kb and ~1 Mb, respectively, whereas r^2 = 0.2 was reached at ~1.1 and ~3.5 Mb for the EU_MES and EU_AND samples, respectively.

Synonymous and missense mutations

The ratios between missense and loss-of-function mutations over synonymous mutations were calculated to reveal patterns of genetic load across genepools and continents. We observed a clear pattern in the genetic load reflecting differences between the Andean and Mesoamerican origins, with the Andean accessions showing a higher genetic load due to the bottleneck before domestication (Fig. 4d). We observed a reduced genetic load in EU_AND for both the loss-of-function (EU_AND *vs* AM_AND; Mann-Whitney p=0.03) and the missense mutations (EU_AND *vs* AM_AND; Mann-Whitney p=0.005). Conversely, we did not observe a reduction of genetic load in EU_MES for both missense and loss-of-function mutations (Fig. 4d). This suggests that the relatively short period of inter-genepool hybridisation, followed by selfing and recombination, promoted the purging of deleterious alleles accumulated in the European Andean pool. The role of hybridisation and subsequent recombination was also supported by the pattern of long-range LD in Europe compared to the Americas (Fig. 4c). The pattern for private alleles (i.e., those not identified in other genepools or populations) in the American and European accessions for low-frequency mutations (< 5%) revealed a 1.44-fold higher frequency of non-synonymous over synonymous mutations in Europe (Supplementary Note 4, SN4_Fig. 43). This may have resulted from the pattern of crop dissemination, which was probably characterised by the exchange of small quantities of seeds and several sequential bottlenecks, followed by rapid population growth at the single-farm level, leading to the fixation of most mutations due to the small population size (i.e., a founder effect). In this demographic context, most mutations would be fixed rapidly at the local level (within the population grown by a single farmer). However, it is also possible that the purging of deleterious mutations, due to hybridisation following seed exchange among farmers and the co-occurrence of different varieties in the same fields (*21*), facilitated the combined

effects of natural and human selection against deleterious recessive alleles and the capture of valuable variants.

Selection and adaptive introgression

We defined putative adaptive introgressed loci (PAIL) as those showing signatures of adaptive introgression meeting the following requirements: (a) an excess of introgression based on Chromopainter (Supplementary Note 6), (b) a signature of selection detected using the hapFLK method, which analyses multiple populations, jointly considering their hierarchical structure (37), and (c) an outlier F_{ST} value between Europe and the Americas, suggesting different patterns of diversity between these regions (Supplementary Notes 6 and 8). Although the hapFLK method allowed us to identify selection signatures across the genome, an outlier F_{ST} value was used to define which selection signature represents significant differentiation at the genomic level between American and European populations, suggesting selection in Europe. The identification of excess introgression independently of hapFLK provides evidence for adaptive introgression and the identification of PAIL. We also considered the occurrence of inter-chromosomal LD across genomic regions, private to European genepools (Fig. 4e, Fig. 5) as an interesting signal to define regions potentially involved in adaptive processes, and we identified, through a GWA analyses, genomic regions associated to flowering and growth habit (Supplementary Note 7). Adaptive introgression appears particularly important for the evolution of the European genotypes of Mesoamerican origin (EU_MES). We identified 44 Andean genomic regions with excess introgression (23 of which showed signals of adaptive introgression) that are shared by > 90% of the European genotypes, spanning all chromosomes except PvChr11 (SN6_Supplemental Dataset 9; F(AND)) and ranging from ~5 to ~118.5 kb in length (Supplementary Note 6). An Andean allele frequency of 96% was detected along a genomic segment of PvChr01 (Chr01:46175616–46294040; SN6_Supplemental Dataset 9) that shows signs of adaptive introgression. This region contains 18 genes including *Phvul.001G203400*, which is orthologous to *OVARIAN TUMOR DOMAIN-CONTAINING DEUBIQUITINATING ENZYME 5* (*OTU5*) (Supplementary Note 8, SN8_ Supplemental Dataset 15; row 16). In *Arabidopsis thaliana*, the function of this gene is to

recalibrate and maintain cellular inorganic phosphate homeostasis (*38*, *39)*. The common bean orthologue may therefore be involved in the phosphate starvation response, making it an interesting candidate for further testing. In the same region on chromosome Pv01, we also identified *Phvul.001G204600* and *Phvul.001G204700* (Supplementary Note 8; SN8_ Supplemental Dataset 15; rows 29 and 30), which are orthologous to *LUMINIDEPENDENS* and *NUTCRACKER*, respectively (Fig. 5).

This adaptive introgression region in the common bean genome is close to known regions associated with flowering time, such as the *fin* locus controlling determinacy, and explaining phenotypic variation also for flowering time (*40*). The region also co-maps with *Phvul.001G189200* (*PvTFL1y*; Pv01:44856139–44857862) (*41*), and shows linkage (*40*) to the *Ppd* locus controlling photoperiod sensitivity (*42*). Wu et al. (*25*) recently identified several markers on chromosome Pv01 associated with flowering under different conditions, with one at ~45.5 Mb on Pv01. In our GWA, we identified a significant association between photoperiod, flowering time (Supplementary Note 7.2) and the marker S01_48049738, which is found ~400 kb downstream of *Phvul.001G221100* (Chr01:47642033–47647745) a gene that has been proposed as a candidate for the common bean *Ppd* locus (*43*, *44*).

Overall, we identified 77 genes that are PAIL. These represent ~11% of all genes (n = 681) showing signatures of selection in Europe (i.e., selection signatures identified with hapFLK and being in an F_{ST} outlier window; n = 354) and/or excessintrogression (n = 404). Accordingly, 277 genes show selection in Europe but not excess introgression, and 327 show excess introgression but not selection in Europe. The 77 PAIL show enrichment in seven Gene Ontology categories including GO:0048523, negative regulation of cellular processes; GO:0010228, vegetative to reproductive phase transition of the meristem; GO:0042445, hormone metabolic processes; GO:0009657, plastid organisation; GO:0042440, pigment metabolic processes; GO:0009733, response to auxin; and GO:0070647, protein modification by small protein conjugation or removal (Supplementary Note 8.2). Enrichment analysis also suggested that flowering has been an important target of adaptive introgression, highlighting the importance of genes that may be associated with adaptation to abiotic and biotic stress.

Adaptive introgression signals were identified in many *P. vulgaris* genes with a putative role related to flowering (Supplementary Note 8.3), including orthologues of genes involved in the four major *A. thaliana* flowering pathways (Fig. 5). Significant examples include *Phvul.009G259400* and *Phvul*.009G259650 (Supplementary Note 8; SN8_ Supplemental Dataset 15; rows 90 and 92), which are orthologues of *A. thaliana LATE ELONGATED HYPOCOTYL* (*LHY*), both located within the same adaptive introgressed region of chromosome PvChr09 (characterised by an Andean allele frequency of 96% in the European genotypes). Notably, the transcription factor encoded by *LHY* is a pivotal oscillator in the morning stage of the circadian clock. It is connected to the indirect suppression of the middle, evening, and night complex genes by *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) (*45*) (Fig. 5), with a putative function in regulating flowering that has also been proposed in the common bean (*46*). In the EU_MES population, these two *LHY* orthologues show private and significant inter-chromosomal LD with *Phvul.011G050600*, which yields GWA signals for flowering time and adaptive introgression (Supplementary Note 8, SN8_ Supplemental Dataset 15; row 97). The latter is an orthologue of the *A. thaliana* genes *VERNALISATION 1* (*VRN1*) and *RELATED TO VERNALISATION1 1* (*RTV1*) (Fig. 5), which are needed to activate the floral integrator genes following long-term exposure to cold temperatures (*47*). The inter-chromosomal LD between these putative flowering genes, which is private to the EU_MES accessions, may be the result of epistatic selection. Analogous examples include *Phvul.001G204600* (Supplementary Note 8, SN8_ Supplemental Dataset 15; row 29) and *Phvul.001G204700* (Supplementary Note 8, SN8_ Supplemental Dataset 15; row 30), which are orthologous to *LUMINIDEPENDENS* (*LD*) and *NUTCRACKER* (*NUC*), respectively. Both are located in a region of PvChr01 as described above, and are in private inter-chromosomal LD with *Phvul.003G137100* (Supplementary Note 8, SN8_ Supplemental Dataset 15; row 38) on PvChr03, which is orthologous to *GATA, NITRATE-INDUCIBLE, CARBONMETABOLISM INVOLVED* (*GNC*), and *CYTOKININ-RESPONSIVE GATA FACTOR 1* (*CGA1*). *LD* is one of the eight genes identified so far in the *A. thaliana* autonomous pathway and it represses *FLOWERING LOCUS C* (*FLC*) to promote the transition from vegetative growth to flowering (Fig 5). *NUC* encodes a transcription factor that positively regulates photoperiodic flowering by modulating sugar transport and metabolism via the

FLOWERING LOCUS T (*FT*) gene (*48*; *49*). The paralogous *GNC* and *CGA1* genes act redundantly to promote greening downstream of the gibberellin signalling network (*50*).

