UC Davis
UC Davis Previously Published Works

Title

Genomic history of the origin and domestication of common bean unveils its closest sister species

Permalink

https://escholarship.org/uc/item/2bn6d9w9

Journal

Genome Biology, 18(1)

ISSN

1474-760X

Authors

Rendón-Anaya, Martha
Montero-Vargas, Josaphat M
Saburido-Álvarez, Soledad
et al.

Publication Date

2017-12-01

DOI

10.1186/s13059-017-1190-6

Peer reviewed

Genome Biology

CrossMark

Genomic history of the origin and domestication of common bean unveils its closest sister species

Martha Rendón-Anaya[1], Josaphat M. Montero-Vargas[2], Soledad Saburido-Álvarez[1], Anna Vlasova[3,4], Salvador Capella-Gutierrez[3,4], José Juan Ordaz-Ortiz[1], O. Mario Aguilar[5], Rosana P. Vianello-Brondani[6], Marta Santalla[7], Luis Delaye[8], Toni Gabaldón[3,4], Paul Gepts[9], Robert Winkler[2], Roderic Guigó[3,4], Alfonso Delgado-Salinas[10] and Alfredo Herrera-Estrella[1*]

Abstract

Background: Modern civilization depends on only a few plant species for its nourishment. These crops were derived via several thousands of years of human selection that transformed wild ancestors into high-yielding domesticated descendants. Among cultivated plants, common bean (*Phaseolus vulgaris* L.) is the most important grain legume. Yet, our understanding of the origins and concurrent shaping of the genome of this crop plant is limited.

Results: We sequenced the genomes of 29 accessions representing 12 *Phaseolus* species. Single nucleotide polymorphism-based phylogenomic analyses, using both the nuclear and chloroplast genomes, allowed us to detect a speciation event, a finding further supported by metabolite profiling. In addition, we identified ~1200 protein coding genes (PCGs) and ~100 long non-coding RNAs with domestication-associated haplotypes. Finally, we describe asymmetric introgression events occurring among common bean subpopulations in Mesoamerica and across hemispheres.

Conclusions: We uncover an unpredicted speciation event in the tropical Andes that gave rise to a sibling species, formerly considered the "wild ancestor" of *P. vulgaris*, which diverged before the split of the Mesoamerican and Andean *P. vulgaris* gene pools. Further, we identify haplotypes strongly associated with genes underlying the emergence of domestication traits. Our findings also reveal the capacity of a predominantly autogamous plant to outcross and fix loci from different populations, even from distant species, which led to the acquisition by domesticated beans of adaptive traits from wild relatives. The occurrence of such adaptive introgressions should be exploited to accelerate breeding programs in the near future.

Keywords: Common bean, Domestication, Genomic introgression, Adaptive traits, Speciation

Background

The transition from hunting–gathering to agriculture is one of the major milestones in human evolution. An important, *sine qua non* consequence of this transition has been the domestication of crop plants and farm animals [1]. Furthermore, domestication provides an experimental model to study evolution in general, with

several advantages, including the existence of ancestral populations, an established time frame (~10,000 years), and identifiable traits under selection for both domesticated plants and animals [2]. In this perspective, *Phaseolus* species are of particular interest because of the multiple domestications that have taken place in this genus. Indeed, of the 70–80 wild species that have been described, no less than five species have been domesticated in contrasting ecogeographic settings: common bean (*P. vulgaris* L.); lima bean (*P. lunatus* L.); runner bean (*P. coccineus* L.); tepary bean (*P. acutifolius* A. Gray); and year bean (*P. dumosus* Macfady). In addition, the

* Correspondence: alfredo.herrera@cinvestav.mx
[1]Laboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del IPN (Cinvestav), 36821 Irapuato, Guanajuato, Mexico
Full list of author information is available at the end of the article

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 2 of 17

first two species were independently domesticated at least twice—in Mesoamerica and in the Andes—implying that some domestication traits may have been selected multiple times, as shown by the determinacy trait in common bean [3]. This is in contrast with other crops that have been subjected to fewer domestication events, such as maize (single domestication [4, 5]) or rice and wheat (three domestications [6–9]). The multiple domestication phenomenon in *Phaseolus* provides an opportunity to examine to what extent similar selection pressures have led to convergent evolution at the molecular level [10]. Conversely, comparative genomics can illustrate the differential genetic control of adaptation to contrasting environments in which the different *Phaseolus* species were domesticated.

Although a New World origin of the genus has been established by phylogenetic studies [11], the geographic origin of *P. vulgaris* has been strongly debated. Initial evidence suggested the Peruvian–Ecuadorian region as the center of origin, given that accessions collected there have an ancient form of the seed storage protein phaseolin [12, 13]. However, based on an analysis of five loci, Bitocchi et al. [14] proposed that common bean originated in Mexico and then colonized the Southern hemisphere, giving rise separately to the Peruvian–Ecuadorian populations and the wild Andean gene pool, both phylogenetically derived from the Mesoamerican clade. Despite the uncertainty regarding the geographic origin of *P. vulgaris*, several lines of evidence from traditional (allozymes or seed proteins) and more recent molecular markers [15–17] converge in the establishment of two geographically and genetically isolated gene pools, one in Mesoamerica and one in the central to southern Andes. From these pools, two independent domestications took place starting ~8000 years ago [18–21], followed by local adaptations and further expansions. Accompanying these processes, not only has the genetic diversity of the domesticated varieties decreased due to domestication bottlenecks, but it would also appear that hybridization events between wild and domesticated populations have occurred through time, as suggested by morphological variation and microsatellite diversity [22–26], displacing the original genetic diversity in these regions [27, 28]. At the same time, however, introgressions from wild relatives may have permitted domesticated varieties to acquire adaptive traits. In this regard, gene flow has been crucial for the adaptation of maize cultivars to different environmental conditions [29], as well as for the introduction of morpho-agronomic traits that increase the commercial value of domesticated varieties of tomato [30]. Accordingly, gene flow also deserves careful examination in the context of common bean evolution.

This study aims at understanding how the current distribution of common bean was originated and how its genome has been shaped by domestication and through gene flow from its close relatives, to become the crop plant we know. Accordingly, we built a robust evolutionary model of common bean lineage divergence and domestication. We relied on the availability of two recently published annotated whole-genome sequences, of Mesoamerican [31] and Andean [32] origins, and re-sequenced ten additional accessions from Mesoamerica (MA) and three from the southern Andes (AN), together with five genotypes from the Peruvian–Ecuadorian area originating in the Amotape–Huancabamba Depression (AH), and 11 Mesoamerican *Phaseolus* species from the Vulgaris, Filiformis, Lunatus, Leptostachyus, Polystachios, and Tuerckheimii phylogenetic groups [11]. We focused on three essential, novel aspects of this model that are key to establishing the basic domestication pattern of common bean. First, a phylogenetic analysis of the presumed ancestral *P. vulgaris*, which has an extraordinarily broad distribution extending from northern Mexico to northwestern Argentina, leads us to postulate that a cluster of wild populations in northern Peru–Ecuador actually represents a sibling species of *P. vulgaris*, which was not domesticated. Second, an analysis of the patterns of allelic admixture identifies signals of asymmetric intra-species and inter-species genomic introgression, which can represent the acquisition of adaptive traits by domesticated beans from its wild relatives. Third, we identify both shared and distinctive haplotypes associated with domestication traits between the Mesoamerican and Andean domestication processes.

Results
Genomic differentiation of *Phaseolus* species
Twenty-nine *Phaseolus* genomes, representing most of the species diversity in the genus, were sequenced at a coverage in the range of 8–20X (Additional file 1: Figure S1; Additional file 2: Tables S1 and S2). According to a previously proposed phylogenetic classification, which divides *Phaseolus* species into two sister clades [11], our sampling covered one of the three well-defined groups from clade A (Tuerckheimii) and had at least one representative species from each group of clade B (comprising all domestication events and having a broader distribution in the Americas), with an intentional bias towards the Vulgaris group. Raw reads were filtered and mapped against the *P. vulgaris* cv. BAT93 reference genome of 556.4 Mb (86% of the theoretical genome length of 650 Mb), as well as to a synteny-based pseudoassembly of BAT93 using the *P. vulgaris* G19833 genome as a scaffold. This pseudoassembly was produced using the SynMap tool at CoGe (https://genomevolution.org/coge/) with ≥ 4 contiguous syntenic CDSs between assembled tracks in order to construct longer

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 3 of 17

chromosomes with more certainty of the order and sense of the scaffolds than in the current BAT93 genome version. The calculated breadths of coverage (number of bases of the reference genome that were covered during the mapping process) were congruent with the phylogenetic closeness of each accession to *P. vulgaris* BAT93 [11], as they were in the range of 56–77% for species from the Filiformis, Lunatus, Leptostachyus, Polystachios, and Tuerckheimii groups; 60–88% for species from the Vulgaris group (e.g. *P. acutifolius* or *P. coccineus*), and up to > 90% for most *P. vulgaris* accessions (Additional file 2: Table S2).

