Title

A reference genome for common bean and genome-wide analysis of dual domestications.

Journal

Nature genetics, 46(7)

ISSN

1061-4036

Authors

Schmutz, Jeremy
McClean, Phillip E
Mamidi, Sujan
et al.

Publication Date

2014-07-01

DOI

10.1038/ng.3008

Peer reviewed

A reference genome for common bean and genome-wide analysis of dual domestications.

Jeremy Schmutz[1,2†*], Phillip McClean[3†*], Sujan Mamidi[3], G. Albert Wu[1,], Steven B. Cannon[4], Jane Grimwood[2], Jerry Jenkins[2], Shengqiang Shu[1], Qijian Song[5], Carolina Chavarro[6], Mirayda Torres-Torres[6], Valerie Geffroy[7,15], Samira Mafi Moghaddam[3], Dongying Gao[6], Brian Abernathy[6], Kerrie Barry[1], Matthew Blair[8], Mark A. Brick[9], Mansi Chovatia[1], Paul Gepts[10] , David M Goodstein[1], Michael Gonzales[6], Uffe Hellsten[1], David L. Hyten[5,^], Gaofeng Jia[5], James D. Kelly[11], Dave Kudrna[12], Rian Lee[3], Manon M.S. Richard[7], Phillip N. Miklas[13] , Juan M. Osorno[3], Josiane Rodrigues[5,^^], Vincent Thareau[7], Carlos A. Urrea[14], Mei Wang[1], Yeisoo Yu[12], Ming Zhang[1], Rod A. Wing[12], Perry B. Cregan[5], Daniel S. Rokhsar[1], Scott A. Jackson[6*]

[1] US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA.
[2] HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA.
[3] North Dakota State University, Department of Plant Sciences, Fargo, North Dakota 58102, USA.
[4] United States Department of Agriculture-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, Iowa 50011, USA.
[5] Soybean Genomics and Improvement Laboratory, United States Department of Agriculture, Agricultural Research Service, Beltsville, Maryland 20705, USA.
[6] Center for Applied Genetic Technologies, University of Georgia, Athens, Georgia 30602, USA.
[7] Institut de Biologie des Plantes (IBP), UMR-CNRS 8618, Université Paris Sud, Bâtiment 630, Saclay Plant Sciences, 91405 Orsay cedex, France.
[8] Tennessee State University, Department of Agricultural and Natural Sciences, Nashville, Tennessee 37209, USA.
[9] Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado, 80523, USA.
[10] Department of Plant Sciences/MS1, University of California, Davis, California 95616, USA.
[11] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan 48824, USA.
[12] BIO5 Institute, University of Arizona, Tucson, Arizona 85721, USA.
[13] Vegetable and Forage Crop Research Unit, USDA-ARS, Prosser, WA 99350, USA.
[14] University of Nebraska, Panhandle Research & Extension Center, Scottsbluff, Nebraska 69361, USA.
[15] Unité Mixte de Recherche de Génétique Végétale, Institut National de la Recherche Agronomique, IDEEV, 91190 Gif-sur-Yvette, France

[^] Present address: Pioneer Hi-Bred International Inc, Johnston, Iowa, USA.
[^^] Present address: Federal University of Viçosa, Viçosa, Brazil.

[†]Shared first authorship
[*]Corresponding authors

Abstract

Common bean (*Phaseolus vulgaris* L.) is the single most important grain legume for human consumption and, due to its ability to fix atmospheric nitrogen via symbioses with soil-borne microorganisms, has a valuable place in sustainable agriculture. We assembled 473 Mb of the 587 Mb common bean genome and genetically anchored 98% of this sequence in 11 chromosome-scale pseudomolecules. We compared the common bean genome against its most economically important relative, soybean, to examine changes in soybean that occurred after its recent whole genome duplication. Using resequencing of 60 wild individuals and 100 landraces from the two, genetically differentiated Mesoamerican and Andean gene pools, we confirmed that common bean underwent two independent

domestications, starting from genetic pools that had diverged prior to human colonization. We found that less than 10% of 74 Mb putatively involved in domestication was shared between the two domestication events which included 1,835 genes for the Mesoamerican and 748 for the Andean gene pools. We used sequence diversity and differentiation estimates to identify a subset of genes linked and associated pathways with an increase in leaf and seed size and combined these results with QTL data from a multi-site association mapping trial of modern Mesomerican cultivars to identify genes that likely contribute to the large seed size of modern, cultivated common bean. Finally, we identified a set of genes affected by domestication that may serve as targets for genomics-enabled crop improvement.