Conclusions

We have shown that adaptive introgression was necessary for the successful dissemination and adaptation of the common bean in Europe. We combined genome resequencing, metabolomics, classical phenotyping and data analysis for chromosome-level genetic assignment and environmental association. Our data indicate that the Andean genepool was the first to be successfully introduced in Europe, most likely from Francisco Pizarro's expedition to northern Peru in 1529. Most of the Andean genetic background of the European common bean was contributed by the A1/Nueva Granada and A3/Chile races after secondary domestication, whereas the more photoperiod-sensitive A2/Peru race contributed little to European common bean germplasm. The secondary domestication of these Andean races, related to the latitudinal expansion of the cultivation areas from the Andean centres of origin, facilitated the successful dissemination of the Andean common bean in the Old World. However, we propose that the adaptive introgression observed in Europe for individuals mainly belonging to the M1/Durango-Jalisco race was an important event that underpinned the successful dissemination of this Mesoamerican ancestry in Europe. Indeed, genomic analysis indicated that Andean genotypes were rapidly disseminated, whereas Mesoamerican genotypes were eventually disseminated in Europe following introgression from the Andean types. In addition to the flowering time data gathered from the European and American accessions, we also identified clear signatures of selection in common bean orthologues of genes representing the major flowering pathway and environmental adaptations, such as the *OTU5* gene involved in the inorganic phosphate starvation response. These are interesting candidate loci for further validation. Finally, we propose that the dissemination of common bean accessions in Europe may have been influenced by political factors and constraints in the 16th century, including the interaction between political and religious powers in Western Europe and the subdivision of the European continent into Islamic and Christian countries.

Acknowledgements

This manuscript is dedicated to our former collaborator Dr. Monica Rossi, who passed away at the age of 44 in 2019, and to Dr. Chris Berry, a colleague and close friend who contributed to the editing of an earlier version of this manuscript and who recently passed away. This work was carried out as part of the BEAN ADAPT project, funded by the ERA-CAPS Programme (2014 Call, Expanding the European Research Area in Molecular Plant Sciences).

S.A. and A.R.F acknowledge funding from the European Commission for the PlantaSyst project (SGA-CSA nos. 664621/739582 under FPA no. 664620)

Author contributions

A.G., P.G., A.R.F., S.A.J. and R.P. conceived and managed the project; E.Be. and R.P. wrote the article; E.Be., A.B., C.X., E.Bi., M.R., S.A., V.D.V., K.N., G.C., G.F., P.L.M., A.G., P.G., A.R.F., S.A.J. and R.P. contributed to the writing and drafting; E.Be., A.B., C.X., E.Bi., M.R., S.A., V.D.V., T.G., K.N., G.C., G.F., L.N., J.J.F., A.C., G.A., P.L.M., G.B., A.G., P.G., A.R.F., S.A.J. and R.P. contributed to the editing of the article; E.Be., A.B., C.X., E.Bi., M.R., S.A., V.D.V., T.G., K.N., G.C., G.F., E.M., E.T., L.N., A.A., C.L., J.J.F., A.C., P.L.M., GB, A.G., A.R.F., S.A.J. and R.P. contributed to the writing and organisation of the supplementary material; E.Be., E.Bi., S.A., T.G., K.N., L.N., G.L., L.J., J.J.F. and A.C. performed DNA extraction, field and greenhouse experiments; C.X., J.H.S. and S.A.J. conducted sequencing and primary bioinformatic analysis; S.A. and A.R.F. conducted metabolomics analysis; A.B., E.Bi., M.R., G.F., E.T., C.L., G.B. and R.P. contributed to data analysis; E.Be., A.B., C.X., E.Bi., M.R., S.A., V.D.V., G.B., A.R.F., S.A.J. and R.P. contributed to coordinate data analysis and data integration. All authors read and approved the article.

Data availability

The raw sequence reads generated and analysed during the current study are available in the Sequence Read Archive (SRA) of the National Center of Biotechnology Information (NCBI) under BioProject number PRJNA573595.

References

1. Cortinovis G, Di Vittori V, Bellucci E, Bitocchi E, Papa R (2020) Adaptation to novel environments during crop diversification. Curr. Opin. Plant Biol., 56:218–222, doi: 10.1016/j.pbi.2019.12.011

2. Crosby AW (1972) The Columbian Exchange; Biological and Cultural Consequences of 1492. Westport, Conn.: Greenwood Pub.Co.

3. Brandenburg J-T, Mary-Huard T, Rigaill G, Hearne SJ, Corti H, Joets J, et al. (2017) Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLoS Genet 13(3): e1006666. https://doi.org/10.1371/journal. pgen.1006666

4. Gepts P (2002) A Comparison between Crop Domestication, Classical Plant Breeding, and Genetic Engineering. Crop Sci., 42: 1780-1790. https://doi.org/10.2135/cropsci2002.1780

5. Gepts P and Bliss FA (1988) Dissemination pathways of common bean (*Phaseolus vulgaris*, Fabaceae), deduced from phaseolin elettrophoretic variability. II Europe and Africa. Econ. Bot. 42, 86-104. https://doi.org/10.1007/BF02859038

6. Santalla M, Rodiño AP, De Ron AM (2002) Allozyme evidence supporting southwester Europe as a secondary center of genetic diversity for common bean. Theor Appl Genet 104:934–944

7. Sicard D, Nanni L, Porfiri O, Bulfon D and PapanR (2005) Genetic diversity of *Phaseolus vulgaris* L. and *P. coccineus* L. landraces in central Italy. Plant Breed., 124: 464-472. https://doi.org/10.1111/j.1439-0523.2005.01137.x

8. Zeven AC, Waninge J, van Hintum T, Singh SP (1999) Phenotypic variation in a core collection of common bean (*Phaseolus vulgaris* L.) in the Netherlands. Euphytica 109:93-106 doi: 10.1023/A:1003665408567

9. Angioi SA, Rau D, Attene G, Nanni L, Bellucci E, Logozzo G, Negri V, Spagnoletti Zeuli PL, Papa R (2010) Beans in Europe: origin and structure of the European landraces of *Phaseolus vulgaris* L. Theor Appl Genet. 121:829-843, doi: 10.1007/ s00122-010-1353-2

10. Angioi SA, Rau D, Nanni L, Bellucci E, Papa R, Attene G (2011) The genetic make-up of the European landraces of the common bean. Plant Genet Resour. 9:197-201 doi: doi:10.1017/S1479262111000190

11. Bellucci, E., Mario Aguilar, O., Alseekh, S., et al. (2021), The INCREASE project: Intelligent Collections of food-legume genetic resources for European agrofood systems. Plant J, 108: 646-660. https://doi.org/10.1111/tpj.15472

12. Bitocchi E, Rau D, Bellucci E, Rodriguez M, Murgia ML, Gioia T, Santo D, Nanni L, Attene G, Papa R (2017) Beans (*Phaseolus* ssp.) as a model for understanding crop evolution. Front Plant Sci. 8:722, doi: 10.3389/fpls.2017.00722

13. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res., 19(9):1655-64

14. Ariani A, Berny Mier y Teran JC, Gepts P (2018) Spatial and temporal scales of range expansion in wild *Phaseolus vulgaris*. Mol. Biol. Evol. 35(1): 119-131 doi: 10.1093/molbev/msx273

15. Schmutz J, McClean P, Mamidi S, et al. (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46, 707–713. https://doi.org/10.1038/ng.3008

16. Rossi M, Bitocchi E, Bellucci E, Nanni L, Rau D, Attene G, Papa R (2009) Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. Evol. Appl. 2, 504–522, doi: 0.1111/j.1752-4571.2009.00082.x

17. Tsitrone A, Kirkpatrick M, Levin DA (2003) A model for chloroplast capture. Evolution 57(8):1776-82. PMID: 14503619 doi: 10.1111/j.0014-3820.2003.tb00585.x

18. Singh SP, Gepts P, Debouck DG (1991) Races of common bean (*Phaseolus vulgaris*, Fabaceae). Econ Bot. 45,379–396, doi: 10.1007/BF02887079

19. Trucchi E, Benazzo A, Lari M, et al. (2021) Ancient genomes reveal early Andean farmers selected common beans while preserving diversity. Nat. Plants 7, 123–128, doi: 10.1038/s41477-021-00848-7

20. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. PLoS Genet. 8(1): e1002453, doi:10.1371/journal.pgen.1002453

21. Zeven AC (1997) The introduction of the common bean (*Phaseolus vulgaris* L.) into western Europe and the phenotypic variation of dry beans collected in The Netherlands in 1946. Euphytica 94, 319–328, doi: 10.1023/A:1002940220241

22. Perale M (2001) Milacis Cultus Aperire Paramus. “De milacis cultura” di Pierio Valeriano, il Primo Testo Europeo Dedicato al Fagiolo. Belluno: Momenti AiCS

23. Myers JR, Formiga AK and Janick J (2022) Iconography of beans and related legumes Following the Columbian Exchange. Front. Plant Sci. 13:851029. doi: 10.3389/fpls.2022.851029

24. Bellucci E, Bitocchi E, Rau D, Rodriguez M, Biagetti E, Giardini A, et al. (2014) Genomics of origin, domestication and evolution of *Phaseolus vulgaris*. Genomics of Plant Genetic Resources, eds R. Tuberosa, A. Graner, and E. Frison (Berlin: Springer), 483–507. doi: 10.1007/978-94-007-7572-5_20.