Based on the collection sites of the *P. vulgaris* accessions, we grouped them in subpopulations (Table 1) that became useful for calculations requiring allele frequencies, such as intra-species and inter-species pairwise absolute genetic divergences (d_{XY}, calculated on 5 Kb non-overlapping windows; [33]) (Fig. 1a). We observed *P. vulgaris* intra-species average divergence values below 0.009 and higher d_{XY} values between *P. vulgaris* and its sister species *P. coccineus*, *P. dumosus*, and *P. costaricensis* Freytag & Debouck, in the range of 0.026–0.03. From this analysis, two contrasting results were noteworthy. First, accessions from the AH zone belonged to a narrowly restricted population, both at the geographic and genetic levels, as they were the least divergent accessions among all comparisons ($d_{XY} = 0.0023$). Second, the AH subpopulation and any other *P. vulgaris* group were equally divergent ($d_{XY} \approx 0.014$), as were the two well-defined sister species, *P. dumosus* and *P. costaricensis* ($d_{XY} = 0.011$). Not only were the d_{XY} values within *P. vulgaris* subpopulations and between *P. vulgaris* and AH accessions different (Kruskal-Wallis *p* value = 0.014), but the comparison of inter-species and intra-species distances (Fig. 1a) indicated that they were all derived from independent populations (Kruskall Wallis *p* value = 0.007). The genome-wide d_{XY} calculation indicates that the AH subpopulations represent a different lineage, divergent enough from *P. vulgaris* and from other close members of the Vulgaris group (*P. dumosus*, *P. costaricensis*, or *P. coccineus*) to be considered a different lineage.
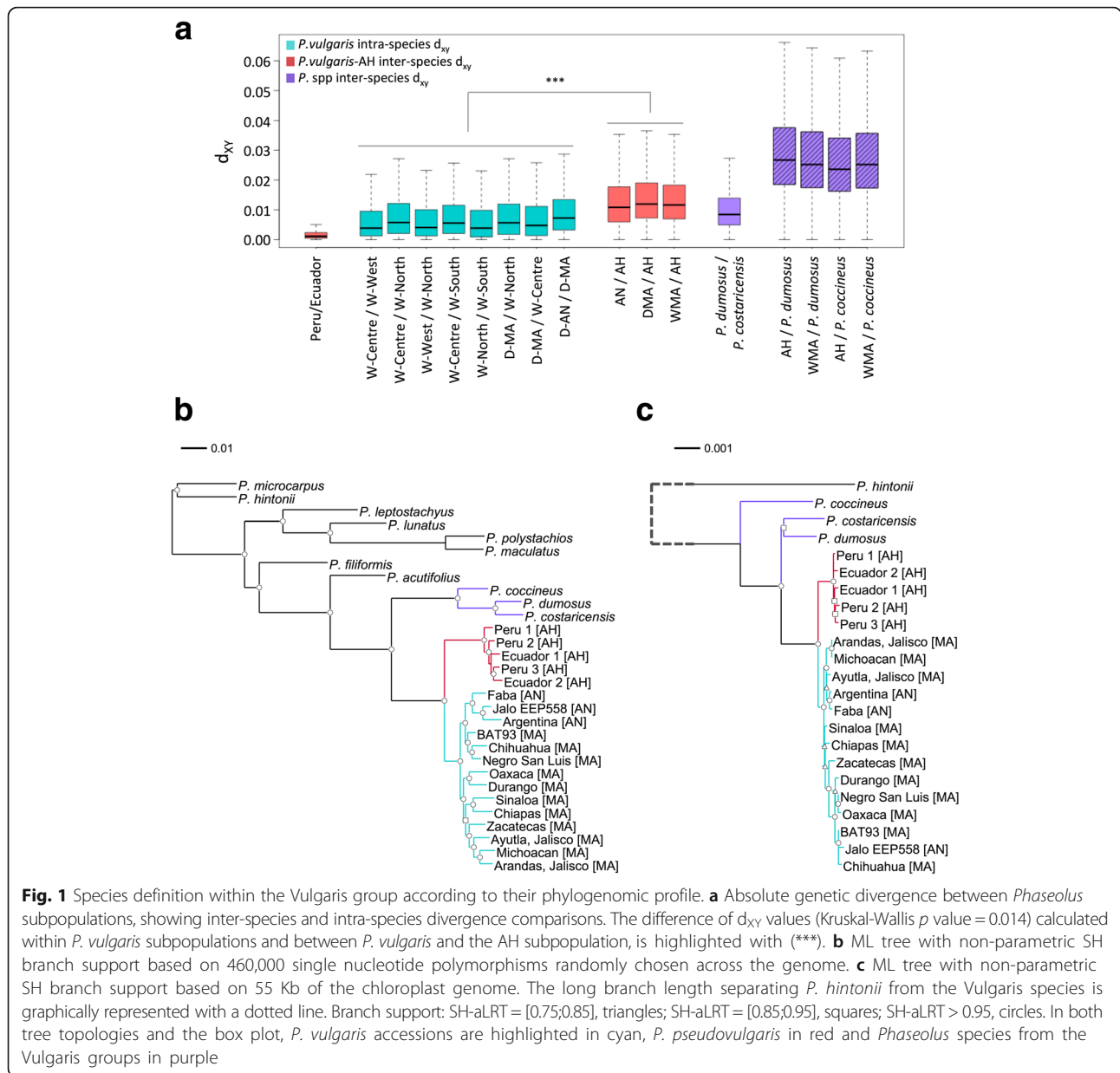
Uncovering the closest sister clade of *P. vulgaris*

We reconstructed the phylogenetic relationships of the species using genome-wide detected single nucleotide polymorphisms (SNPs) ($17.2 e^6$ SNPs were detected, and filtered to remove unique polymorphisms to a total of $7.4e^6$; Additional file 2: Table S3). This analysis uncovered an unpredicted novel lineage within the Vulgaris group (Fig. 1b). In contrast to previous reports in which wild accessions from northern Peru and Ecuador formed a clade derived from MA wild subpopulations [14], our maximum-likelihood tree (produced with PhyML using aLRT non-parametric SH branch support [34]) placed these individuals in a separate clade (SH-aLRT = 1), sister to all *P. vulgaris* genotypes, Andean and Mesoamerican. This signal remained consistent in each linkage group when individual phylogenies were reconstructed using specific SNPs for each chromosome ($6.1e^5$ SNPs on average per chromosome; Additional file 1: Figures S2–S4). We further corroborated this evolutionary relationship using a sequenced 55-Kb chloroplast genome fragment (cpDNA). The phylogenetic signal resulting from the cpDNA (Fig. 1c) was consistent with that observed for nuclear markers (Fig. 1b), pointing to a divergence of the AH genotypes and the *P. vulgaris* lineage that predated the split of the MA and AN wild gene pools. Moreover, these results confirm the hybrid speciation between *P. vulgaris* (maternal contributor) and *P. coccineus* (paternal contributor) that gave rise to *P. dumosus* [35, 36]. The high nuclear affinity between *P. dumosus* and *P. coccineus* and the remarkable chloroplast closeness of *P. dumosus* and *P. vulgaris* indicates that the cpDNA is indeed telling a linear evolutionary history that does not seem to be strongly influenced by recombination events in the organelle.

To provide a temporal frame of the divergence of the AH/*P. vulgaris* lineages, coalescent simulations with an uncorrelated lognormal relaxed molecular clock were performed using both nuclear (two independent sets of 150 and 170 genes) and plastid markers (55 Kb of the chloroplast genome). Several time priors were tested according to previous reports

Table 1 *Phaseolus* subpopulations, grouped based on their sites of collection

Subpopulation	Accessions	Origin	Type
North	Sinaloa; Durango	Mesoamerican	Wild
Center	Zacatecas; Jalisco	Mesoamerican	Wild
South	Chiapas; Oaxaca	Mesoamerican	Wild
West	Michoacán; Jalisco	Mesoamerican	Wild
DMA	BAT93; Negro San Luis; Chihuahua	Mesoamerican	Domesticated
AN	Jalo EEP558; Faba Andecha; G19901	Andean	Domesticated + Wild
AH	G21244; G21245; G23587; G23724; G23582	Amotape–Huancabamba Zone in Peru/Ecuador	Wild

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 4 of 17

Fig. 1 Species definition within the Vulgaris group according to their phylogenomic profile. **a** Absolute genetic divergence between *Phaseolus* subpopulations, showing inter-species and intra-species divergence comparisons. The difference of d_{XY} values (Kruskal-Wallis *p* value = 0.014) calculated within *P. vulgaris* subpopulations and between *P. vulgaris* and the AH subpopulation, is highlighted with (***). **b** ML tree with non-parametric SH branch support based on 460,000 single nucleotide polymorphisms randomly chosen across the genome. **c** ML tree with non-parametric SH branch support based on 55 Kb of the chloroplast genome. The long branch length separating *P. hintonii* from the Vulgaris species is graphically represented with a dotted line. Branch support: SH-aLRT = [0.75;0.85], triangles; SH-aLRT = [0.85;0.95], squares; SH-aLRT > 0.95, circles. In both tree topologies and the box plot, *P. vulgaris* accessions are highlighted in cyan, *P. pseudovulgaris* in red and *Phaseolus* species from the Vulgaris groups in purple

[11, 32], e.g. 0.165 Mya of divergence between the MA and AN *P. vulgaris* gene pools (Additional file 3: Tables S4 and S5); clock rates and μ values were adjusted for each dataset according to the calculated pairwise absolute genetic divergences (Additional file 3: Tables S4 and S5). Our results corroborated an early split of the AH lineage: the chloroplast genome shows a divergence time of 0.9 My [0.5–1.4, 95% Highest Posterior Density interval (HPD)] while the nuclear markers show a divergence time of 0.26 My (0.02–0.7, 95% HPD), much earlier than the separation of the MA and AN gene pools (0.2 My with 0.07–0.3 95% HPD for the plastid markers; 0.002 My

with 1.5E-4–5.9E-3 95% HPD for the nuclear gene sets (Additional file 3: Tables S6–S11; Figures S5–S10). The discrepancies observed in the coalescent results (older divergence using cpDNA than nuclear DNA) are most likely attributable to nuclear recombination events, as has been documented in other plant models [37–39].

Altogether, these results provide further support to our phylogenomic inferences and indicate that the AH group should be considered a separate lineage within the Vulgaris group. Based on these findings and the following data, we henceforth denote the AH group as "*Phaseolus pseudovulgaris*."

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 5 of 17

Metabolomic profiling differentiates *Phaseolus* species and implicates flavonoid production as a phenomenon accompanying species radiation

Similar morphology appears to have hidden evolutionary relationships between *Phaseolus* lineages, particularly when *P. vulgaris* and *P. pseudovulgaris* are contrasted. To establish a phenotypic discrimination of the species, other than morphological traits, we used high-throughput, non-targeted mass fingerprinting [40, 41]. Combining direct-injection electrospray mass spectrometry (DIESI-MS) and hierarchical cluster analysis (HCA), we constructed metabolic profiles of *P. vulgaris*, *P. pseudovulgaris*, and *P. coccineus* accessions from young trifoliate leaves (Fig. 2). More than 1000 different mass to charge signals (*m/z*) were recovered, representing the "metabolic space" of each accession. After mass error removal and signal filtering, 318 high quality mass signals of metabolites were kept for further analyses. HCA of the 100 most abundant metabolites correctly isolated *P. coccineus* as the outgroup and discriminated *P. vulgaris* accessions into wild or domesticated types. More importantly, the *P. vulgaris* accessions were separated from their sibling species, placing these accessions in two independent clades

(Additional file 1: Figures S11 and S12). Using a data mining approach [42], we identified the 30 variables that best explained the metabolic differences between the common bean populations. The dendrogram constructed from those variables replicated the phylogeny described in the previous section, with bootstrap and approximately unbiased (AU) probabilities supporting the topology (Fig. 2).

Using high-resolution liquid chromatography–mass spectrometry (LC-MS) data, we identified 44 metabolites, 25% of which were among the 100 variables that best explained inter-species differences. Most of the metabolite diversity in this set corresponded to flavonoids, such as the isobars of luteolin and kaempferol, or the coumarin derivative 4-methylumberlliferone (Additional file 2: Table S12) that play crucial roles during legume-microbe interactions in the rhizosphere (reviewed by [43]).