Introduction

Common bean (*Phaseolus vulgaris* L.) is a crop of major societal importance and an important source of protein and essential nutrients. Worldwide, common bean is the most consumed legume, providing up to 15% of total daily calories and 36% of total daily protein in parts of Africa and the Americas (http://fao.faostat.org). More than 200 million people in sub-Saharan Africa depend on common bean as a primary staple. It has many health-beneficial[1,2] nutrients whose concentrations are heritable[3] and increasing the concentrations of these nutrients is a breeding objective worldwide[4].

Multiple lines of evidence have shown that wild common bean is organized in two geographically isolated and genetically differentiated wild gene pools (Mesoamerican (MA) and Andean) that diverged from a common ancestral wild population more than 100,000 years ago (ya)[5]. From these wild gene pools, nearly 8,000 ya, common bean was independently domesticated in Mexico and in South America[6-9] followed by local adaptations resulting in landraces with distinct characteristics. In Mexico, common bean was likely domesticated concurrently with maize as part of the "milpa" cropping system (featuring common bean along with maize and squash), which was adopted throughout the Americas[10]. Domestication led to morphological changes including increased seed and leaf sizes, changes in growth habit and photoperiod responses[11], and seed coat color and pattern variation that distinguish culturally adapted classes of beans[12].

Independent domestications, starting from distinct gene pools from a single species, provide experimental replication not typically found in domestication or evolutionary studies. It is possible to deduce the domestication history at a genome-wide scale and examine the roles of parallel evolution and introgression during the domestication of two independent lineages within a single species. Here, in order to understand the history of these complicated domestication events and their implications for modern bean crop improvement, we report a genome sequence for an Andean ecotype of common bean and an analysis of genetic variation in accessions ranging from Mexico to the species southern range in Argentina. In addition, comparative genomics to soybean (*Glycine max*), a closely related crop, revealed effects of shared and lineage-dependent polyploidies on gene fractionation and recent transposable element expansion in common bean.

Reference genome and analysis

To obtain a high quality reference genome, we sequenced an inbred landrace line of *Phaseolus vulgaris* (G19833) derived from the Andean pool (Race Peru) using a whole genome shotgun sequencing strategy that combined multiple linear libraries (18.6x assembled) and ten paired libraries of varying insert sizes (1.8x assembled) sequenced with

the Roche 454 platform, and 24.1 GB of Illumina-sequenced fragment libraries. For longer-range linkage we also end-sequenced three fosmid libraries and two BAC libraries on the Sanger platform (0.54x long-insert pairs) for a total assembled sequence coverage level of 21.0x (Supplementary Table S1). The resulting assembled sequences were organized into 11 chromosomal pseudomolecules by integration with a dense GoldenGate and Infinium-based SNP map of 7,015 markers typed on 267 F_2 lines from a Stampede x Red Hawk cross and a similar set of Infinium markers and 261 SSRs typed on 88 F_5-derived recombinant inbred lines (RILs) derived from the same cross (Cregan and Song, unpublished). Additional refinements to the pseudomolecules were made based on synteny with soybean (*Glycine max*), where allowed by available map data. Almost all of these changes were made in pericentromeric regions, where recombination is generally too limited to resolve the ordering and orientation of small scaffolds. The pseudomolecules include 468.2 Mb of mapped sequence in 240 scaffolds. The total release includes 472.5 Mb of the ~587 Mb[13] genome with half of the assembled nucleotides in contigs longer than 39.5 kb (i.e., "contig N50") (Supplementary Table S3). To annotate the chromosomal assembly we combined Sanger-derived EST resources and a substantial amount of new RNA-seq reads (727 M reads from 11 tissues and developmental stages; Supplementary Table S4) with homology-based and *de novo* gene prediction approaches. The resulting annotation includes 27,197 protein-coding loci, including 4,491 alternative transcripts (Supplementary Table S5), an underestimate that will increase with additional transcriptomes and analyses. Most, 91%, of these genes are retained in synteny blocks with *G. max* (Supplementary Material S5).