25. Wu J, Wang L, Fu J, et al. (2020) Resequencing of 683 common bean genotypes identifies yield component trait associations across a north–south cline. Nat Genet 52, 118–125. https://doi.org/10.1038/s41588-019-0546-0

26. Ongaro L., Scliar M.O, Flores R., et al (2019). The Genomic Impact of European Colonization of the Americas 3974-3986.e4 Curr. Biol., 29 (2019), 10.1016/j.cub.2019.09.076

27. Sherpa, S, Blum, MGB, Capblancq, T, Cumer, T, Rioux, D, Després, L. Unravelling the invasion history of the Asian tiger mosquito in Europe. Mol Ecol. 2019; 28: 2360– 2377. https://doi.org/10.1111/mec.15071

28. S.J. Micheletti, K. Bryc, S.G. Ancona Esselmann, W.A. Freyman, M.E. Moreno, G.D. Poznik, A.J. Shastri, S. Beleza, J.L. Mountain, 23andMe Research Team Genetic Consequences of the Transatlantic Slave Trade in the Americas. Am. J. Hum. Genet., 107 (2020), pp. 265-277

29. Rispe, C., Legeai, F., Nabity, P.D. et al. The genome sequence of the grape phylloxera provides insights into the evolution, adaptation, and invasion routes of an iconic pest. BMC Biol 18, 90 (2020). https://doi.org/10.1186/s12915-020-00820-5

30. Fontaine, M.C.; Labbé, F.; Dussert, Y.; Delière, L.; Richart-Cervera, S.; Giraud, T.; Delmotte, F. Europe as a Bridgehead in the Worldwide Invasion History of Grapevine Downy Mildew, Plasmopara viticola. Curr. Biol. 2021, 31, 2155–2166.

31. 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. PNAS, USA 109: E788–E796

32. Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E., et al. (2013) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytol. 197, 300–313, doi: 10.1111/j.1469-8137.2012.04377.x

33. Whitlock, M.C. (2008), Evolutionary inference from QST. Molecular Ecology, 17: 1885-1896. https://doi.org/10.1111/j.1365-294X.2008.03712.x

34. Beleggia R, Rau D, Laidò G, Platani C, Nigro F, Fragasso M, De Vita P, Scossa F, Fernie AR, Nikoloski Z, Papa R. Evolutionary Metabolomics Reveals Domestication-Associated Changes in Tetraploid Wheat Kernels. Mol Biol Evol. 2016 Jul;33(7):1740-53. doi: 10.1093/molbev/msw050. Epub 2016 Apr 5. PMID: 27189559; PMCID: PMC4915355.

35. Meyer RS and Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet. 14(12):840-52. doi: 10.1038/nrg3605. PMID: 24240513

36. Chakraborty R and Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. PNAS, USA 85 (23) 9119-9123; doi: 10.1073/pnas.85.23.9119

37. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics, 193(3):929-41

38. Yen MR, Suen DF, Hsu FM, Tsai YH, Fu H, Schmidt W, Chen PY (2017) Deubiquitinating enzyme OTU5 contributes to DNA methylation patterns and is critical for phosphate nutrition signals. Plant Physiol. 175(4):1826-1838, doi: 10.1104/pp.17.01188

39. Suen DF, Tsai YH, Cheng YT, Radjacommare R, Ahirwar RN, Fu H, Schmidt W (2018) The deubiquitinase OTU5 regulates root responses to phosphate starvation. Plant Physiol. 176(3):2441-2455, doi: 10.1104/pp.17.01525

40. Koinange, E.M.K., Singh, S.P. and Gepts, P. (1996), Genetic Control of the Domestication Syndrome in Common Bean. Crop Science, 36: 1037-1045 cropsci1996.0011183X003600040037x. https://doi.org/10.2135/cropsci1996.0011183X003600040037x

41. Repinski SL, Kwak M, Gepts P. The common bean growth habit gene PvTFL1y is a functional homolog of Arabidopsis TFL1. Theor Appl Genet. 2012 May;124(8):1539-47. doi: 10.1007/s00122-012-1808-8. Epub 2012 Feb 14. PMID: 22331140.

42. Wallace DH, Yourstone KS, Masaya PN, Zobel RW. 1993. Photoperiod gene control over partitioning between reproductive and vegetative growth. Theoretical and Applied Genetics 86, 6–16.

43. Kamfwa, K., Cichy, K.A. and Kelly, J.D. (2015), Genome-Wide Association Study of Agronomic Traits in Common Bean. The Plant Genome, 8: plantgenome2014.09.0059. https://doi.org/10.3835/plantgenome2014.09.0059

44. James L Weller, Jacqueline K Vander Schoor, Emilie C Perez-Wright, Valérie Hecht, Ana M González, Carmen Capel, Fernando J Yuste-Lisbona, Rafael Lozano, Marta Santalla, Parallel origins of photoperiod adaptation following dual domestications of common bean, Journal of Experimental Botany, Volume 70, Issue 4, 1 February 2019, Pages 1209–1219, https://doi.org/10.1093/jxb/ery455

45. Adams S, Manfield I, Stockley P, Carré IA (2015) Revised morning loops of the *Arabidopsis* circadian clock based on analyses of direct regulatory interactions. Plos One 10(12): e0143943, doi: 10.1371/journal.pone.0143943

46. Kaldis, AD., Kousidis, P., Kesanopoulos, K. et al. Light and circadian regulation in the expression of LHY and Lhcb genes in Phaseolus vulgaris. Plant Mol Biol 52, 981–997 (2003). https://doi.org/10.1023/A:1025433529082.

47. Heo JB, Sung S, Assmann SM (2012) Ca2+-dependent GTPase, extra-large G protein 2 (XLG2), promotes activation of DNA-binding protein related to vernalization 1 (RTV1), leading to activation of floral integrator genes and early flowering in *Arabidopsis*. J. Biol. Chem. 287(11):8242-8253, doi: 10.1074/jbc.M111.317412

48. King RW, Hisamatsu T, Goldschmidt EE, Blundell C (2008) The nature of floral signals in *Arabidopsis*. I. Photosynthesis and a far-red photoresponse independently regulate flowering by

increasing expression of FLOWERING LOCUS T (FT). J. Exp. Bot. 59(14):3811–3820, doi: 10.1093/jxb/ern231

49. Seo PJ, Ryu J, Kang SK, Park CM (2011) Modulation of sugar metabolism by an INDETERMINATE DOMAIN transcription factor contributes to photoperiodic flowering in *Arabidopsis*. Plant J. 65(3):418-429, doi: 10.1111/j.1365-313X.2010.04432.x

50. Richter R, Bastakis E, Schwechheimer C (2013) Cross-repressive interactions between SOC1 and the GATAs GNC and GNL/CGA1 in the control of greening, cold tolerance, and flowering time in *Arabidopsis*. Plant Physiol. 162(4):1992-2004; doi: 10.1104/pp.113.219238

Figure legends

Figure 1. Population structure of common bean in America and Europe.

a, Admixture analysis (K = 2) showing inferred ancestry in the American (AM; left) and European (EU; right) accessions, with the identification of two gene pools (identified as clusters 1 and 2) that show correspondence to the two main common bean gene pools based on our passport data (cluster 1 – Andean, cluster 2 – Mesoamerican), and several intermediates and admixed genotypes in Europe. **b,** Admixture plots for the AM Mesoamerican accessions (K = 2) grouped by geographic origin (i.e., latitude and state), which identifies two main subgroups (M1 and M2). **c,** Admixture plots for the AM Andean accessions (K = 4) grouped by geographic origin (i.e., latitude and state), which identifies three Andean genetic subgroups (A1, A2 and A3). A fourth cluster in four accessions, based on the whole-set ADMIXTURE analysis (K = 2), was induced by the occurrence of Mesoamerican alleles with AM_M1/AM_M2 components (supplementary Note 4). **d,** Neighbour-joining tree and seed pictures of the 66 pure American accessions. **e,** Spatial interpolation of the geographic distributions of the EU Mesoamerican (M1 and M2) and EU Andean (A1, A2 and A3) ancestry components in Europe, as inferred by ChromoPainter analysis. **f,** Primary and secondary domestications of Mesoamerican and Andean genetic groups/races in America. Loss of photoperiod sensitivity during secondary domestication was a relevant factor for the introduction of the Andean A1/Nueva Granada and A3/Chile and for the Mesoamerican M2/ Mesoamerica ancestries in Europe (solid arrow). Genetic group M1 (Durango-Jalisco race) was successfully introduced into Europe after introgression from other genetic groups characterised by little or no photoperiod sensitivity (dashed arrow). Genetic group A2 (Peru race) was not introduced into Europe due to its high photoperiod sensitivity (discontinuous and truncated line).