Asymmetric intra-species and inter-species genomic flow in Mesoamerica and across hemispheres

In spite of its preferential autogamy, *P. vulgaris* cannot be considered to have a closed reproductive system, as it maintains outcrossing rates in the range of 1–70%, depending
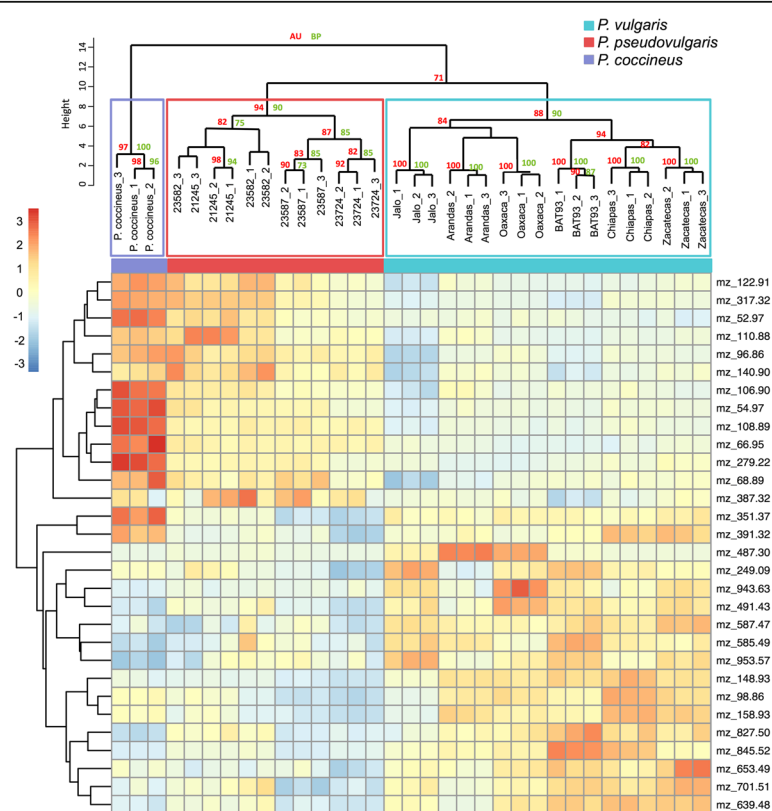

Fig. 2 Metabolomic profiles of *Phaseolus* species. The *heatmap* shows the 30 most informative mass signals from extracts of young trifoliate leaves that explain inter-species differences between *P. vulgaris*, *P. pseudovulgaris*, and *P. coccineus*. The associated horizontal dendrogram reproduces the phylogeny of the accessions, while the vertical dendrogram clusters mass signals according to their abundance. Approximately unbiased probabilities (AU) and bootstrap support (BP) ≥ 70 are displayed in the horizontal dendrogram

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 6 of 17

upon the experimental conditions [44, 45]. Therefore, combining the dynamic estimator of the degree of introgression between subpopulations (f_d, a modified version of Patterson's D statistic) and the absolute genetic divergence (d_{XY}) [46–48], we looked for allelic admixture signals within and between *Phaseolus* species. Several triads ($P_1P_2P_3O$) were considered to estimate the f_d parameter, permuting the donor (P_3) and receptor (P_1,P_2) subpopulations (Table 1) and fixing *P. hintonii* as the outgroup (O). We defined introgressed blocks as those windows that

belong to the top 5% f_d outliers and, at the same time, display d_{XY} values smaller than the mean d_{XY} across the whole genome. Genomic windows displaying such traits were condensed into larger blocks that, in several cases, were close to the size of recombination units.

We observed a clear tendency of increased introgression signals as we compared phylogenetically closer subpopulations. That is, intraspecific introgression occurred with higher frequency than interspecific introgression (Figs. 3a and 4c–m). The f_d values were close to 0.3
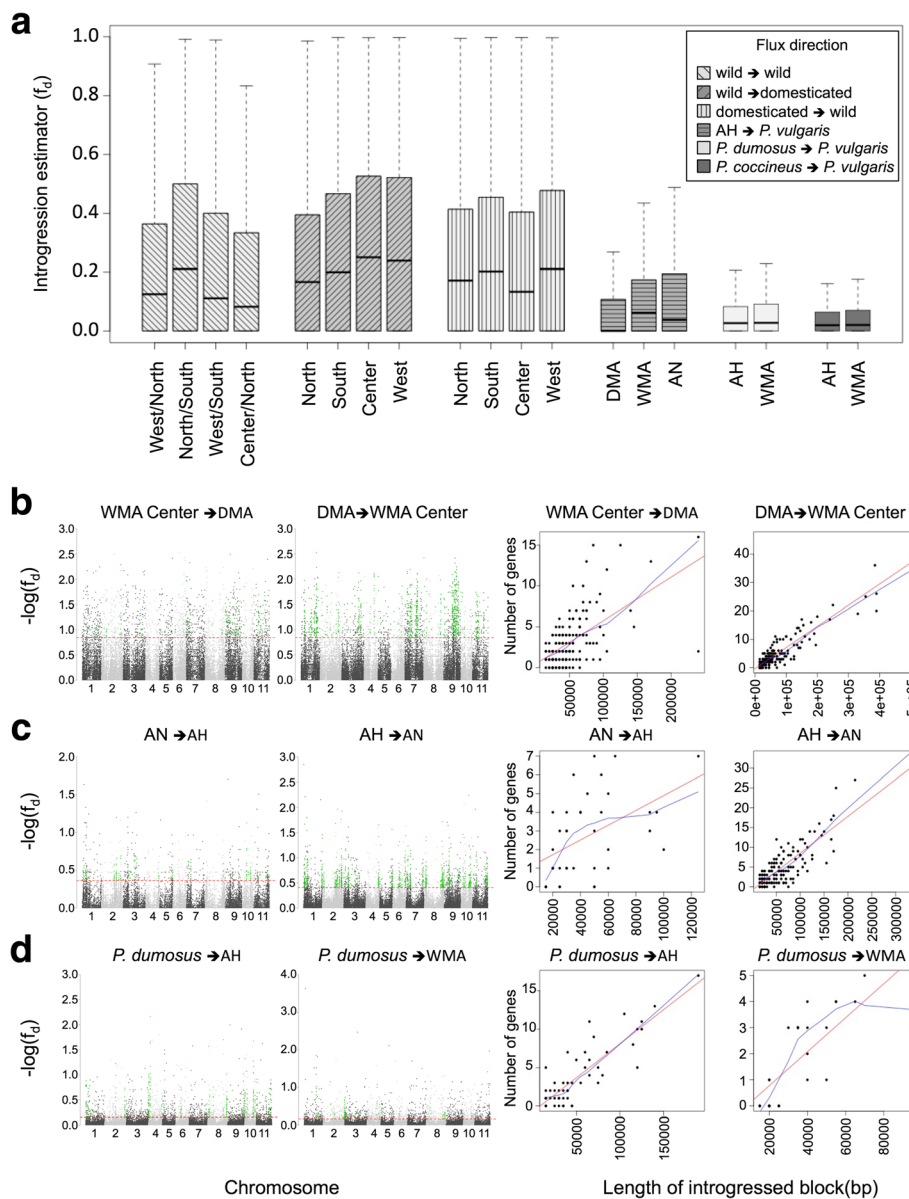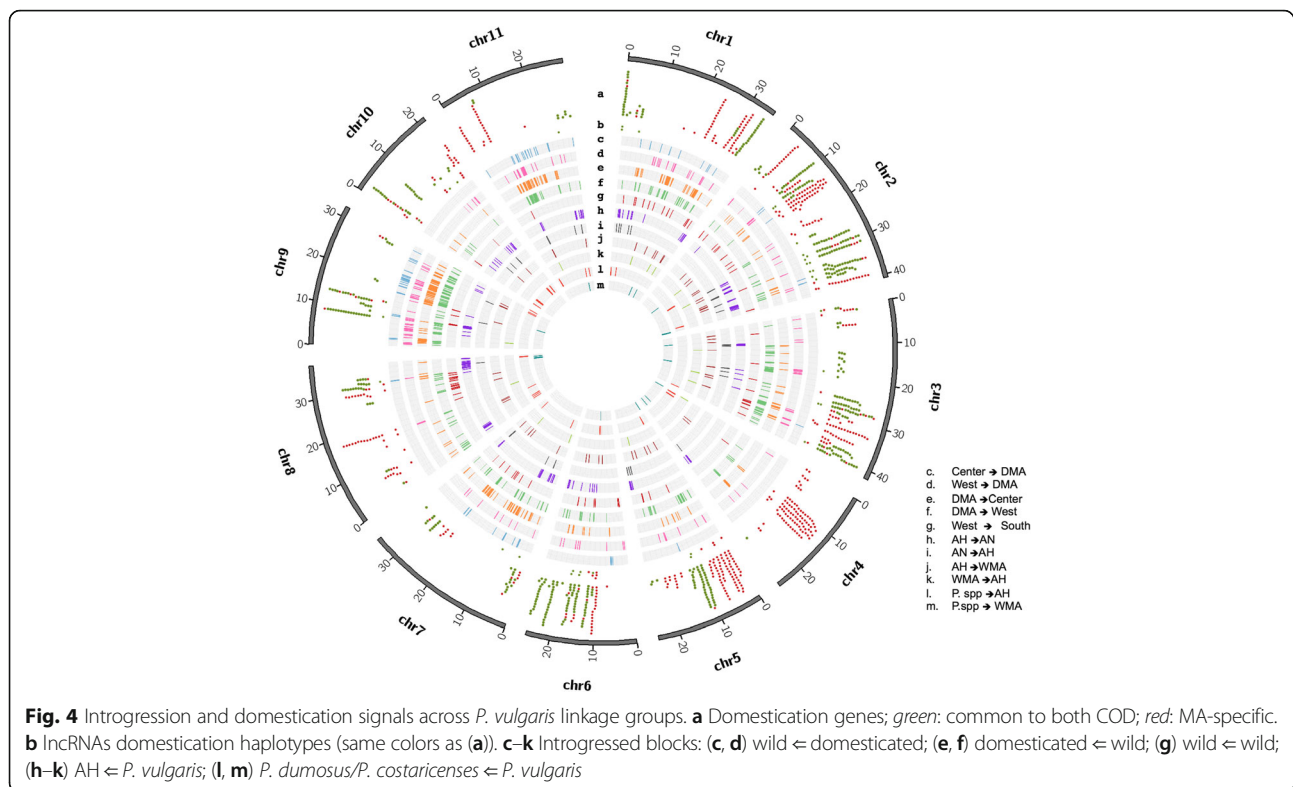