Recent transposable element activity: We identified significant recent transposable element (TE) activity and transposon expansions (Supplementary Figs. S17-19). Although recently diverged repeats could not be assembled directly from Roche 454 pyrosequencing, extensive BAC- and fosmid-end sequence data and a dense genetic map allowed us to position 99.6% of genic sequences, and to link into those genes embedded in regions dense with transposable elements (Supplementary Figs. S1-11). The centromere/pericentromeric regions cover ~54% of the genome and are primarily repetitive but still contain 26.5% of the genes. Similar to other sequenced genomes[14,15], these pericentromeric genomic regions are recombinationally inert. Using a threshold of 2 Mb/cM to identify transitions into pericentromeric regions, the pericentromeres span ~54% of the genome, and have an average of 4,350 kb/cM versus 220 kb/cM in the euchromatic arms (Supplementary Table S7). The pericentromeres are primarily repetitive; but due to their size, they still contain 26.5% of the genes.

The majority of the repetitive elements within the genome were LTR retrotransposons and we identified and classified 2,668 complete LTR retrotransposons into 165 families, including 65 Ty1-copia, 78 Ty3-gypsy and 22 unclassified families (Supplemental Table S8). Although there were ancient elements which inserted into the genome more than 10 million year ago (MYA), ~75% (2,011/2,668) of the LTR retroelements integrated into *Phaseolus* within the last 2 MY (Supplemental Fig. S17). Notably, the insertion times of 20% (543/2,668) of the elements were more recent than 0.5 MYA—this is likely an underestimate, as our sequencing approach would bias against the assembly of completely identical LTRs. These results were similar to that in soybean[15] and suggested LTR retrotransposons exhibited recent amplifications in both legumes. The 165 LTR retrotransposon families varied in copy number of complete elements as more than

78% (130/165) families had fewer than 10 complete retroelements; however, 11 families had more than 50 complete elements and contain 63% (1,690/2,668) of the complete elements in the *Phaseolus* genome. Some families show extremely high copy numbers, *e.g.,* family pvRetroS2 contains 446 complete elements (likely an underestimate as some elements will not have been assembled distinctly).

Dense clusters of resistance genes: The majority of putative plant resistance genes contain nucleotide binding and leucine-rich repeat domains, collectively known as NB-LRR (NL) genes[16]. We identified 376 NL encoding genes, of which 106 contained an N-terminal toll/interleukin 1 receptor (TIR)-like domain (TNLs), and 108 of which contained an N-terminal coiled-coil (CC) domain (CNL) (Supplemental Table S10). The majority of NL sequences are physically organized in complex clusters, often localized at the ends of chromosomes (Supplemental Fig. S20). In particular, three large clusters are localized at the ends of chromosomes Pv04, Pv10, and Pv11 and contain more than 40 NLs that are enriched for CNL (Pv04, Pv11) or TNL (Pv10) genes that co-localize with previously mapped disease resistance genes[17-22]. Local tandem duplications and ectopic recombination between clusters are involved in the evolution of these NL gene clusters[23].

Comparison of genome changes in duplicated sister legume species

Phaseolus (common bean) and *Glycine* (soybean) diverged ~19.2 MYA but share a whole genome duplication (WGD) ~56.5 MYA[24]. The *Glycine* lineage experienced an independent WGD ~10 MYA[15]. These events are evident in plots of synonymous changes in coding sequences (Ks) between and within these genomes (Supplemental Fig. S21), which also show that *Phaseolus* has evolved more rapidly than *Glycine* since their last common ancestor. Assuming a divergence time of ~19.2 MYA[24], the Ks (synonymous substitution) rate for *Phaseolus* is 1.4 times higher than *Glycine* (8.4635×10^{-9} versus 5.8594×10^{-9} substitutions/year).