Figure 2. Genetic structure, molecular phenotyping and flowering data.

a, b, Molecular phenotypes (PCA1 from 1493 putative secondary metabolites, $H^2 > 0.65$ over the entire dataset) of (a) 94 American accessions and (b) 96 European accessions confirm the subdivision into the two main groups based on the admixture coefficient (derived from nuclear genomic data, K = 2). Intermediate phenotypes and genotypes are seen in Europe. **c,** Violin plots showing the distribution of

PCA1 values related to secondary metabolites showing high heritability (H^2 > 0.65) by genetic subgroups in the American and European accessions. PCA1 was used as a representative molecular phenotype, and it explains 25.7% of the total variance for these traits. **d,** Violin plots showing the distribution of PCA1 values related to the days to flowering (DTF) and photoperiod sensitivity (PS) by genetic subgroups in the American and European accessions. PCA1 was used as a representative phenotypic trait for DTF and PS, and it explains 68.8% of the total variance for these traits. **e,** Proportions of the genetic memberships – P(AM_A1), P(AM_A2), P(AM_A3), P(AM_M1), P(AM_M2), P(SAND), and P(SMES) – inferred from the donor accessions and composing the American and European accessions (grouped as mainly AM_A1, AM_A2, AM_A3_AM_M1, AM_M2, EU_A1, EU_A3, EU_M1, EU_M2, and EU_MIX) are shown in the pie charts below the corresponding groups and flowering data (number and percentage of individuals with delayed or no flowering) in northern and southern Europe, related to the corresponding groups.

Figure 3. Mapping introgression in the European common bean using ChromoPainter.

a, Proportion of introgressed genome in the Mesoamerican (EU_MES; n = 43) and Andean (EU_AND; n = 71) groups. **b, c,** Boxplots showing the median length of the introgressed blocks identified in each of the EU_AND and EU_MES accessions across all of the chromosomes (b) and the median length of the introgressed blocks identified in each of the EU_AND and EU_MES individuals by chromosome (c).

Figure 4. Boxplots of θπ averaged over 100-kb non-overlapping sliding windows, linkage disequilibrium (LD) decay and inter-chromosomal linkage disequilibrium and genetic load.

a, Genetic diversity computed using whole chromosomes and the unmasked dataset. **b,** Genetic diversity computed after the admixture masking process using whole chromosomes and LD decay according to the physical distance. **c,** Comparative LD decay in the American and European accessions. **d,** Genome-wide measure of genetic load in the American and European accessions. The ratios are shown for missense (up) and loss-of-function (down) over synonymous mutations in the different groups. AM_M* and AM_A* are the admixed American accessions (not pure American individuals). **e,** Private inter-chromosomal LD in American and European accessions (left), in the Mesoamerican and

Andean European accessions (middle), and considering genomic regions under selection (S) in the Mesoamerican and Andean European accessions (right).

Figure 5. Candidate genes for adaptation.

Schematic representation of the regulatory networks underlying the four major flowering pathways in *Arabidopsis thaliana*. The genes involved in the photoperiod, vernalisation, autonomous and gibberellin pathways that lead to the transition from vegetative to flowering are shown below the corresponding pathway. Additional genes belonging to secondary pathways and that interact with the main regulatory flowering networks are shown in italic. *Phaseolus vulgaris* orthologues were identified using OrthoFinder. Genes with signatures of selection and adaptive introgression, and those located in GWA peaks for days to flowering and growth habit in common bean, are highlighted as follows: yellow hexagons – common bean orthologues of *LHY* (Phvul.009G259400, Phvul.009G259650) and VRN1 and RTV1 (Phvul.011G050600) showing private inter-chromosomal linkage disequilibrium (LD) in the EU_M pool (*Supplementary Note 7*); pink hexagons – common bean orthologues of *LD* (Phvul.001G204600), *NUC* (Phvul.001G204700), *CGA1* and *GNC* (Phvul.003G137100) showing private inter-chromosomal LD in the EU_M pool (*Supplementary Note 7*); red outlines – at least one orthologous gene in common bean showing signature of selection, introgression and with a significant differentiation (F_{ST} index) between American and European accessions ($p < 0.05$); orange outlines – at least one orthologous gene in common bean showing a signature of selection with no significant F_{ST} ($p < 0.05$); blue asterisks – at least one orthologous gene in common bean showing a signature of introgression; dashed blue outlines – at least one orthologous gene in common bean located within 50 kb centered on a significant GWA peak for days to flowering (DTF); dashed green outlines – at least one orthologous gene in common bean located within 50 kb centered on a significant GWA peak for growth habit (GH); arrows – positive regulation of gene expression; truncated arrows – repression of gene expression; solid lines – direct interactions; dashed lines – indirect interactions in *A. thaliana*. The candidate genes for adaptation or post-domestication of the common bean in European environments, orthologous to those involved in flowering-related pathways, are shown in parentheses: *UBP12/13* (Phvul.007G234000); *LHY* (Phvul.009G259400, Phvul.009G259650); *LUX* (Phvul.011G062100);

PIL5 (Phvul.001G168700); *CIB2* (Phvul.008G133600); *LRB1* (Phvul.006G109600); *DRIP1/2* (Phvul.001G157400, Phvul.007G177500); *VRN1*, *RTV1* (Phvul.011G050600); *UBC1/2* (Phvul.003G191900); *LD* (Phvul.001G204600); *TFL1*, *ATC* (Phvul.001G189200); *GA2OX4* (Phvul.006G120700); *CGA1*, *GNC* (Phvul.003G137100); *GAI*, *RGA1*, *RGL1*, *RGL2* (Phvul.001G230500); *LMI1* (Phvul.001G184800, Phvul.001G184900); *SIC* (Phvul.008G182500); *CRP* (Phvul.008G142400); *MYB30* (Phvul.008G041500); *NUC* (Phvul.001G154800, Phvul.001G204700, Phvul.011G074100); *SUC9* (Phvul.004G085100, Phvul.004G085400, Phvul.004G085594); *CYP715A1* (Phvul.007G071500). Refer to SN8_ Supplemental Dataset 15 for detailed information on each candidate gene (i.e., selection, top selection, F_{ST}, introgression and GWA data).

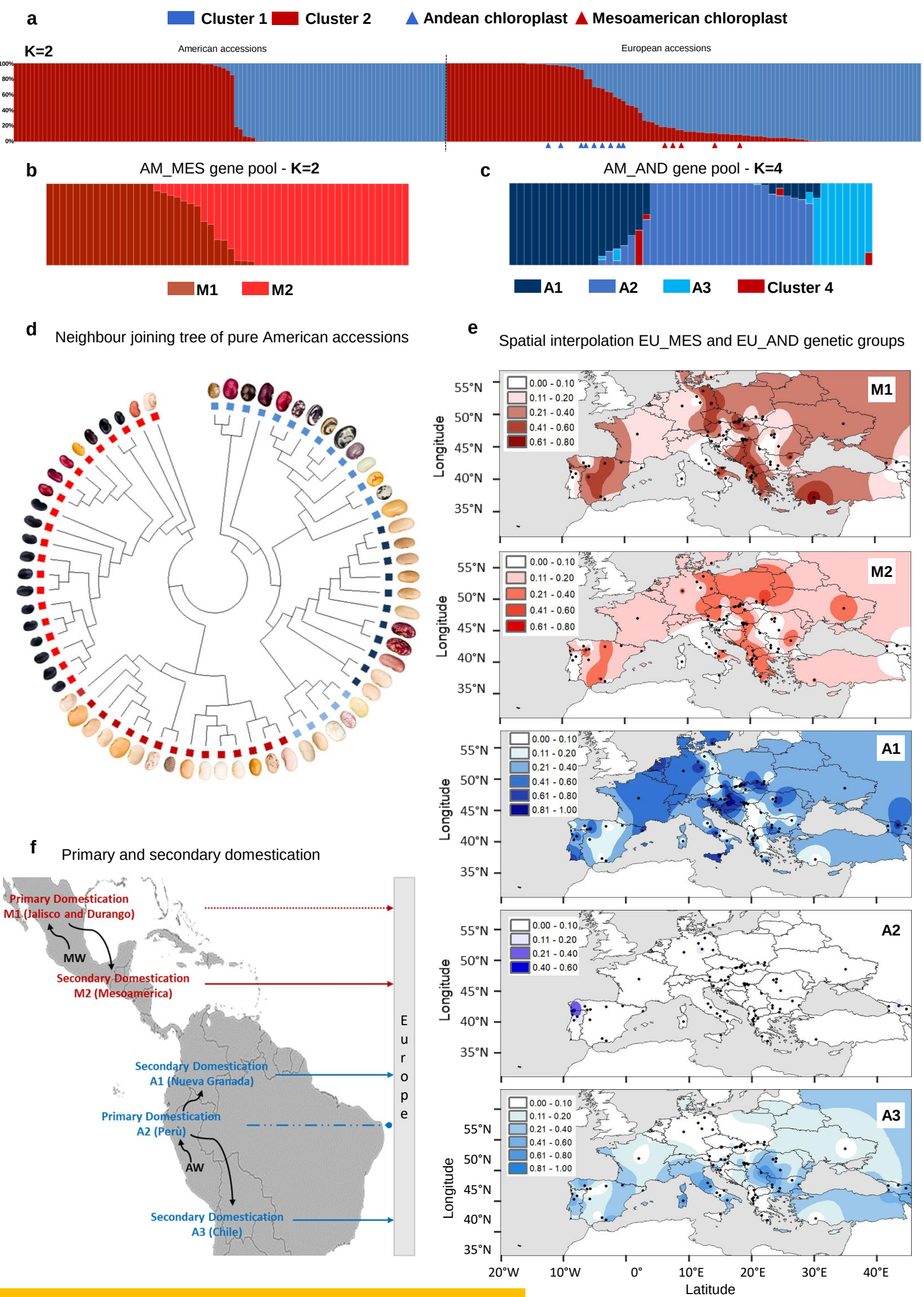

Figure 1. Population structure of common bean in America and Europe.