Fig. 3 Introgression rate (f_d) depends on phylogenetic distance between subpopulations. **a** Global f_d estimations for different triads of *Phaseolus* subpopulations. **b**–**d** Introgression signal across the linkage groups divided into 5-Kb non-overlapping windows is represented in *Manhattan plots* (*left panels*); the *red threshold lines* show the top 5% f_d outliers in each comparison, and strong signals of introgression ($f_d + d_{XY}$) are highlighted in *green*. The number of genes encoded in each introgressed block is represented in *scatterplots* (*right panels* – *colored lines*: linear [*red*] and local [*blue*] regressions). In (**d**), the donor subpopulation is conformed by *P. dumosus* and *P. costaricensis*

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 7 of 17

Fig. 4 Introgression and domestication signals across *P. vulgaris* linkage groups. **a** Domestication genes; *green*: common to both COD; *red*: MA-specific. **b** lncRNAs domestication haplotypes (same colors as (**a**)). **c–k** Introgressed blocks: (**c, d**) wild ⇐ domesticated; (**e, f**) domesticated ⇐ wild; (**g**) wild ⇐ wild; (**h–k**) AH ⇐ *P. vulgaris*; (**l, m**) *P. dumosus/P. costaricensis* ⇐ *P. vulgaris*

Legend within figure:
c. Center → DMA
d. West → DMA
e. DMA → Center
f. DMA → West
g. West → South
h. AH → AN
i. AN → AH
j. AH → WMA
k. WMA → AH
l. P. spp → AH
m. P.spp → WMA

between *P. vulgaris* subpopulations, regardless of their wild or domesticated origin, whereas inter-species f_d values dropped to 0.05–0.1. While admixture can occur in both directions, we observed a larger genomic contribution in terms of the total length of introgressed tracks and transferred protein coding genes (PCGs) from domesticated into wild subpopulations (5.7–17.1 Mb) than from wild into domesticated genotypes (4.1–8.2 Mb; Additional file 1: Figures S13–S16). This was particularly evident when we took the Central subpopulation (Zacatecas and Jalisco) of wild MA genotypes as the receptor in the triad (Fig. 3b), which is consistent with local records that place Zacatecas and Jalisco among the most important states that produce common bean in Mexico and the widespread distribution of wild *P. vulgaris*, especially in Jalisco [49].

Interestingly, we detected asymmetric gene flow between the Northern and Southern American hemispheres, taking the AH genotypes as an intermediate subpopulation (Additional file 1: Figure S17). Although introgression was detected in both directions, from AH to MA and from AH to AN, it was stronger towards the Andean accessions (Fig. 3c), a fact that could be explained by the geographic closeness of the populations along the Andean corridor (e.g. [50]) and the lower levels of genetic diversity in the southern Andean region [14] that contribute to the maintenance of long introgressed blocks. Furthermore,

the AH subpopulation appeared to be preferentially autogamous based on two observations: first, genetic diversity in the Amotape–Huancabamba Depression group was lower than in any other tested subpopulation ($\pi_{AH} = 1.7e^{-3}$, $\pi_{South} = 5.7e^{-3}$, $\pi_{West} = 3.3e^{-3}$, $\pi_{North} = 6.6e^{-3}$, $\pi_{Center} = 5.4e^{-3}$, $\pi_{DMA} = 4.3e^{-3}$, $\pi_{AN} = 6.1e^{-3}$; Additional file 1: Fig. S18); second, introgression signals were weaker when these genotypes were permuted as receptors in the test triad (Fig. 3c). These results indicate that while *P. vulgaris* plants growing in the Southern hemisphere could be cross-pollinated by their AH wild neighbors, this did not occur generally in the opposite direction. It is also possible that interspecific introgressions from sister species that reached the Southern hemisphere, such as *P. coccineus* or *P. dumosus* (Fig. 3d), enhanced the differentiation of *P. pseudovulgaris* from *P. vulgaris* (Additional file 1: Figure S19).

Parallel domestication events share signatures of selection in both common bean gene pools

An advantage of our approach was that whole-genome re-sequencing of individuals allowed us to define haplotypes across linkage groups and test their association with domesticated phenotypes in a case-control design. We looked for haplotype clusters (i.e. clustering of haplotypes on a localized basis: at the position of each genetic marker, haplotypes are clustered according to

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 8 of 17

their similarity in the vicinity of the position) in 19 accessions including all *P. vulgaris* and *P. pseudovulgaris*. Two combinations of phenotypes (cases) were evaluated, one including the three domesticated accessions from MA and the second adding the two AN domesticated cultivars to define those haplotypes common to both domestication processes; the rest of the accessions were kept as controls. To test the haplotype–phenotype association, we followed the two-layer hidden Markov model with linear approximation implemented in hapQTL [51]. Haplotypes strongly associated with the domestication process were defined as such if their calculated Bayes factors were higher than those obtained after permuting case-control labels. The effect of such haplotypes (e.g. altering coding sequences, 5'/3' UTRs or introns) was further evaluated. We selected as domestication candidates those genes that contained at least two markers with high association factors that were affecting regulatory regions, had non-synonymous effects on the coding sequence, or altered splicing sites or stop codons.

Following this procedure, we identified 599 genes with haplotypes shared between domesticated genotypes from MA and AN, and 628 genes with haplotypes specific to MA domesticated accessions (Fig. 4a; Additional file 4: Table S13; Additional file 5: Table S14). Similarly, 52 and 45 long non-coding RNAs (lncRNAs) with domestication-associated haplotypes were shared by the two centers of domestication (CODs) and within MA, respectively (Fig. 4b). These observations indicate that domestication has affected PCGs and regulatory elements, whose functions and targets should be further explored.

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

The functional descriptions of PCGs transferred by introgression events revealed several pathways of potential importance for crop improvement. First, GO terms related to cell wall biogenesis and organization, and pectin and cell wall polysaccharide metabolic processes, were enriched among introgressed genes transferred from *P. coccineus* and *P. dumosus/P. costaricensis* into *P. vulgaris*, which could have contributed to the acquisition of pathogen resistance in *P. vulgaris* [52]. Enrichment of functional terms associated with hormone-mediated signaling pathways, reproductive processes, post-embryonic development, and the formation of reproductive organs was associated with gene flow among *P. vulgaris* subpopulations. Contrary to the mobility of genes behind reproductive processes within *P. vulgaris* subpopulations, such categories were not statistically enriched when *P. pseudovulgaris* and *P. vulgaris* were evaluated. Interestingly, as reported in other crops [53], genes involved in biotic and abiotic stress responses were transferred in most of the *P. vulgaris* triads in both directions (Additional file 1:

Figures S20 and S21; Additional file 6: Tables S15; Additional file 7: Table S16; Additional file 8: Table S17), implying that the continuous movement of such loci favored the adaptation of common bean to different habitats. Genes within these categories corresponded to WRKYs, leucine-rich repeat receptor kinases, and pathogenesis-related proteins, among others (Fig. 5a and c).

Screening of protein definitions associated with the domestication gene candidates identified 21 disease resistance genes and several significantly enriched GO categories (Fisher's exact test, $p < 0.05$) that could be easily linked to the emergence of domestication traits (Fig. 5b; Additional file 9: Table S18; Additional file 10: Table S19). For instance, haplotypes common to both CODs affected components of the sucrose/starch biosynthetic pathway (directly related to starch content in the seeds), the regulation of reproductive processes (involving homologs of transcription factors such as WOX2 for embryonic patterning, or GTE1, which promotes seed germination), inflorescence development and meristem determinacy (e.g. homologs of the homeobox gene *KNOTTED1*). In addition, we identified genes particularly marked by SNPs in their regulatory regions, such as a homolog of *NODULATION-SIGNALLING PATHWAY 2* (*NSP2*), which is involved in Nod factor signaling in legumes, and several genes related to the dormancy and photoperiod sensitivity pathways (Fig. 5d). Other enriched categories such as chromatin assembly, nucleosome organization, and the regulation of histone methylation, were suggestive of epigenetic control in the emergence of domestication traits, which should be explored further. Among genes with MA haplotypes, we identified enriched GO categories particularly related to the development of reproductive structures or other organ formation (including homologs of transcription factors KAN2 or AS1), and genes directly involved in auxin transport and homeostasis or nodulation (*EARLY NODULIN 93*).

Discussion

P. pseudovulgaris emerged by allopatric speciation in the Amotape–Huancabamba Depression before the split of both *P. vulgaris* gene pools