We identified orthologous *Phaseolus* and *Glycine* genes using synteny and Ks as criteria. Consistent with earlier work, there was extensive synteny between *Phaseolus* and *Glycine*, except in pericentromeric regions where microcolinearity is often stretched out and thinned due to genomic expansion in one or both genomes. Typically, two chromosomal blocks in *Glycine* map to a single region of *Phaseolus* due to the most recent WGD in *Glycine* (Fig. 1)[15,25,26]. Most of the *Phaseolus* genes (91%; 24,861) are in identifiable synteny blocks in *Glycine*; and 57% are in synteny blocks within *Phaseolus* itself—a result of the ancient 55 MYA WGD. Within synteny blocks, the recent *Glycine*:*Glycine* duplication has a mean of 33 genes/block, whereas the older shared *Phaseolus*:*Glycine* WGD event has an average of 14 genes/block.

Evolution of common bean gene pools

Mesoamerica has been suggested to be the center of origin for common bean, ultimately leading to distinct modern wild Andean and Mesoamerican (MA) gene pools[7]. To investigate the differentiation of these wild populations, we performed pooled resequencing of 30 individuals each from MA and Andean wild populations (Fig. 2). The MA wild population (π/bp=0.0061; Θ/bp=0.0041) is more diverse than the Andean wild population (π/bp=0.0014; Θ/bp=0.0013). We used 663,0000 polymorphic sites (at least 5 kb from a gene and not in a repeat) to estimate demographic parameters using the joint allele

frequency spectrum (dadi)[27] (Supplemental Section 6). The strong fixation index F_{ST} ~0.34 between these two wild populations indicates that they show substantial allelic differentiation from each other. We estimate the divergence of the two wild pools occurred ~165 kya, with an ancestral effective population size of 168,000. This date is older than, but within 95% confidence intervals of a previous estimate of ~110 kya, based on 13 loci from 24 wild genotypes[5] but younger than other estimates of ~500 kya[28]. The whole genome analysis resulted in a much tighter confidence interval of 146-184 kya.

Demographic inference for the wild Andean gene pool suggests that it was derived from the wild Mesoamerican population with a founding population of only a few thousand individuals (Fig. 3a, Supplemental Section 6). The wild Andean population shows no appreciable growth in effective population size for ~76 ky after founding, although there was continual asymmetric gene flow between the two wild pools with a higher Mesoamerican-to-Andean migration rate (Supplemental Table S12). The Andean population then underwent an exponential growth phase which lasted from ~90 kya until the present. The strong pre-domestication bottleneck in the Andean population has been observed in previous analyses[7,29,30] in contrast, however, no detectable bottleneck was found for the wild Mesoamerican gene pool.

Domestication of common bean

To characterize diversity and differentiation within and between the Mesoamerican and Andean landraces (early domesticates), we sequenced four pooled populations representing distinct MA landraces, and two pooled populations representing distinct Andean landraces (n=7-26). These represent subpopulations from Mexico, Central America, and South America with low levels of admixture (Supplemental Fig. S24). Since the four MA and two Andean landrace populations are representative of the diversity of the original domestication populations, we combined SNP data from these populations to create a composite MA and a composite Andean landrace SNP data set, respectively, for further analysis. This allowed us to distinguish selection from random fixation across the genome[31,32] and to search for signals associated with the domestication events. The number of SNPs ranged from 8,890,318 for wild MA to 1,397,405 for the Andean landrace subpopulation from Peru (Supplemental Table S14), and ~16% of these SNPs are within genes.

To characterize variation among the populations, we calculated diversity (π) and population differentiation (F_{ST}) statistics using data averaged over 10kb windows with a 2kb slide (10kb/2kb; Supplemental Table S15). While the MA landraces are less diverse than MA wilds, Andean landrace populations are more diverse than the Andean wild population, possibly due to admixture with MA and/or de novo mutation within the Andean gene pool. Diversity is further reduced within the MA Central American and southern Andean landraces, suggesting that these subpopulations underwent additional selection that may correspond to local adaptation.