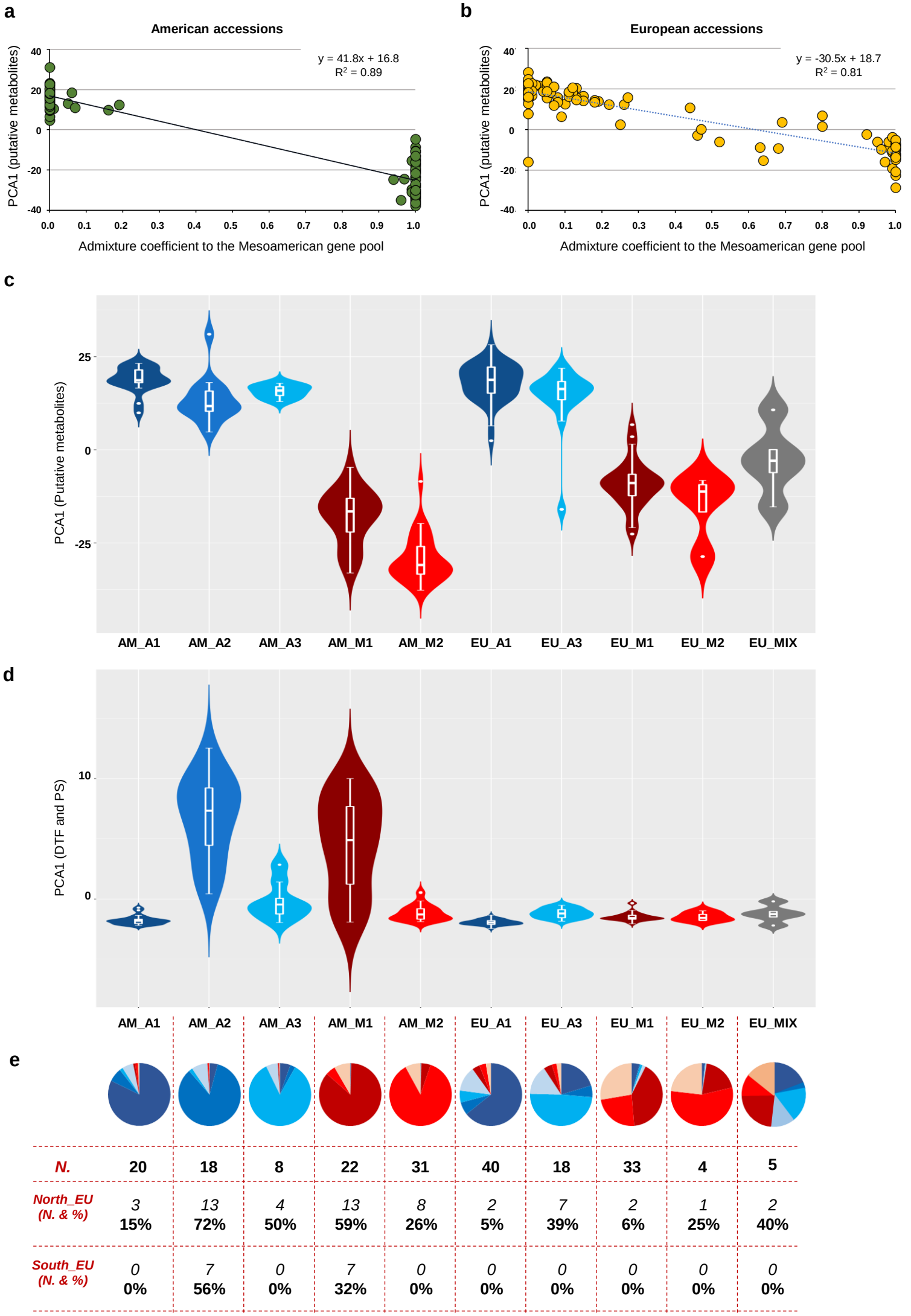

Figure 2. Genetic structure, molecular phenotyping and flowering data.

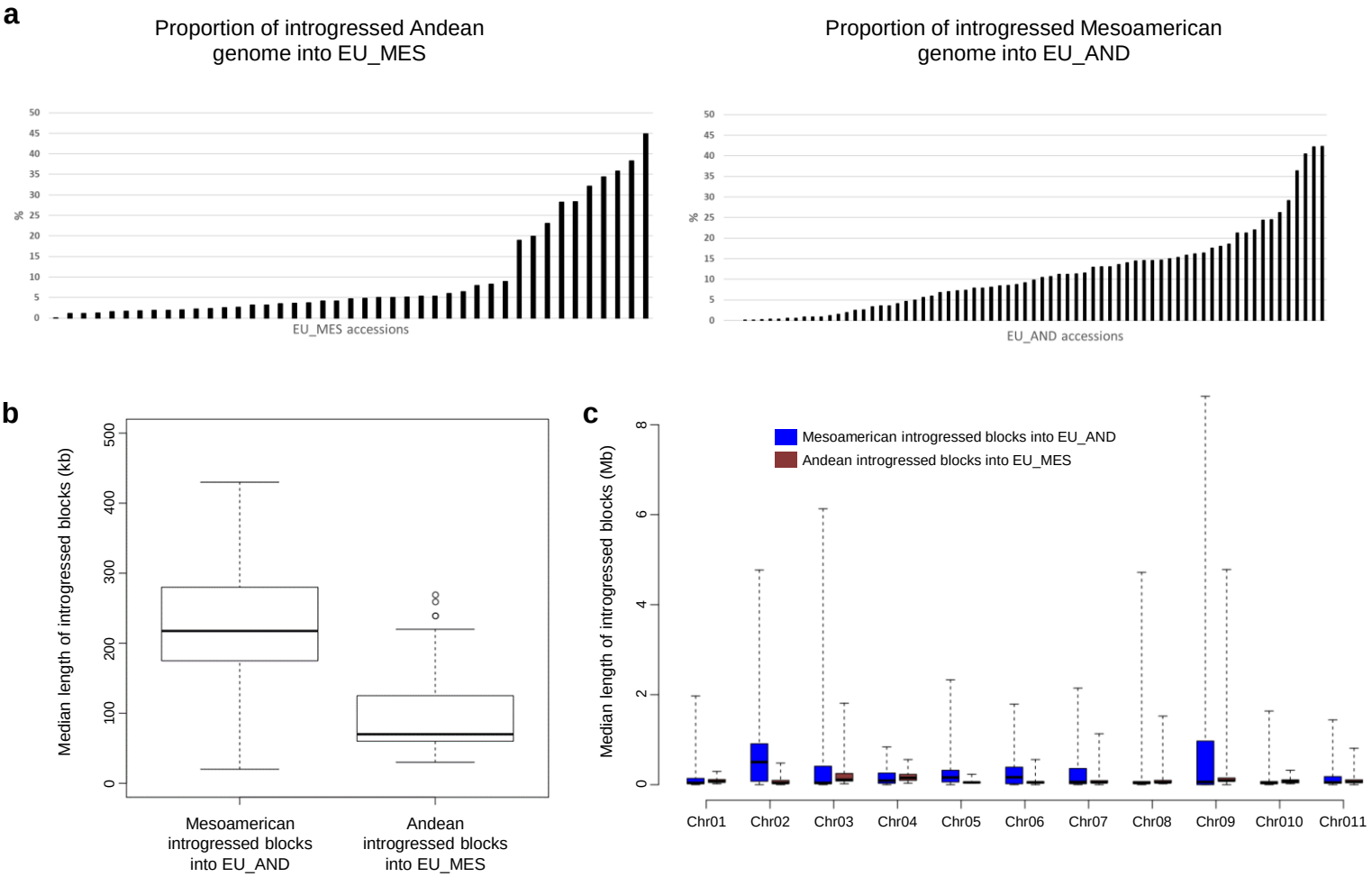

Figure 3. Mapping the introgression in the European common bean using Chromopainter.