Our combined genomic and phenotypic data support the Mesoamerican origin of common bean but focus further questions on the diversification steps immediately before and after *P. vulgaris* speciation. We supply strong evidence pointing to an early speciation event in the western tropical Andes, which clarifies most of the discrepancies introduced by noisy phylogenetic signals of genotypes collected in the northern Peru–Ecuador area of Amotape–Huancabamba [14]. Indeed, this enclosed area has been described as a transition zone between the Northern and Central Andes, where climate dynamics and oro-geographic conditions have produced the highest degree of plant species diversity and endemism along the
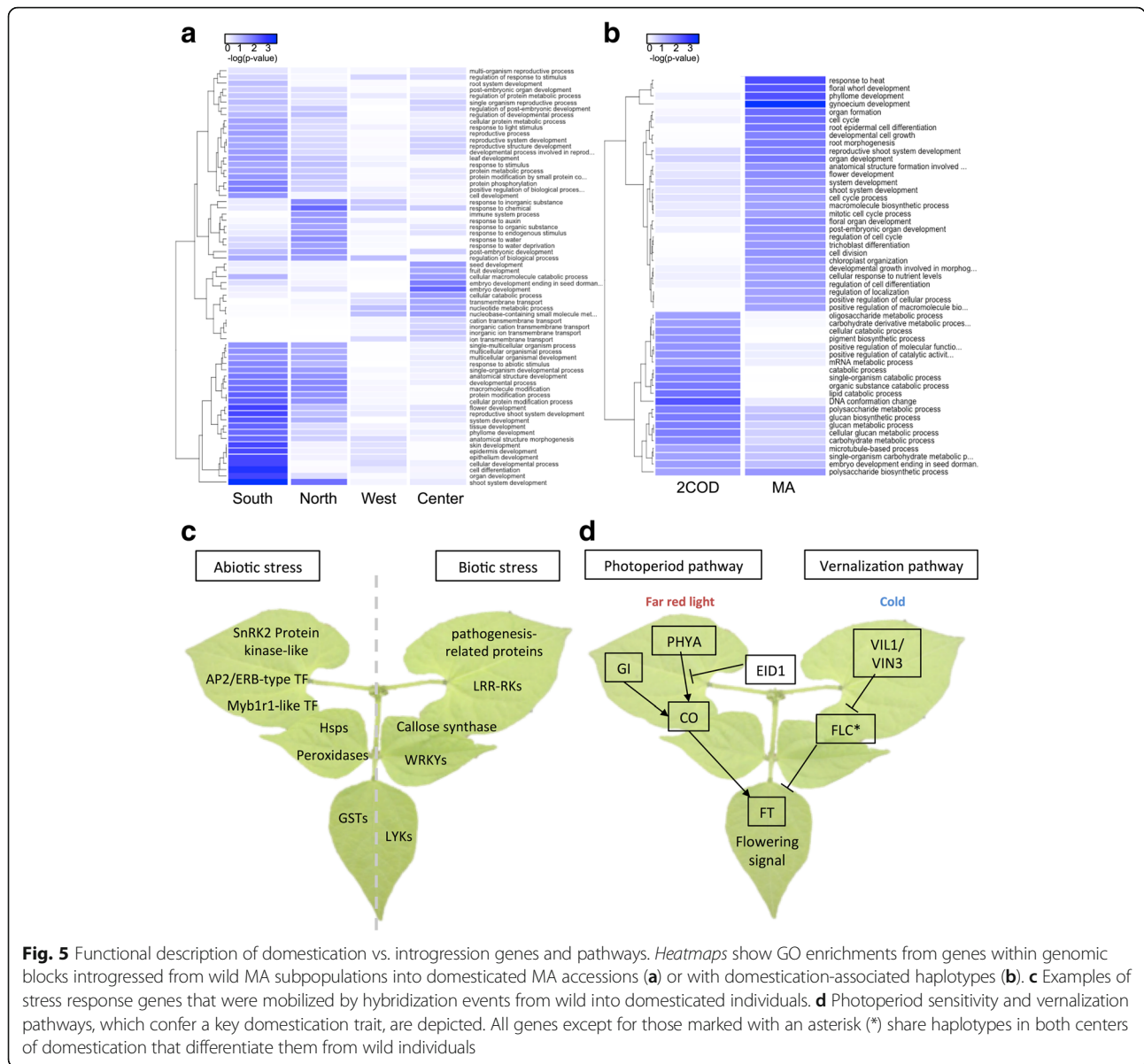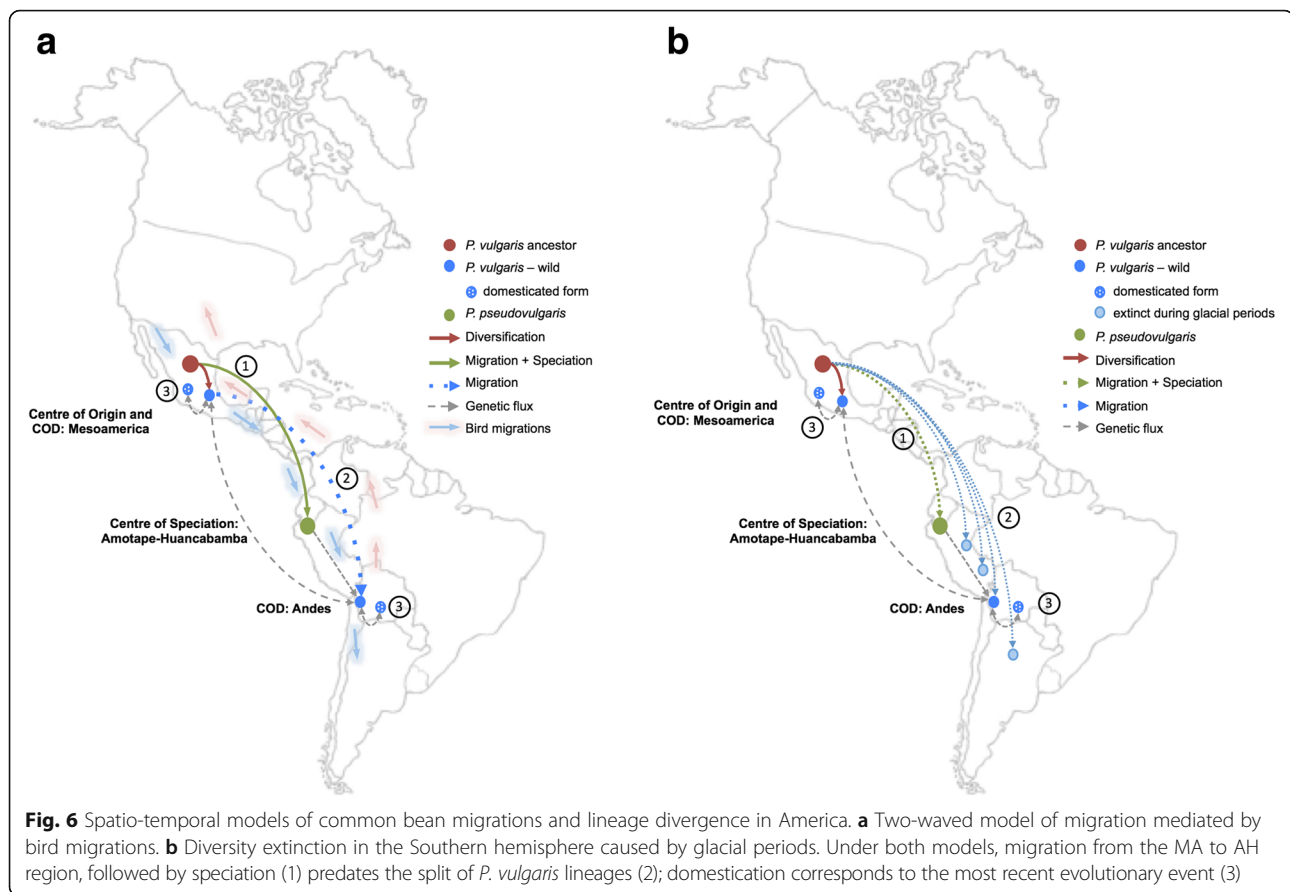
Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 9 of 17

Fig. 5 Functional description of domestication vs. introgression genes and pathways. *Heatmaps* show GO enrichments from genes within genomic blocks introgressed from wild MA subpopulations into domesticated MA accessions (**a**) or with domestication-associated haplotypes (**b**). **c** Examples of stress response genes that were mobilized by hybridization events from wild into domesticated individuals. **d** Photoperiod sensitivity and vernalization pathways, which confer a key domestication trait, are depicted. All genes except for those marked with an asterisk (*) share haplotypes in both centers of domestication that differentiate them from wild individuals

Andean Mountains [54–56]. Furthermore, the AH region represents a contrasting environment compared with that of other wild *P. vulgaris* in the Andes, including populations in Colombia and Venezuela (Mesoamerican gene pool [57]) and those in central and southern Peru, Bolivia, and Argentina (Andean gene pool [50]). Following this line of evidence, we propose a two-waved migration event in which *P. vulgaris*—or an ancestral form of the species—dispersed from Mesoamerica, reaching the AH zone between Northern and Central Andes, where it remained isolated and underwent allopatric speciation (Fig. 6a). This could have occurred through seed dispersal by birds [57, 58] following migration routes along the narrow isthmus connecting North and South America and following the Andean corridor of mountains but not necessarily

reaching the western side of Peru and Ecuador when migrating from south to north [59]. Hundreds of thousands of years later, a small population of *P. vulgaris* with Mesoamerican genetic background likely invaded the Central and Southern Andes, giving rise to the second gene pool that was later domesticated. A second spatiotemporal model considers glacial periods in the Southern Andes during the Pleistocene (Fig. 6b). The fact that other *Phaseolus* species from the Vulgaris group can be found in Central America, reaching Colombia, suggests that seeds from this group, including the ancestral lineage of *P. vulgaris*, spread into South America across the Isthmus of Panama or by long-distance dispersal waves that have been dated for terrestrial organisms at 20 Mya and 6 Mya [60]. Climatic changes during the Pleistocene resulted in

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 10 of 17

Fig. 6 Spatio-temporal models of common bean migrations and lineage divergence in America. **a** Two-waved model of migration mediated by bird migrations. **b** Diversity extinction in the Southern hemisphere caused by glacial periods. Under both models, migration from the MA to AH region, followed by speciation (1) predates the split of *P. vulgaris* lineages (2); domestication corresponds to the most recent evolutionary event (3)

recurrent expansions and contractions of *Phaseolus* populations, placing them in glacial refuges and limiting gene flow between them. The Amotape–Huancabamba Depression could have been an important glacial refuge, favoring the isolation and diversification of the AH populations around 1 Mya, which coincides with our cpDNA coalescent results. The outcome of a glacial period spanning 140–180 Kya (reviewed by [61]) could have been a small remaining founder population of *P. vulgaris* in South America that was domesticated afterwards, as this glacial period matches the age of the split between the MA and AN gene pools and the suggested bottleneck duration in the Andean wild population [32].

In terms of phenotypic distinctiveness, we identified secondary metabolites such as flavonoids as variables that allow inter-species discrimination. It is noteworthy that among them we found luteolin, a strong inducer of Nod gene expression [62, 63], a chemo-attractant and a growth regulator of rhizobia [64]; kaempferol, a flavonol involved in the regulation of auxin transport in response to rhizobia [65]; and 4-methylumbelliferone, implicated in controlling lateral root formation [66]. In this regard, *Phaseolus* species have the capacity to select their symbionts from coexisting soil bacteria [67, 68]. A survey of nodule bacteria [69] revealed a clear preference for

nodulation by *Bradyrhizobium* in most *Phaseolus* species and a shift to *Rhizobium* nodulation in the Vulgaris clade. These observations suggest that alterations in legume-root nodule symbiosis and symbiont preference shifts have accompanied *Phaseolus* speciation and diversification in the Americas.

Insights into genetic barriers for reproduction in the genus
Given the short time period separating the AH *Phaseolus* populations from their *P. vulgaris* relatives from MA and AN, reproductive barriers have not been fully established, as reflected by introgression signatures from AH into Andean genotypes (Fig. 3) and the observation of weedy populations in the AH Zone [70]. This could be attributable to the morphological similarity between species (Additional file 2: Table S20). At the same time, the limited outcrossing of MA genotypes with AH, but not necessarily between AH and AN genotypes, could be due to the geographic barriers imposed by the Andean corridor; however, this does not exclude the possibility of outcrossing between the AH subpopulations and neighboring *P. vulgaris* from Venezuela and Colombia. The selective gene flow to the AN genotypes is also in agreement with a previous report of unsuccessful crosses between a genotype from Cajamarca, Peru (G21245),

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 11 of 17

included in our sampling, and two tester lines from MA (G04830) and AN (G00122) [71]. In the same report, crosses of other genotypes showed hybrid weakness; accession G21245 was successfully crossed with 36% and 75% of the tested MA and AN genotypes, respectively, showing an asymmetric reproductive barrier [71].