Multiple results point to independent domestication events in the MA and Andean gene pools, a feature observed for only a few modern crops. We characterized common bean domestication at the genomic level by comparing wild and landrace populations across 10kb/2kb sliding windows and selecting windows that met strict composite criterion: upper 90% of the population's empirical distribution for both the $\pi_{wild}/\pi_{landrace}$ ratio and F_{ST} (Fig. 3b, 4). We observed 930 MA windows (totaling 74Mb) with both low diversity and high differentiation. Since low diversity and high differentiation are two features of selection[33],

we consider these selection windows. Of these, 209 windows longer than 100kb accounted for 70.1% of the total selection distance. Among the 750 Andean selection windows exhibiting low diversity and high differentiation, 172 windows longer than 100kb covered 69.8% of the total selection distance (60 Mb). As expected for independent MA and Andean domestication events, these selection regions were distinct. Within the MA landrace population, Pv02, Pv07, and Pv09 accounted for 43% of the length (32.338 Mb), with 33.3% of chromosome Pv09 showing signatures of selection, whereas the Andean domestication primarily involved Pv01, Pv02, and Pv10 (Fig. 4). Interestingly, only 7.234 Mb of the predicted domestication regions are common between the two gene pools suggesting different genetic routes to domestication.

We identified candidate genes associated with domestication using the same criteria applied to the discovery selection windows (upper 90% of the pool's empirical distribution for both the $\pi_{wild}/\pi_{landrace}$ ratio and F_{ST}). We identified 1,835 MA and 748 Andean domestication candidate genes (Supplemental Table S16 and S17), and all candidates have a negative Tajima's D value indicating positive selection. Most notably, only 59 of the candidate genes (3% of the MA, 8% of the Andean candidates) are common between the two landrace populations. Among the 59 common candidates, the mean F_{ST} was 0.67, suggesting selection on different alleles or the appearance of unique mutations in the two gene pools. This is consistent with the *PvTFL1y* determinancy locus that was independently derived in each gene pool[34], but contrasts with rice where a domestication locus appeared uniquely in one gene pool, indica or japonica and was transferred to the other pools[35]. Most MA candidates (n=1,561; 85%) are located in 10kb selection windows, whereas only 48.1% of the Andean candidates are within such windows. The effects of domestication were uneven across the MA subpopulations: we detected only 418 candidates in the MA Central American landrace population, compared to 1,424 candidates within the MA Mexican landraces. The fact that only 33 of these genes are shared between these two subpopulations indicates unique evolutionary trajectories among subpopulations within the MA gene pools Within the Andean gene pool, none of the candidate genes from the northern and southern Andean landrace populations are shared. These results demonstrate that the sexually compatible MA and Andean lineages with similar morphologies and life cycles underwent independent selection upon distinct sets of genes. This is in contrast to rice, where many major domestication genes were shared by gene flow between the indica and japonica types[36].

Domestication had distinct effects on genes involved in flowering[37] in the two gene pools. While the principal floral integrator genes *SOC1* and *FT*[38] are not candidates in either pool, 25 MA and 13 Andean genes that are in pathways that control these two genes are domestication candidates. For example, within the vernalization pathway, orthologs of *VRN1* (*Phvul.003G033400*) and *VRN2* (*Phvul.002G000500*) are MA candidates and *FRL1* (*Phvul.006G053200*) and *TFL2* (*Phvul.009G117500*) are Andean candidates. *COP1* is a photoperiod pathway regulator that controls *FT* through *CO*. The MA ortholog of *COP1* is a MA candidate while *Phvul.006G165300*, a *CUL4* ortholog that encodes a protein that is part of a complex that along with *COP1* regulates *CO*[39], is an Andean candidate. This demonstrates independent selection upon different members of the same protein complex. The only shared domestication candidates are *Phvul.007065600*, an ortholog of *AGL42*, which regulates flowering through the gibberellin pathway, and *Phvul.009G203400*, an ortholog of *FUL* that regulates *SOC1*.

Increased plant size is typically associated with plant domestication[40], and multiple MA candidates fall into this category. *Phvul.*011G213300 is an ortholog of Arabidopsis *BB*, a component of the ubiquitin ligase degradation pathway that controls flower and stem size[41], while *Phvul.*009G040200 is a *BIN4* ortholog that regulates cell expansion and final plant size[42]. Multiple candidates are also components of nitrogen metabolism,which directly affects plant size. The MA candidate *Phvul.008G168000* encodes nitrate reductase, a critical element for plant and seed growth, which genetically maps to the SW8.2 seed weight QTL[43]. Other nitrogen metabolism candidates include the MA (*Phvul.005G132200*) and Andean (*Phvul.002G242900*) nitrogen transporters, and the MA asparagine synthase (*Phvul.006G069300*).