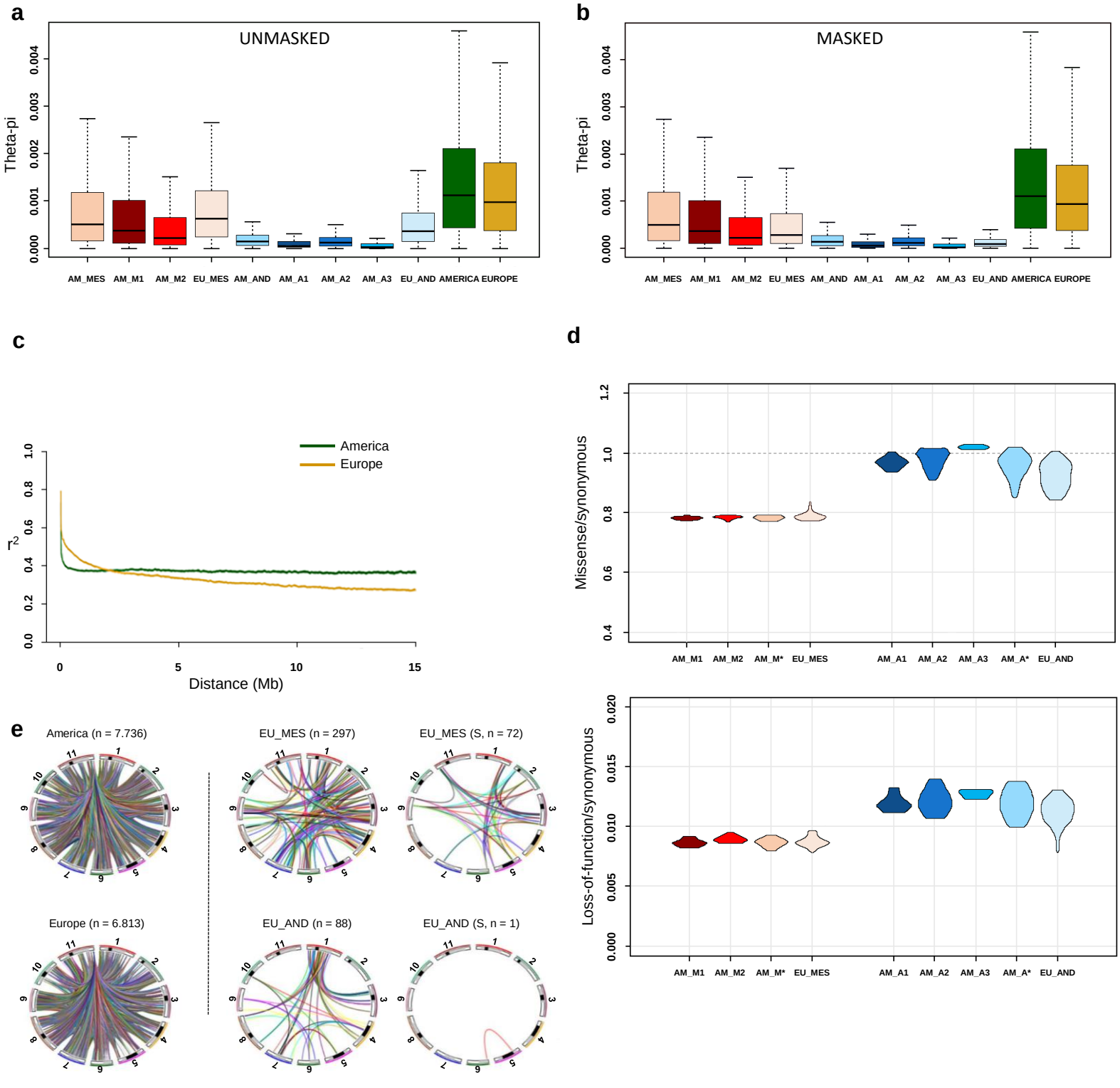

Figure 4. Boxplots of θπ averaged over 100-kb non-overlapping sliding windows, linkage disequilibrium decay and inter-chromosomal linkage disequilibrium and genetic load.

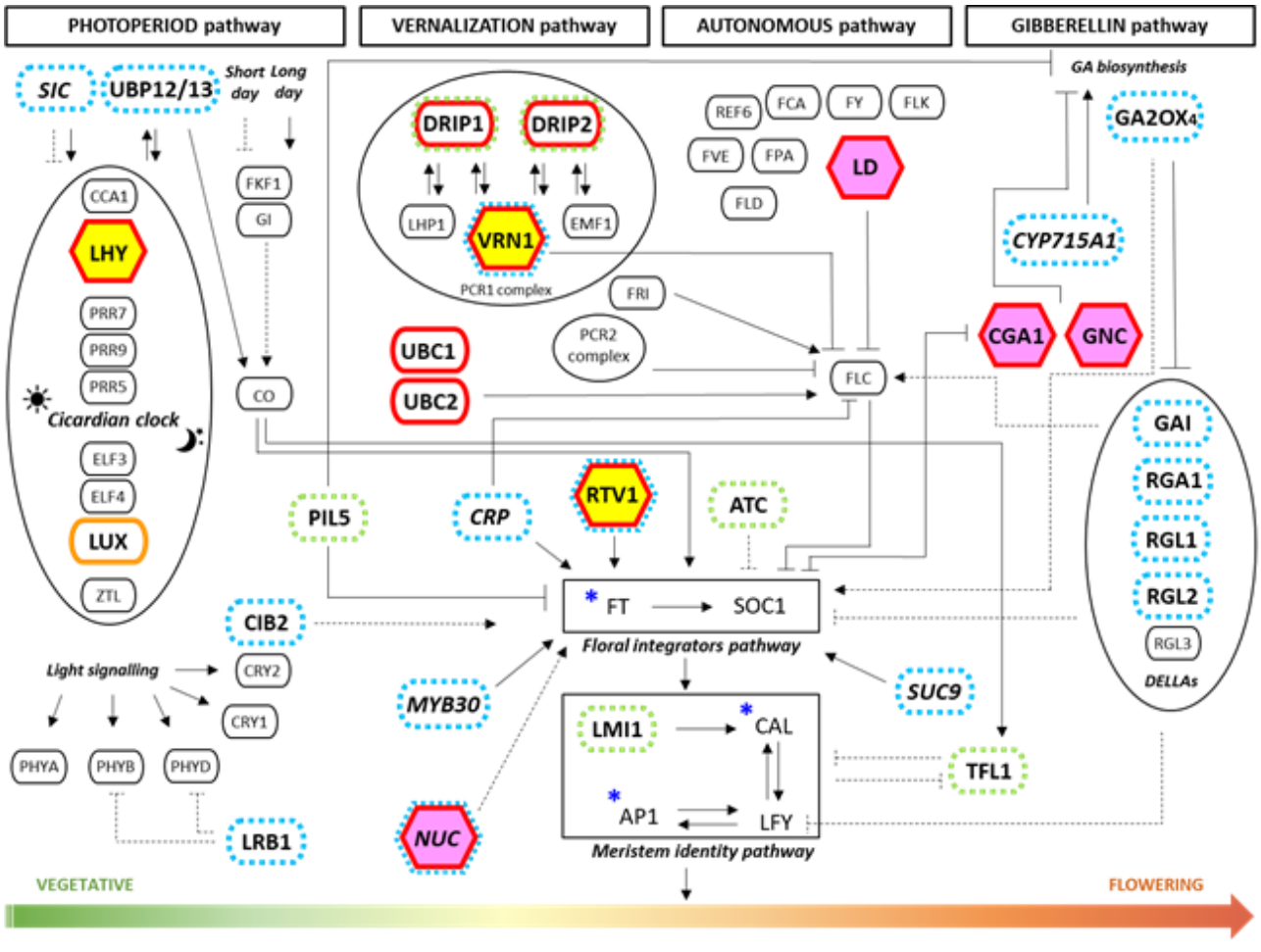

Figure 5. Candidate genes for adaptation.

Methods

Plant materials

Original seeds for 218 common bean accessions (*P. vulgaris*) were collected from international gene banks or individual institutional collections. We produced 199 single seed descent (SSD) lines by performing at least three cycles of self-fertilisation. For the remaining 19 accessions, one seed per accession was sampled directly from original seeds provided by the donor.

Experimental design and phenotyping

Plants were grown across 10 different environments in fields and greenhouses, applying long-day (7), short-day (2) and intermediate photoperiod conditions. During the summers of 2016 and 2017, four field trials were carried out in Italy (Villa d'Agri, Marsicovetere, Potenza) and in Germany (Gatersleben IPK) (Supplementary Note 2.1). Six additional greenhouse experiments were performed under controlled conditions in Golm (Potsdam, Germany), Potenza (Italy) and Villaviciosa (Spain) during 2016, 2017 and 2018 (Supplementary Note 2.2).

Classical phenotyping was carried out on the 199 SSD lines, focusing on two main traits: days to flowering (DTF), defined as the number of days from sowing until 50% of plants showed at least one open flower; and growth habit (GH), defined as determinacy *vs* indeterminacy on a single plant basis. Photoperiod sensitivity (PS) was calculated as the ratio between DTF in long-day and short-day experiments.