The lack of an introgression signal affecting loci implicated in reproductive processes between *P. vulgaris* and *P. pseudovulgaris* gives further support to the reproductive isolation of accessions endemic to the AH Zone. Alleles that determine species phenotypes tend to introgress at a very low frequency [72]. Therefore, the observation that genes involved in reproductive processes are not transferred strongly suggests that these loci are important for the establishment of reproductive barriers between these close species. Previous studies have attempted to identify the genetic sources of incompatibility between the AN and MA gene pools, so far attributable to the root-expressed and shoot-expressed semi-dominant alleles DOSAGE-DEPENDENT LETHAL 1 (*DL1*) and *DL2* [73]. The differential introgression that we observe within *P. vulgaris* and between *P. vulgaris* and *P. pseudovulgaris* might provide insight into the genetic basis of the reproductive isolation in the genus.

Modern cultivars result from the combined outcome of domestication and adaptive introgressions

Standing genetic diversity is a prerequisite for more rapid adaptation in response to selection pressures and constitutes the raw material to develop improved breeds or cultivars [74]. A survey of the use of wild germplasm in crop improvements over the last decades [53], which included use in rice, wheat, maize, cassava, potato, and bean, among other crops, revealed that over 80% of the reported beneficial traits conferred by genes from wild relatives are involved in pest and disease resistance. Clearly, systematic efforts to bring genetic diversity from wild relatives into crop plants to incorporate a wider range of useful adaptations are required to increase the resiliency and productivity of crops.

Our data on common bean permit us to define introgression signals and differential haplotypes that, for the first time, can be combined to define domestication and putative adaptation loci. We confirmed a remarkable asymmetry of gene flow between wild and domesticated common bean subpopulations, as previously measured using microsatellite diversity [27]. Introgression signals between wild accessions might be disrupted as a consequence of the frequent hybridizations that have maintained high levels of genetic diversity (Additional file 1: Figure S13). In addition, introgression in domesticated genotypes from wild neighbors could be limited through selection against hybrids where wild traits, which are dominant or semi-dominant [75], are easily recognized by farmers. We cannot exclude, however, that

the high introgression signals $(f_d + d_{XY})$ between our defined subpopulations could actually underestimate the spans in each triad, which might cover larger portions of the genome if homogenous populations were tested.

As expected [28], domestication gene candidates do not overlap with introgressed regions since selective sweeps, i.e. homozygous regions that are rich in domestication-associated haplotypes, do not display signals of introgression. This is consistent with observations in other crops, such as maize, where domestication genes act as barrier loci for introgression events [29]. There is strikingly little overlap (65 PCGs) between our domestication PCG candidates and those reported previously [32]. This could be due to the absence of gene flow estimations in the previous model, as some of the reported loci that differentiate landraces from wild genotypes may not be the outcome of artificial selection, but rather represent admixture with other gene pools. Indeed, 40% of the published domestication gene candidates [32] recognized in the BAT93 gene set were within introgressed tracks in our test triads, implying that those might be neutral loci easily transferred between common bean subpopulations.

Conclusions

First, the studies reported here demonstrate—based on genomic and metabolomics data, and reproductive isolation—how the ancestral nature of the AH wild populations in northern Peru and Ecuador derived from a dissemination and speciation event that preceded the *P. vulgaris* speciation event and the split of the latter species into two major geographic gene pools (Mesoamerican versus Andean), based on structural genomic, reproductive isolation, and metabolomic data. Second, they provide a genome-wide picture of the importance of gene flow in common bean, a predominantly autogamous species, in providing local adaptation both in wild and domesticated populations. They also confirm the predominance of the domesticated to wild gene flow. Third, they illustrate how the domesticated genome of common bean has been shaped not only by selection under domestication but also by gene flow from other common bean population and closely related species, like *P. dumosus* and *P. coccineus*. This gene flow may have led to adaptive genomic introgressions permitting the adaptation of cultivars to environments outside their centers of domestication. This capacity, as well as the high frequency of stress response genes, should be exploited to enrich the genetic diversity base of breeding programs.

Methods
Plant material
Phaseolus vulgaris cv. BAT93 is a breeding line developed at the International Center for Tropical Agriculture (CIAT, Cali, Colombia) and derived from a double cross

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 12 of 17

involving four Mesoamerican genotypes: (Veranic × Tlal-nepantla 64) × (Negro Jamapa × Tara). Its genome se-quence was recently published [31] and was defined as our reference for downstream analyses. The biological material collected for this analysis included other im-portant *P. vulgaris* accessions: eight wild Mesoamerican genotypes, selected according to their geographical dis-tribution along the Mexican territory; one landrace from Chihuahua (Mexico); Jalo EEP558, a selection from the Andean landrace Jalo obtained by R. Guazelli at the Estação Experimental de Pato de Minas (Minas Gerais, Brazil); Faba Andecha, a Spanish cultivar of Andean ori-gin selected based on its domesticated traits; an Andean wild accession from Argentina (G19901); and five acces-sions from Peru and Ecuador considered by other authors to represent the ancestral form of the species based on of their phaseolin isoform (PhI, [6]), all collected in the con-strained location of the Amotape–Huancabamba Deflec-tion [48]. Outside the *P. vulgaris* species, we selected 11 additional species covering most of the clade diversity of the genus, according to [11]. These species corresponded to the Tuerckheimii group (*P. hintonii* A. Delgado) and the unclassified group (*P. microcarpus* Mart.) from clade A and, from clade B, the groups Filiformis (*P. filiformis* Benth.), Lunatus (*P. lunatus* – lima bean), Polystachios (*P. polystachios* Britton and *P. maculatus* Scheele), Leptosta-chyus (*P. leptostachyus* Benth.), and Vulgaris (*P. coccineus*, *P. dumosus*, *P. costaricensis*, and *P. acutifolius*). Plants were grown under greenhouse conditions and young trifoliate leaves were collected for DNA extraction.

DNA/RNA sequencing and mapping
DNA libraries were constructed and sequenced from both ends (paired-end reads) using the HiSeq (Illumina) technology at the Genomic Services Laboratory of LANGEBIO-CINVESTAV, Mexico. Reads of high quality (FastQC and FastxToolkit) were mapped with BWA v0.7.9a [76] using default parameters against the *P. vul-garis* BAT93 reference genome, as well as to a syneny-based pseudoassembly produced with SynMap at CoGe ([77]; https://genomevolution.org/coge/) of BAT93, tak-ing the G19833 genome as scaffold with at least four contiguous syntenic CDSs between assembled tracks.

Nuclear and chloroplast phylogenetic profiles
For each sequenced accession, individual–specific con-sensus sequences were generated and small variants (SNPs) were identified with ANGSD v0.614 [78] with the following options: major and minor alleles were in-ferred with doMaf = 1and doMajorMinor = 1; the geno-type likelihood was calculated using the samtools method with GL = 1; positions at each chromosome covered in all 30 genotypes were controlled with minInd = 30; SNP-pval = 1e-6. Depth adjustments for SNP calling and consensus

sequence reconstruction were done taking into account the sequencing depth of each accession: for all but four *P. vulgaris* accessions (Zacatecas, Oaxaca, Michoacán, Jalisco-Arandas) for which the depth threshold was set at five reads, a minimum of ten reads was required. Called SNPs in positions that were covered in all accessions were considered for further analyses.

From the collections of SNPs for each chromosome, singletons (unique SNPs for a particular genotype) were removed to avoid noisy signals derived from long-branch attraction effects (Additional file 2: Table S3). The filtered polymorphisms were then used to reconstruct phylogen-etic trees based on the maximum likelihood (ML) ap-proach. ML trees were reconstructed using the best-fitting evolutionary model, selected with PhyML v.3 [34] and using aLRT non-parametric SH branch support.

A 55-Kb chloroplast sequence was derived from scaf-fold00910 of the current BAT93 assembly, which was BLAST searched against the available genomic sequence of the plastid from *P. vulgaris* Negro Jamapa [79], dis-playing 99% identity. The consensus sequence of this scaffold was obtained as described above for the acces-sions belonging to the Vulgaris group and *P. hintonii*, as the outgroup. The 55-Kb plastid tracks were aligned and cleaned with TrimAl v1.3 [80]; the corresponding tree topology was constructed with the ML approach imple-mented in PhyML, using aLRT non-parametric SH branch support.

Coalescent simulations
To obtain a temporal frame of the divergence between AH genotypes and the *P. vulgaris* clade, we conducted coalescent simulations using the same 55-Kb chloroplast sequence fragment as used in the phylogenetic analysis to avoid noisy signals from recombination events in the nuclear markers. We used the Bayesian approach imple-mented in BEAUti and BEAST v2.3.0 [81], considering only six genotypes: BAT93 and Jalo EEP558 (as represen-tative genotypes of the MA and Andean genepools, re-spectively), one accession from Peru (G21245), *P. dumosus*, *P. costaricensis*, and *P. coccineus*. See Additional file 3 for more details on the priors for the simulations.

Introgression signal
We combined two different parameters [48], the dy-namic estimator of the degree of introgression between subpopulations (f_d) and the absolute genetic distance (d_{XY}). In principle, genomic regions that behave as f_d outliers can be distinguished as introgressed from ances-tral variation if the absolute genetic distance d_{XY} is also reduced between a donor (P_3) and a receptor population (P_2), given that in the presence of gene flow, genomic windows coalesce more recently than the species split, so the magnitude of reduction in P_2-P_3 d_{XY} is greater

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 13 of 17

than in the absence of recombination and hybridization. The f estimator was derived from the ABBA-BABA D statistic (Eq. 1a), and it assumes unidirectional gene flow from P_3 to P_2 (i.e. P_3 is the donor and P_2 is the recipient). In the case of the dynamic estimator f_d, the denominator is calculated by defining a donor population (P_D) for each site independently. For each site, P_D is the population (either P_2 or P_3) that has the higher frequency of the derived allele, thus maximizing the denominator and eliminating f estimates greater than 1 (Eq. 1b):

$$
\text{a.} \quad D(P_1, P_2, P_3, O) = \frac{\sum C_{ABBA}(i) - C_{BABA}(i)}{\sum C_{ABBA}(i) + C_{BABA}(i)}
$$

$$
\text{b.} \quad \hat{f}_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)}
$$

$$(1)$$

Eq. 1. Introgression estimators. (**a**) Patterson's D statistic. $C_{ABBA}(i)$ and $C_{BABA}(i)$ are counts of either 1 or 0, depending on whether the pattern ABBA or BABA is observed at site I in the genomic block. P_1/P_2: receptor populations; P_3: donor population; O: outgroup species. (**b**) Dynamic estimator. S: the difference between sums of ABBAs and BABAs, calculated using the frequency of the derived allele at each site in each population rather than binary counts; P_D: the population (either P_2 or P_3) with the higher frequency of the derived allele that maximizes the denominator.

Introgressed blocks that belong to the top 5% f_d outliers that, at the same time, display d_{XY} values smaller than the mean d_{XY} across the whole genome were condensed using a custom R script to define genomic windows of at least three 5-Kb neighboring blocks. The parameters f_d, d_{XY}, π, and D were calculated for 5-Kb non-overlapping windows along the 11 linkage groups of the synteny-based pseudoassembly of BAT93, using the pipeline reported by [48] and available at http://datadryad.org/resource/doi:10.5061/dryad.j1rm6.