Increased seed size is a major phenotypic shift associated with domestication of common bean[44] and other legumes[45] and distinguishes the many types of beans that humans consume. We surveyed the MA domestication candidates for genes previously shown to be associated with seed weight[46] and used the whole genome sequence for a genome-wide association analysis (GWAS) to understand the genetic architecture of seed weight in modern MA cultivars. We found 15 candidate genes previously shown to be involved in seed weight (SI Table 19)**.** Among these are nearly all the components of the cytokinin synthesis and multiple-component phosphorelay regulatory system (Supplemental Fig. S24). Included are *Phvul.002G082400*, which encodes a protein that transmits the phospo-signal to response regulators, and three type-B response regulator transcription factors (*Phvul.003G017000*, *Phvul.003G110100*, and *Phvul.009G088900*) which in turn activate a number of downstream genes[47]. An additional candidate, *Phvul.01G038800*, has orthologs that encode cytokinin oxidase/dehydrogenase proteins, which function by regulating the pathway by degrading active cytokinin. The relevance of these as seed weight candidates is supported by work in Arabidopsis where orthologs of these cytokinin pathway candidates have been shown by transgenic studies to regulate seed size/weight[46]. In contrast, however, none of these genes are Andean domestication candidates.

Seed weight GWAS analysis confirmed three of these domestication candidates. It was not possible to confirm the other twelve candidates with the GWAS because the MA domestication reduced diversity to near homozygosity such that associations could not be found. The GWAS was able to place 75 domestication candidates within 50kb of a significant (P<1.0E-04) seed weight SNP, and a significant SNP was found within eight candidates (Supplemental Table 21). One sweep window on Pv07 (9.662-10.662 Mb) contains 33 candidates, and is located in a GWAS peak that exhibited extensive linkage disequilibrium (Fig. 5b). Using GWAS, we also detected seed weight candidate genes that resulted from modern breeding of common bean. These included 15 improvement-related genes previously shown to be associated with seed weight, five of which function in the cytokinin regulation/degradation pathway (Supplemental Table S22). Finally, three genes in complete linkage disequilibrium and equally significant (P=6.3E-06), are located in a Pv07 seed weight QTL that has been replicated in many experiments[48].

Impact on society and agriculture

Common bean is the most important grain legume for human consumption and is an especially important nutrient dense food in developing parts of the world. Improvement of common bean will require a more fundamental understanding of the genetic basis of how it responds to biotic and abiotic stresses. The clustering of resistance genes in a few genomic

locations suggests that stacking resistances between clusters should be relatively easy but stacking multiple resistances located within a single physical cluster, and then combining these traits by breeding, may prove more challenging. The observation that the dual domestications of common bean share few common selective sweeps leads us to posit that domestication, previously thought to be typically associated with selection at a few major loci, can also be derived via multiple genetic pathways effecting similar/same phenotypes (*e.g.*, seed size). In addition, the lack of correspondence between domestication selective sweeps and genetic bottlenecks imposed by breeding indicates that domestication traits were fixed early and that subsequent selection was likely on traits for local adaptation and desired seed/plant traits. Together, these findings provide information on regions of the genome that have been intensely selected either during domestication or early improvement and thus provide targets for future crop improvement efforts, as valuable alleles will have been lost during early selection.

Methods summary:

Sequencing. The majority of the *de novo* genome sequencing reads was collected with standard sequencing protocols for Roche 454 XLR and Illumina HiSeq2000. Fosmid and BAC End sequences were collected using standard protocols on ABI 3730XL capillary sequencing. 160 *P. vulgaris* race and wild genotypes were combined in 8 pools, unamplified libraries made using standard JGI protocols and data collected on Hiseq2000.

Construction of genetic map. A set of 992,682 SNPs were obtained from sequencing diverse common bean genotypes and filtering the reads against an early *P. vulgaris* assembly. These were typed on an Illumina Infinium BeadChip (5,232 SNPS) on 267 F_2 progeny from the cross Stampede x Red Hawk. From the *P. vulgaris* assembly V0.9, an additional BeadChip (5,514 SNPS) was designed, and the same population was typed. 261 simple sequence repeat (SSR) markers and both BeadChips were used to type 88 F_5-derived recombinant inbred lines (RILs) from the same cross. The final linkage map constructed with JoinMap 4.0[49] software contains 7,015 SNP (6,535 beadchip and 484 Illumina GoldenGate) and SSR markers arranged in 11 *P. vulgaris* linkage groups designated by 25 framework markers.