Descriptive statistics were calculated for the different phenotypic traits using R (https://cran.r-project.org/) or JMP 7.0.0 (*51*). The restricted maximum likelihood (REML) model implemented in JMP 7.00 was used to calculate least square means (LSM) and the best linear unbiased predictors (BLUPs) of each genotype. The REML model was also used to calculate the broad-sense heritability (h^2B) for each quantitative trait by considering genotypes and environments as random effects. The distribution of DTF in each environment and Pearson's pairwise correlation between environments were calculated using the *corrplot* and *PerformanceAnalytics* R packages (*52*, *53*).

Molecular phenotyping of the 199 accessions was performed on the first trifoliate fully-expanded leaves harvested under long-day conditions in three biological replicates. Secondary metabolites were measured as previously described (*54*). For non-targeted metabolomics, chromatogram peak detection and integration were achieved using GeneData REFINER MS 10.0 (http://www.genedata.com). To explore the

molecular phenotypic diversity, we performed non-targeted metabolic fingerprinting by high-throughput LC-MS analysis. Mass signals that were not detected in ≥ 50% of the samples and/or those with a peak intensity ≤ 1000 were excluded. Heritability was analysed as stated above, setting genotype and continent as random effects. The heritability was calculated based on 190 accessions (94 from the Americas and 96 from Europe) with more than one replicate.

Sequencing, variant calling and annotation

Genomic DNA was extracted from frozen young leaves of the 199 SSD lines grown in greenhouse and directly from seeds of the remaining 19 accessions using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). It was sheared using a Covaris E220 device to fragments of ~550 bp and PCR-free libraries were constructed using the KAPA HyperPrep Kit. Paired-end sequencing libraries were sequenced on Illumina HiSeq2500 or HiSeq4000 devices and labelled with different barcodes.

Sequencing data were aligned to the common bean reference genome v2.0 (*15*) using BWA-mem v0.7.15 (*55*). Unmapped reads were mapped to the *P. vulgaris* chloroplast genome (NCBI NC_009259). In both cases, SNPs were called using SAMtools (*56*) and Genome Analysis ToolKit (GATK) v3.6 (*57*; *58*). In SAMtools, duplicated reads were removed with *rmdup* and SNPs were discovered with *mpileup* for filtered high-quality alignments (–q 10) and bases (–Q 20) and then genotyped with BCFtools (*59*). In GATK, duplicated reads were sorted and filtered with Picard v2.4.1 (http://broadinstitute.github.io/picard). Variants were then called according to GATK best practices and pre-filtered using the recommended parameters for hard filtering. Chromosomal and overlapping SNPs reported by both methods were retained and the genotypes produced by GATK were selected. We applied additional filtering with VCFtools (*607*) (--minDP 3 --max-missing 0.5 --maf 0.05) and excluded SNPs with proportions of heterozygous genotypes > 0.01. SNPs were annotated with SnpEff v4.3s (*61*).

Population structure analysis

Population structure analysis was applied to the SAMtools/GATK overlapping SNP callset, followed by filtering to retain only genomic positions with a QUAL score ≥ 30 and a global depth of coverage between 1/3 and 4 times the mean value. Individual genotypes called using two reads or fewer were marked as missing data. Imputation and phasing were performed with Beagle (*62*). ADMIXTURE v1.3 (*13*) was used for population structure analysis. The unphased variants were filtered by taking one SNP every 250 kb using

VCFtools. In ADMIXTURE, we varied K from 1 to 20 in 20 replicates and applied the analysis independently over the whole sample of American and European (n = 218) accessions or using the American accessions only (AM, n = 104). We dealt with potential cryptic population structures within each pool as previously described (*16*, *31*, *32*).

Population structure was inferred from chloroplast data using Bayesian Analysis of Population Structure (BAPS) v5.3 (*63*, *64*). Mixture analysis was used to determine the most probable K value according to the data. *Clustering with linked loci* analysis was chosen to account for the linkage between sites. Ten repetitions of the algorithm for each K value from 2 to 20 were applied. The relationships between genotypes were determined using the neighbour-joining method in MEGA X (*65*) with a bootstrap value of 1000. Gaps and missing data were excluded.

Chromosome painting

Chromosome painting was applied to the phased variants in ChromoPainter v2.0 (*20*). The effective population size (Ne) and mutation rates (Mu) were estimated individually for each accession using 10 iterations of the expectation–maximisation algorithm in ChromoPainter. The estimated parameters were fixed in a new round of analysis producing the final chromosome painting of the recipient haplotypes. Donor individuals were chosen as follows, according to their ancestry proportion inferred by admixture: (a) Mesoamerican individuals with a q value > 0.99 in the admixture run (K = 3 using all American accessions), and (b) Andean individuals with a q value consistently > 0.99 from K = 2 to K = 4 in the admixture run restricted to Andean accessions. Donors were subdivided into the five groups inferred by ADMIXTURE (AM_M1, AM_M2, AM_A1, AM_A2 and AM_A3) and were used to estimate their contribution to the ancestry of each SNP in the recipient individuals. Individual SNP probabilities were then combined in 10-kb non-overlapping sliding windows along chromosomes and each window in each recipient haplotype was assigned to one of the five donor groups if a probability ≥ 0.8 was observed (Supplementary Note 4.2). The total proportion of genetic material from the seven groups or “unknown” (genotypes assigned to none of the groups) was computed for each recipient individual and for each chromosome (both pairs). The final assignment of each recipient accession to the genepools was based on (a) the total proportion of windows attributed to Mesoamerica or Andes, and (b) the number of chromosomes assigned to the two genepools following the majority rule criterion.

The attribution of each genomic window to the seven groups was also used to estimate the length of the introgressed blocks within each European accession. Each haplotype of the EU_AND accessions was traversed, merging consecutive windows attributed to any of the Mesoamerican clusters. Bedtools (*66*) was used to join windows within a maximum distance between elements of 50 kb to deal with artificially broken introgressed blocks. The length of each Mesoamerican block in each EU_AND individual was recorded for each chromosome and was then filtered to remove blocks composed of single windows (10 kb). The final within-individual distribution of lengths was characterised by the median value due to the non-normal distribution of the data.

For spatial analyses, the ecological data (resolution ~1 km^2) were downloaded from WorldClim (http://www.worldclim.org) (*67*) for a total of 19 bioclimatic variables and 24 monthly variables (SN4_ Supplemental Dataset 7). The vegan *R* package (*68*) was used to calculate the geographical and ecological distances, the Mantel statistics, and the spatial autocorrelation. Initially, the Mantel statistics were tested by 10^3 permutations and the autocorrelogram was calculated between 10 distance classes of nearly 540 km each, determining the significance of the correlation in each class by 9999 permutations. We then applied environmental association analysis with a multivariate correlation analysis between the p values (proportion of membership to the five genetic groups) assigned to each European accession and the ecological variables registered at the collection site.

Genetic diversity

The genetic diversity within groups of accessions, defined according to their geographic origin and genepool, was quantified using the theta estimator θπ (*69*). The --site-pi VCFtools flag was used to obtain a per-SNP estimate that was subsequently filtered according to the genome annotation, including only positions located: (a) in callable regions, (b) in coding regions, and/or (c) in neutral regions (Supplementary Note 4.3). The per-site θπ estimate was then summed up and divided by the size of each specific region to calculate a global estimate. A raw estimate of θπ along chromosomes, averaged over 100-kb non-overlapping windows, was also computed to highlight chromosomal regions with different levels of genetic diversity. To evaluate the stability of the θπ estimate at different missing data levels, a masked dataset was obtained by filtering introgressed alleles (identified by ChromoPainter) and alleles with an ambiguous assignment, within European accessions. The --site-pi and --missing-site commands in VCFtools were used to obtain a per-site

θπ estimate and the proportion of missing data for each position, respectively. The global within-group θπ was computed for the callable, coding and neutral genomic partitions, excluding regions with an average (over SNP) minimum mean proportion of non-masked individuals (PIND) from 0% to 100%.

To detect patterns of private alleles, missense and synonymous variants were screened in American and European accessions (Supplementary Note 4.3). Variants that were private to the European or American groups were retained and divided into those with low (< 5%) and medium-high (> 5%) within-sample frequencies. The genomic coordinates related to private alleles segregating at different frequencies in the American and European groups of accessions were intersected with the gene annotations, and the burden of missense and synonymous mutations was recorded for each gene element.

The magnitude of the genomic differentiation between and within America and Europe was evaluated using the Weir & Cockerham estimator F_{ST} (*70*). We estimated the baseline differentiation between and within the two continents. In addition, the F_{ST} was then computed in 10-kb non-overlapping sliding windows between each pair of groups using VCFtools. The mean and the interquartile range (IQR) of the windows-based distribution were used as a point estimate of the differentiation between groups and to evaluate its dispersion.