Selection during domestication

Given that our sampling and sequencing strategies produced whole genomes from individuals belonging to different locations, we computed haplotype probabilities to identify haplotype clusters strongly associated with the domesticated phenotype. Since our sampling was biased to MA collections, particularly for wild *P. vulgaris* genotypes, we were not able to distinguish domestication haplotypes unique to the Andean cultivars, as we had one single wild accession from this area. However, we could differentiate haplotype clusters shared both by MA-domesticated and AN-domesticated cultivars, and those that emerged exclusively during the domestication process in MA.

For this purpose, we used the complete collection of SNPs of each *P. vulgaris* accession, including the genotypes from northern Peru and Ecuador as part of the wild subpopulation (19 genotypes in total) that were identified (Additional file 2: Table S3). The lists of non-unique SNPs from each chromosome were converted into tped files and then to bimbam format using Plink [82]. The resulting files were used as input for hapQTLv0.99 [51], a haplotype association method that relies on a hidden Markov model, and is suitable for large datasets to infer ancestral haplotypes and their loadings at each marker for each individual. With this algorithm, the local haplotype sharing (LHS)—the probability of two diploid individuals descending from the same ancestral haplotypes and thus a natural extension of identity by descent —can be quantified using the loadings. By testing whether the genetic similarity is associated with a particular phenotype, hapQTL is able to identify associations at each (core) marker between local haplotypes and phenotypes. For all hapQTL independent runs at each chromosome, we used two upper-layer clusters, two lower-layer clusters, and 20 steps in the EM runs using linear approximation; the rest of the parameters were kept as default. Two combinations of phenotypes were defined: (1) BAT93, Negro San Luis and Chihuahua labeled as "cases" of domestication in MA (DMA) and the other 16 genotypes (wild MA, AN and AH) as "controls;" and (2) BAT93, Negro San Luis, Chihuhahua, Jalo EEP558, and Faba Andecha labeled as "cases" of domestication in both COD, and the rest (wild MA, wild AN and AH) as "controls." For each domestication phenotype, we permuted case–control labels once and computed Bayes factors, treating these as Bayes factors under the null. Based on the permutation tests, Bayes factors (bf1 and bf2) were filtered as follows: both COD, bf1 ≥ 3 and bf2 ≥ 3.5; DMA, bf1 ≥ 3.3 and bf2 ≥ 3.

Once selected based on their Bayes factors, SNPs were evaluated with SnpEff [83] to identify those markers located in the coding sequences (exons), regulatory regions (5′/3′ UTRs), or introns. We selected as domestication candidates those genes that contained at least two SNPs with high association factors to any domestication phenotype, and were affecting regulatory regions, had non-synonymous effects on the coding sequence, or altered splicing sites or stop codons.

Gene Ontology enrichments

The functional description of genes falling within introgressed genomic windows and with domestication haplotypes was analyzed [31]. Gene Ontology enrichments for each case were performed using the topGO package implemented in Bioconductor [84], using the classic Fisher's exact test with a maximum *p* value of 0.05.

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 14 of 17

Sample preparation and extraction for metabolomic profiling

Young trifoliate leaves from *P. vulgaris*, *P. pseudovulgaris*, and *P. coccineus* were collected and immediately frozen in liquid nitrogen. The leaves were then lyophilized and finely ground (<300 μm) using a Mixer Mill MM 400 (Retsch®). Subsequently, extracts were prepared mixing 50 mg plant powder in 1000 μL methanol and formic acid solution (75% *v/v* and 0.15% *v/v*, respectively). The mixture was sonicated for 15 min in a water bath at maximum frequency and centrifuged at 10,000 *g* for 10 min at 4 °C. The supernatant was filtered through a 0.22-μm filter before analysis by DIESI-MS. All samples were prepared by triplicate and analyzed immediately.

Mass spectrometry

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

Non-targeted metabolite profiling

For non-targeted metabolite profiling, samples previously extracted with methanol were reconstituted in a mixture of methanol/de-ionized water/formic acid (75:24.85:0.15 [*v/v/v*]) and filtered through a 0.2-μm filter. Chromatographic separation was achieved on an Acquity UPLC System (Waters, Milford) using a BEH C18 2.1 × 50 mm, 1.7-um column maintained at 40 °C. Samples were injected (10 μL) and elution of compounds was performed at a flow rate of 0.5 mL/min as follows: mobile phase A: de-ionized water containing 0.1% formic acid; mobile phase B: acetonitrile containing 0.15% formic acid. The gradient program was isocratic for the first 30 s, then a linear gradient increase to 30% of solvent B at 2 min, 40% of B at 4 min, 40% of B at 6 min, and 70% of B at 10 min, with 1 min for column washing and 4 min for column re-equilibration. The mass spectrometer comprised an orthogonal QTOF Synapt G1 (Waters, UK) operated under the following conditions: electrospray ionization in positive mode, capillary voltage at 3.0 kV, cone voltage 46 V, extractor voltage 4.0 V, source and desolvation temperature were 120 and 300 °C, respectively. Cone and desolvation gas flow was nitrogen at a flow rate of 20 L/h and 800 L/

min. Leucine-enkephaline (M + H) + = 556.2771 was infused at a flow rate of 5 μL/min at concentration of 2 ng/mL during acquisition as internal mass calibrant to correct for mass drift.

Metabolomic data analysis

Prior to data analysis, the.*raw* native data format for each spectrum were transformed to standard mass spectrometry.*mzML* format employing msconvert [85]. The spectrum data were then processed using a workflow designed in R (http://www.rproject.org) with the package MALDIquant [86] as follows:.*mzML* data import, summarizing all scans of each sample, smoothing by a Savitzky–Golay filter, and peak alignment/detection for comparison of peaks across different spectra. In total, 318 high quality intensity values of ions were used for statistical analysis. We employed a hierarchical clustering analysis (HCA) approach for the generation of metabolic heatmaps to evaluate the differences in the fingerprinting data. To find the most important ions, we generated a Random Forest Tree model for classification in the R package "Rattle" [87]. LC-MS/MS data were analyzed using MS-672DIAL software v2.06 [88]. Peak annotation was performed comparing fragment mass spectra with MassBank, ReSpect ESI, and MS/MS libraries in positive ion mode.

Additional files

Additional file 1: Supplementary figures **S1–S4** and **S11–S21**. (PDF 5628 kb)

Additional file 2: Tables S1–S3. Table S12 and **Table S20**. (PDF 175 kb)

Additional file 3: Supplementary text, **Tables S4–S11** and **Figures S5–S10**. Coalescent simulations. (PDF 693 kb)

Additional file 4: Table S13. Domestication gene candidates common to 2CODs. (XLSX 92 kb)

Additional file 5: Table S14. Domestication gene candidates specific to Mesoamerican varieties. (XLSX 100 kb)

Additional file 6: Table S15. PCGs and functional categories introgressing between wild subpopulations. (XLSX 109 kb)

Additional file 7: Table S16. PCGs and functional categories introgressing from wild into domesticated subpopulations. (XLSX 99 kb)

Additional file 8: Table S17. PCGs and functional categories introgressing from domesticated into wild subpopulations. (XLSX 149 kb)

Additional file 9: Table S18. GO terms associated to domestication PCGs common to 2CODs. (XLSX 52 kb)

Additional file 10: Table S19. GO terms associated to domestication PCGs specific to Mesoamerican varieties. (XLSX 56 kb)

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 15 of 17

Availability of data and materials

Raw sequence reads were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), primary accession numbers PRJNA349045 (SRS1938383 *P. vulgaris* G23582; SRS1938423 *P. vulgaris* G21245; SRS1938488 *P. vulgaris* G23724; SRS1938489 *P. vulgaris* G23587; SRS1938492 *P. vulgaris* G23939; SRS1938493 *P. vulgaris* G19901; SRS1938369 *P. polystachios*; SRS1938370 *P. hintonii*; SRS1938371 *P. leptostachyus*; SRS1938372 *P. maculatus*; SRS1938373 *P. microcarpus*; SRS1938374 *P. lunatus*; SRS1938375 *P. acutifolius*; SRS1938376 *P. filifromis*; SRS1938377 *P. coccineus*; SRS1938378 *P. dumosus* G36043; SRS1938379 *P. costaricensis* G40811A) and PRJNA221782 (SRS605147 *P. vulgaris* Faba Andecha; SRS605153 *P. vulgaris* G23556; SRS605149 *P. vulgaris* G24392; SRS605152 *P. vulgaris* G23463; SRS605154 *P. vulgaris* G24377; SRS605155 *P. vulgaris* G50368; SRS605156 *P. vulgaris* G23551; SRS605150 *P. vulgaris* G12967; SRS605157 *P. vulgaris* G23550; SRS605151 *P. vulgaris* G24594; SRS605148 *P. vulgaris* G21244; SRX362852 *P. vulgaris* Jalo EEP558; SRS620372 *P. coccineus*). The BAT93 genome can be accessed and browsed at http://denovo.cnag.cat/genomes/bean.

Authors' contributions

AH-E conceived and led the study. LD, RW, TG, RG, JJO-O, and AD-S suggested strategies; MR-A and AH-E wrote the paper with significant contributions by AD-S and PG; MR-A, JMM-V, SS-A, generated and collected data; MR-A, JMM-V, AV, and SC-G analyzed data. PG, RV-B, MS, and MA provided materials. RV-B, MS, MA, RG, and AH-E coordinated teams in the participating countries. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Laboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del IPN (Cinvestav), 36821 Irapuato, Guanajuato, Mexico. [2]Departamento de Biotecnología y Bioquímica, Unidad Irapuato, Cinvestav, 36821 Irapuato, Guanajuato, Mexico. [3]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain. [4]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, 08003 Barcelona, Spain. [5]Instituto de Biotecnología y Biología Molecular (IBBM), UNLP-CONICET, 1900 La Plata, Argentina. [6]EMBRAPA Rice and Beans, Biotechnology Laboratory, Santo Antônio de Goiás, GO 75375-000, Brazil. [7]Mision Biológica de Galicia (MBG)-National Spanish Research Council (CSIC), 36080 Pontevedra, Spain. [8]Departamento de Ingeniería Genética, Unidad Irapuato, Cinvestav, Irapuato, Guanajuato, Mexico. [9]Department of Plant Sciences, University of California, Davis, CA 95616-8780, USA. [10]Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico.