Assembly. A total of 49,412,786 sequence reads (21.02x assembled sequence coverage) were assembled using our modified version of Arachne[50] v.20071016 to form 1,627 scaffolds (42,447 contigs) totaling 474.3 Mb of sequence. A total of 71 misjoins were identified in the initial assembly and were broken. The combination of the genetic map described here was used to identify 248 joins to form the 11 chromosomes. The remaining scaffolds were screened for contamination to produce a final assembly consisting of 708 scaffolds (31,391contigs) containing 472.5 MB of the *P.vulgaris* genome.

Annotation. PERTRAN (S. Shu) was used to construct RNA-seq transcript assemblies from about 727 million reads of G19833 paired-end Illumina RNA-seq reads. PASA[51] was used to build transcript assemblies from 79,630 available ESTs (see SI Table 4) and 43,627 RNA-seq transcript assemblies. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of peptides from other plant genomes to a repeat-soft-masked genome using RepeatMasker[52]. Gene models were predicted by homology-based predictors, FGENESH+[53], FGENESH_EST (similar to FGENESH+, EST as splice site and intron input instead of peptide/translated ORF), and GenomeScan[54]. The best-scored predictions for each locus were selected using multiple positive factors including EST and peptide support, and one negative factor: overlap with repeats. PASA-improved gene models were filtered based on peptide homology or on EST evidence with Pfam TE domain models removed. The final gene set has 27,197 protein coding genes and 31,638 protein coding transcripts.

Comparisons to other legumes
Synteny blocks within and between the *Phaseolus* and *Glycine* genomes were identified with DAGchainer [55], based on chromosomal locations of genic anchor points determined using the NCBI blastp program (E-value <= 1e-10), filtered to the top reciprocal best matches per chromosome pair. The Ks for gene pairs were calculated, using in-frame CDS alignments, and Ks calculated using the codeml from PAML, version 4.4[56].

Developing race pools
126 wild and 179 landrace *P. vulgaris* genotypes were screened with indel markers[57] and analyzed using STRUCTURE[58]. Two subpopulations best fit the wild genotype data, while the landraces were represented by six subpopulations. Based on the results, 30 individuals each were pooled to create the wild Mesoamerican and wild Andean sequencing pools. Three of the six landrace pool populations were of Mexican origin (Mx1, n=25; Mx2, n=7; Mx3, n=16), one was of Central American origin (CA, n=26), and two were of Andean origin (S Andes, n=9; N Andes, n=17). Sequence data among selected pools were combined to represent the following populations: MA landraces (Mx1, Mx2, Mx3, CA), Andean landraces (S Andes, N Andes), and Mexican landraces (Mx1, Mx2, Mx3).

Identifying selection windows and domestication genes
For all individual and composite pools, the number of SNPs, nucleotide diversity [π[59]; θ_w[60]], and Tajima's D[61] were calculated for each 10kb/2kb sliding window and for each gene. Differentiation between pooled populations was measured using F_{ST}. Windows or genes were considered under selection during domestication if both the $\pi_{wild}/\pi_{landrace}$ ratio and F_{ST} statistics were in the upper 90% of their empirical distributions.

Association mapping
280 modern Mesoamerican varieties were grown in replicated trials in four locations (ND, MI, NE, CO) in the United States and genotyped with 34,799 SNPs. SNP loci associated with seed weight were discovered using a mixed linear model accounting for kinship using an identity-by-state estimate and population structure using principal components.

Acknowledgements:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research was funded by the National Science Foundation (DBI 0822258), USDA-NIFA 2006-35300-17266 to SAJ, and USDA-CSREES (2009-01860) and (2009-01929) to SAJ and PM, respectively.

Data deposition:

Assembly and annotation is available from http://www.phytozome.net/commonbean.php and is deposited in Genbank under accession ANNZ00000000.