Comparison of the genetic structure, molecular phenotype and flowering data

Analysis of variance (ANOVA) was used to evaluate differences between the genetic subgroups using the first principal component related to DTF and PS (Supplementary Note 2.4) as representative phenotypic traits.. Principal component 1 was obtained from the secondary metabolites with a high hereditability of $H^2 > 0.65$ (Supplementary Note 2.3) and this was used as a phenotype for comparison between the genetic subgroups.

Tagging the signatures of adaptation in Europe

The presence of *excess of introgression and selection* was investigated in Europe. To detect deviations from the frequencies expected in the absence of demographic and selection forces, the ChromoPainter output was parsed by tracing the assignment of each SNP to the corresponding Mesoamerican or Andean groups. For each SNP, we computed the proportion of haplotypes assigned to the Mesoamerican or Andean groups. We extracted the genomic coordinates of SNPs showing an unexpected proportion of Andean alleles (threshold: EU_A, 71 × 2 = 142 haplotypes, plus 50% of the Mesoamerican ones, EU_M, 43 × 2 × 0.5 = 43, Fobs ≥

0.811). The putative SNP targets of Mesoamerican introgression events were identified according to the same rationale (threshold: EU_M, 43 × 2 = 86, plus the 50% of Andean ones, EU_A, 71 × 2 × 0.5 = 71, Fobs = 0.688). The Bedtools slop -b 2500 and merge -d 10000 functions were used to pass from SNP point coordinates to 5-kb regions and then merge into larger genomic blocks if the relative distance between them was < 10 kb. Only genomic regions supported by at least three SNPs were retained (Supplementary Note 6.1)

The hapFLK (*37*) method was used to identify selection signatures. The local genomic differentiation along chromosomes, as measured by haplotypic F_{ST}, was compared to the expectation given by the inferred genomic relationships between groups, considering the genetic drift within groups. Accessions were subdivided into the AM_A (n = 30), AM_M (n = 36), EU_A (n = 71) and EU_M (n = 43) groups and VCFtools was used to sample a single SNP every 250 kb. This set of SNPs was used to construct a neighbour-joining tree and a kinship matrix according to the Reynolds' genetic distance matrix between the four groups of accessions, constituting a genome-wide estimate of population structure. The hapFLK statistics were then computed independently on each chromosome over the complete SNP dataset and averaged over 20 expectation–maximisation cycles to fit the LD model. Initial analysis fixed the number of haplotype clusters to five. A second run was conducted, selecting the appropriate number of haplotype clusters based on the fastPHASE (*71*) cross-validation procedure, implemented in the *imputeqc* R package (https://github.com/inzilico/imputeqc). VCFtools was used to extract a subset of SNPs spaced at least 100 kb apart on each chromosome, and five independent copies of the SNP set were generated, randomly masking 10% of the variants. We then used fastPHASE v1.4.8 to impute the missing genotypes in each dataset, setting the number of haplotype clusters to 5, 10, 20, 30, 40 and 50. The EstimateQuality function was used to compute the proportion of wrongly imputed genotypes (Wp) for each combination, and the K value, minimising the mean Wp proportion across the five SNP set replicates, was selected as the most supported number of haplotype clusters. The analysis was replicated using all or only American accessions. The scaling_chi2_hapflk.py script was used to scale hapFLK values and compute the corresponding p values. The significant SNPs ($p < 10^{-3}$; FDR < 0.05) were extracted and bedtools was used to create a 10-kb region centred on each significant SNP and to merge overlapping regions within a maximum distance of 5 kb. The two sets of regions were merged forming the extended set (constituted by the union of the two sets) and the restricted set (containing only regions supported by both runs). To pinpoint putative regions under selection

in Europe, the extended and restricted sets were intersected with the F_{ST} window analysis, and only regions containing at least one F_{ST} window located in the top 5% or top 1% were retained.

Linkage disequilibrium (LD)

The relationship between LD and physical distance along chromosomes was evaluated in America and Europe, and successively within the American subgroups. PopLDdecay (*72*) was used to compute the correlation (r^2) between allele frequencies at pairs of SNPs along the chromosomes, setting a minimum minor allele frequency of 0.1 and a maximum distance between SNPs of 5 Mb.

The level of inter-chromosomal LD was also evaluated. VCFtools was used to sample one SNP every 10 kb and compute r^2 between pairs of markers located on different chromosomes. The analysis was performed independently over the American subgroups, using only SNPs that were segregating within each group of accessions with a minor allele frequency > 0.05, and only pairs of SNPs showing an r^2 value ≥ 0.8 were retained. Multiple pairs of SNPs pointing to the same chromosomal regions were merged if within a distance of 100 kb from each other and only pairs of regions spanning at least 500 kb on each side were retained. The whole analysis was also repeated including only SNPs falling in the putative regions under selection, decreasing the minimum width of retained regions from 500 kb to 50 kb. Link plots showing high-LD regions were produced using Rcircos (*73*).

Genome-wide association study (GWAS)

GWAS was carried out for the growth habit, flowering time and photoperiod sensitivity data (Supplementary Note 2.4). First we ran a single-locus mixed linear model (MLM) in the R package MVP (*74*) (https://github.com/XiaoleiLiuBio/MVP). Bonferroni correction at $\alpha = 0.01$ was applied as the significance threshold for each trait. The analysis was then conducted using the multi-locus stepwise linear mixed-model (MLMM) [*75*] (https://github.com/Gregor-Mendel-Institute/MultLocMixMod). By applying a stepwise approach, this includes the most significant SNPs as cofactors in the mixed model. The mBonf criterion was used to identify the optimal results with Bonferroni correction at $\alpha = 0.01$.

Investigation on the putative function of candidate genes for adaptation

Common bean genes orthologous to *A. thaliana* and legume genes were identified using Orthofinder (*76*). The putative function of poorly-characterised genes was predicted based on orthologous relationships and literature screening. The orthologue and known genes involved in DTF, GH and PS were screened against

the GWAS results. Genes within 50 kb on either side of each significant SNP associated with DTF and GH, and genes located within selection scan and introgression scan regions, were investigated by GO term enrichment analysis (biological process, cellular component and molecular function) using Metascape (*77*) (http://metascape.org).

References

51. JMP®, Version 7.0.0. SAS Institute Inc., Cary, NC, 1989–2012

52. Wei T, Simko V (2017) R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

53. Peterson BG, Peter CP, Boudt K, Bennett R, Ulrich J, Zivot E, Cornilly D, Hung E, Lestel M, Balkissoon K, Wuertz D, Christidis AA, Martin RD, Zhou Z, Shea JM (2020) PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis (Version 2.0.4). Available from https://github.com/braverock/PerformanceAnalytics

54. Perez de Souza L, Scossa F, Proost S, Bitocchi E, Papa R, Tohge T and Fernie AR (2019) Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (*Phaseolus vulgaris*) metabolome. Plant J, 97: 1132-1153. https://doi.org/10.1111/tpj.14178

55. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN].

56. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-2993.

57. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-1303.

58. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11 10 11-33.

59. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) GigaScience, Volume 10, Issue 2, giab008, https://doi.org/10.1093/gigascience/giab008

60. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G (2011) The variant call format and VCFtools. Bioinformatics. 2011, 27(15):2156-8.

61. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly 6:80-92.

62. Browning S R and Browning B L (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097. doi:10.1086/521987

63. Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163(1):367-74. doi: 10.1093/genetics/163.1.367. PMID: 12586722; PMCID: PMC1462429.

64. Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9, 539 https://doi.org/10.1186/1471-2105-9-539

65. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35(6):1547–49. doi: 10.1093/MOLBEV/MSY096.

66. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, Volume 26, Issue 6, 15, 841–842, https://doi.org/10.1093/bioinformatics/btq033

67. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol., 25, 1965–1978.

68. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2020) vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

69. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics, 105(2):437-60.

70. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. evolution, 1358-1370.

71. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet., 78(4):629-44.

72. Zhang C, Dong SS, Xu JY, He WM, Yang TL (2019) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics, Volume 35, Issue 10, 1786–1788, https://doi.org/10.1093/bioinformatics/bty875

73. Zhang H, Meltzer P, Davis S (2013) RCircos: an R package for Circos 2D track plots BMC bioinformatics, 14(1):244.

74. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat. genet., 44(7), 821.

75. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. genet., 44(7), 825.

76. Di Vittori V, Bitocchi E, Rodriguez M, Alseekh S, Bellucci E, Nanni L, Gioia T, Marzario S, Logozzo G, Rossato M, De Quattro C, Murgia ML, Ferreira JJ, Campa A, Xu C, Fiorani F, Sampathkumar A, Fröhlich A, Attene G, Delledonne M, Usadel B, Fernie AR, Rau D, Papa R (2021) Pod indehiscence in common bean is associated with the fine regulation of PvMYB26. J. Exp. Bot., Volume 72, Issue 5, 27 February 2021, Pages 1617–1633, https://doi.org/10.1093/jxb/eraa553

77. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanasichuk O, Benner C, Chanda SK (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. 10, 1523. https://doi.org/10.1038/s41467-019-09234-6