References

1. Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci U S A. 2014;111:6139–46.
2. Gepts P. The contribution of genetic and genomic approaches to plant domestication studies. Curr Opin Plant Biol. 2014;18:51–9.
3. Kwak M, Toro O, Debouck D, Gepts P. Multiple origins of the determinate growth habit in domesticated common bean (*Phaseolus vulgaris* L.). Ann Bot. 2012;110:1573–80.
4. Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, Gopalakrishnan S, Ross-Ibarra J, Gilbert MT, et al. Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. Curr Biol. 2016;26:3195–201.
5. Xue S, Bradbury PJ, Casstevens T, Holland JB. Genetic architecture of domestication-related traits in maize. Genetics. 2016;204:99–113.
6. Dubcovsky J, Dvorak J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science. 2007;316:1862–6.
7. Sang T, Ge S. Understanding rice domestication and implications for cultivar improvement. Curr Opin Plant Biol. 2013;16:139–46.
8. Gornicki P, Zhu H, Wang J, Challa GS, Zhang Z, Gill BS, et al. The chloroplast view of the evolution of polyploid wheat. New Phytol. 2014;204:704–14.
9. Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat Genet. 2016;48:1083–8.
10. Lenser T, Theißen G. Molecular mechanisms involved in convergent crop domestication. Trends Plant Sci. 2013;18:704–14.
11. Delgado-Salinas A, Bibler R, Lavin M. Phylogeny of the genus *Phaseolus* (Leguminosae): a recent diversification in an ancient landscape. Syst Bot. 2006;31:779–91.
12. Kami J, Becerra Velásquez V, Debouck DG, Gepts P. Identification of presumed ancestral DNA sequences of phaseolin in *Phaseolus vulgaris*. Proc Natl Acad Sci U S A. 1995;92:1101–4.
13. Kwak M, Gepts P. Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet. 2009;118:979–92.
14. Bitocchi E, Nanni L, Bellucci E, Rossi M, Giardini A, Zeuli PS, et al. Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Proc Natl Acad Sci U S A. 2012;109:E788–96.
15. McClean P, Gepts P, Kami J. Genomics and genetic diversity in common bean. In: Wilson RF, Stalker HT, Brummer EC, editors. Legume crop genomics. Champaign: AOCS Press; 2004. p. 61–82.
16. Cortés AJ, Chavarro MC, Blair MW. SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet. 2011;123:827–45.
17. Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, et al. Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytol. 2013;197:300–13.
18. Kaplan L, Lynch TF, Smith Jr CE. Early cultivated beans (*Phaseolus vulgaris*) from an Intermontane Peruvian Valley. Science. 1973;179:76–7.
19. Gepts P, Bliss FA. F1 hybrid weakness in the common bean: differential geographic origin suggests two gene pools in cultivated bean germplasm. J Hered. 1985;76:447–50.
20. Gepts P. Origin and evolution of common bean: past events and recent trends. Hortic Sci. 1998;33:1124–30.
21. Kaplan L, Lynch TF. *Phaseolus* (Fabaceae) in archaeology: AMS radiocarbon dates and their significance for pre-Colombian agriculture. Econ Bot. 1999;53:261–72.
22. Beebe S, Toro O, González AV, Chacón MI, Debouck DG. Wild-weed-crop complexes of common bean (*Phaseolus vulgaris* L., Fabaceae) in the Andes of Peru and Colombia, and their implications for conservation and breeding. Genet Resour Crop Ev. 1997;44:73–91.
23. Payró de la Cruz E, Gepts P, Colunga GarcíaMarín P, Zizumbo-Villareal D. Spatial distribution of genetic diversity in wild populations of *Phaseolus vulgaris* L. from Guanajuato and Michoacán, México. Genet Res Crop Ev. 2005;52:589–99.
24. Zizumbo-Villarreal D, Colunga GarcíaMarín P, Payró de la Cruz E, Delgado-Valerio P, Gepts P. Population structure and evolutionary dynamics of

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 16 of 17

wild-weedy-domesticated complexes of common bean in a Mesoamerican region. Crop Sci. 2005;45:1073–83.

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

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

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

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

29. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The genomic signature of crop-wild introgression in maize. PLoS Genet. 2013;9:e1003477.

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

31. Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, Hernández-Oñate M, Minoche AE, Erb I, et al. Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biol. 2016;17:32.

32. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46:707–13.

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

34. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.

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

36. Mina-Vargas AM, McKeown PC, Flanagan NS, Debouck DG, Kilian A, Hodkinson TR, et al. Origin of year-long bean (Phaseolus dumosus Macfady, Fabaceae) from reticulated hybridization events between multiple Phaseolus species. Ann Bot. 2016;118:957–69.

37. Blanco-Pastor JL, Vargas P, Pfeil BE. Coalescent simulations reveal hybridization and incomplete lineage sorting in Mediterranean Linaria. PLoS One. 2012;7:e39089.

38. Huang DI, Hefer CA, Kolosova N, Douglas CJ, Cronk QC. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, Populus balsamifera and P. trichocarpa (Salicaceae). New Phytol. 2014;204:693–703.

39. Ruhsam M, Rai HS, Mathews S, Ross TG, Graham SW, Raubeson LA, et al. Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in Araucaria? Mol Ecol Resour. 2015;15:1067–78.

40. Montero-Vargas JM, González-González LH, Gálvez-Ponce E, Ramírez-Chávez E, Molina-Torres J, Chagolla A, et al. Metabolic phenotyping for the classification of coffee trees and the exploration of selection markers. Mol Biosyst. 2013;9:693–9.

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

42. Winkler R. An evolving computational platform for biological mass spectrometry: workflows, statistics and data mining with MASSyPup64. PeerJ. 2015;3:e1401.

43. Reddy P, Rendón-Anaya M, Soto del Rio MD, Khandual S. Flavonoids as signaling molecules and regulators of root nodule development. Dyn Soil Dyn Plant. 2007;1:83–94.

44. Wells WC, Isom WH, Waines JG. Outcrossing rates of six common bean lines. Crop Sci. 1988;28:177–8.

45. Ferreira J, de Souza Carneiro JE, Teixeira AL, de Lanes FF, Cecon PR, Borém A. Gene flow in common bean (*Phaseolus vulgaris* L.). Euphytica. 2007;153:165–70.

46. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328:710–22.

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

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

49. SAGARPA. Estudio de gran visión y factibilidad económica y financiera para el desarrollo de infraestructura de almacenamiento y distribución de granos y oleaginosas para el mediano y largo plazo a nivel nacional. 2014. http://www.sagarpa.gob.mx/agronegocios/Documents/Estudios_promercado/GRANOS.pdf.

50. Freyre R, Ríos R, Guzmán L, Debouck DG, Gepts P. Ecogeographic distribution of *Phaseolus* spp. (Fabaceae) in Bolivia. Econ Bot. 1996;50:195–215.

51. Xu H, Guan Y. Detecting local haplotype sharing and haplotype association. Genetics. 2014;197:823–38.

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

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

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

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

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

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

58. Viana DS, Santamaría L, Figuerola J. Migratory birds as global dispersal vectors. Trends Ecol Evol. 2016;31:763–75.

59. La Sorte FA, Fink D, Hochachka WM, Kelling S. Convergence of broad-scale migration strategies in terrestrial birds. Proc R Soc B. 2016;283:20152588.

60. Bacon CD, Silvestro D, Jaramillo C, Smith BT, Chakrabarty P, Antonelli A. Biological evidence supports an early and complex emergence of the Isthmus of Panama. Proc Natl Acad Sci U S A. 2015;112:6110–5.

61. Hain MP, Sigman DM, Haug GH. The biological pump in the past. In: Holland HD, Turekian KK, editors. Treatise on Geochemistry, Edition 2, Chapter: 8.18. Amsterdam, NL: Elsevier; 2014. p. 485–517.

62. Hungria M, Johnston AW, Phillips DA. Effects of flavonoids released naturally from bean (*Phaseolus vulgaris*) on nodD-regulated gene transcription in *Rhizobium leguminosarum* bv. *phaseoli*. Mol Plant Microbe In. 1992;5:199–203.

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

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

65. Ng JL, Hassan S, Truong TT, Hocart CH, Laffont C, Frugier F, et al. Flavonoids and auxin transport inhibitors rescue symbiotic Nnodulation in the *Medicago truncatula* cytokinin perception mutant cre1. Plant Cell. 2015;27:2210–26.

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

67. Aguilar OM, Riva O, Peltzer E. Analysis of *Rhizobium etli* and of its symbiosis with wild *Phaseolus vulgaris* supports coevolution in centers of host diversification. Proc Natl Acad Sci U S A. 2004;101:13548–53.

68. Ribeiro RA, Ormeño-Orrillo E, Dall'Agnol RF, Graham PH, Martinez-Romero E, Hungria M. Novel *Rhizobium* lineages isolated from root nodules of the common bean (*Phaseolus vulgaris* L.) in Andean and Mesoamerican areas. Res Microbiol. 2013;164:740–8.

69. Servín-Garcidueñas LE, Zayas-Del Moral A, Ormeño-Orrillo E, Rogel MA, Delgado-Salinas A, Sánchez F, et al. Symbiont shift towards *Rhizobium* nodulation in a group of phylogenetically related Phaseolus species. Mol Phylogenet Evol. 2014;79:1–11.

70. Debouck DG, Castillo TR, Tohme JM. Observations on little-known *Phaseolus* germplasm of Ecuador. Plant Genet Resour Newsl. 1989;80:15–21.

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

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

73. Hannah MA, Krämer KM, Geffroy V, Kopka J, Blair MW, Erban A, et al. Hybrid weakness controlled by the dosage-dependent lethal (DL) gene system in common bean (*Phaseolus vulgaris*) is caused by a shoot-derived inhibitory signal leading to salicylic acid-associated root death. New Phytol. 2007;176:537–49.

Rendón-Anaya *et al. Genome Biology* (2017) 18:60

Page 17 of 17

74. Gepts P, Papa R. Evolution during domestication, Encyclopedia of Life Sciences. London: Macmillan Publishers, Nature Publishing Group; 2002.
75. Koinange EMK, Singh SP, Gepts P. Genetic control of the domestication syndrome in common-bean. Crop Sci. 1996;36:1037–45.
76. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754–60.
77. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, et al. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol. 2008;148:1772–81.
78. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15:356.
79. Guo X, Castillo-Ramírez S, González V, Bustos P, Fernández-Vázquez JL, Santamaría RI, et al. Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts. BMC Genomics. 2007;8:228.
80. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
81. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73.
82. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007;81:559–75.
83. Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
84. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0. 2016.
85. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012;30:918–20.
86. Gibb S, Korbinian S. MALDIquant: A versatile R package for the analysis of mass spectrometry data. Bioinformatics. 2012;28:2270–1.
87. Williams G. Data Mining with Rattle and R: The Art of Excavating Data for 1282 Knowledge Discovery. New York: Springer; 2011.
88. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015;12:523–6.