Author contributions:
J.S., P.M., D.R. and S.A.J. conceived the study and jointly wrote the paper with S.B.C. Genomic clones and DNAs were provided by R.A.W, Y.Y., D.K, R.L. and M.B. The following analyses and authors are as follows: repeat annotation, D.G.; resistance genes, V.G., M.M.S.R. and V.T.; genetic mapping, P.B.C., Q.S., J.R., D.L.H. and G.F.; sequencing/assembly/annotation, J.G., J.J., S.S., K.B., M.C., D.M.G., U.H., M.W. and M.Z.; comparative/population/evolutionary analyses, S.M., G.A.W., S.B.C., C.C., S.M.M., B.A. and M.G.; and GWAS, S.M.M., M.A.B., P.G., J.K., P.N.M., J.M.O. and C.A.U.

Author information:

Correspondence and requests for materials should be addressed to sjackson@uga.edu, jschmutz@hudsonalpha.org and phillip.mcclean@ndsu.edu.

References

1 Anderson, J. W. *et al.* Hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men. *The American journal of clinical nutrition* **40**, 1146-1155 (1984).
2 Geil, P. & Anderson, J. Nutrition and health implications of dry beans: a review. *Journal of the American College of Nutrition* **13**, 549-558 (1994).
3 Cichy, K. A., Caldas, G. V., Snapp, S. S. & Blair, M. W. QTL analysis of seed iron, zinc, and phosphorus levels in an Andean bean population. *Crop science* **49**, 1742-1750 (2009).
4 Beebe, S. Common bean breeding in the tropics. *Plant Breed Rev* **36**, 357-426 (2012).
5 Mamidi, S. *et al.* Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. *Heredity* **111**, 267-276 (2013).
6 Bitocchi, E. *et al.* Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytologist* **197**, 300-313 (2013).
7 Bitocchi, E. *et al.* Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proceedings of the National Academy of Sciences* **109**, E788-E796 (2012).
8 Gepts, P., Osborn, T., Rashka, K. & Bliss, F. Phaseolin-protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): evidence for multiple centers of domestication. *Economic botany* **40**, 451-468 (1986).
9 Mamidi, S. *et al.* Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Functional Plant Biology* **38**, 953-967 (2011).
10 Zizumbo-Villarreal, D. & Colunga-GarcíaMarín, P. Origin of agriculture and plant domestication in West Mesoamerica. *Genetic Resources and Crop Evolution* **57**, 813-825 (2010).
11 Singh, S. P., Gepts, P. & Debouck, D. G. Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Economic Botany* **45**, 379-396 (1991).
12 McClean, P., Lee, R., Otto, C., Gepts, P. & Bassett, M. Molecular and phenotypic mapping of genes controlling seed coat pattern and color in common bean (*Phaseolus vulgaris* L.). *Journal of Heredity* **93**, 148-152 (2002).
13 Bennett, M. & Leitch, I. *Plant DNA C-values database (release 6.0, Dec. 2012)* <http://www.kew.org/cvalues/> (2012).

14 Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
15 Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).
16 Meyers, B. C., Kaushik, S. & Nandety, R. S. Evolving disease resistance genes. *Current opinion in plant biology* **8**, 129-134 (2005).
17 Geffroy, V. *et al.* Molecular analysis of a large subtelomeric nucleotide-binding-site–leucine-rich-repeat family in two representative genotypes of the major gene pools of *Phaseolus vulgaris*. *Genetics* **181**, 405-419 (2009).
18 Geffroy, V. *et al.* Identification of an ancestral resistance gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen Colletotrichum lindemuthianum. *Molecular plant-microbe interactions* **12**, 774-784 (1999).
19 Innes, R. W. *et al.* Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiology* **148**, 1740-1759 (2008).
20 Chen, N. W. G. *et al.* Specific resistances against Pseudomonas syringae effectors AvrB and AvrRpm1 have evolved differently in common bean (*Phaseolus vulgaris*), soybean (*Glycine max*), and *Arabidopsis thaliana*. *New Phytologist* **187**, 941-956 (2010).
21 Geffroy, V. *et al.* A family of LRR sequences in the vicinity of the Co-2 locus for anthracnose resistance in *Phaseolus vulgaris* and its potential use in marker-assisted selection. *Theoretical and Applied Genetics* **96**, 494-502 (1998).
22 Miklas, P. N., Kelly, J. D., Beebe, S. E. & Blair, M. W. Common bean breeding for resistance against biotic and abiotic stresses: From classical to MAS breeding. *Euphytica* **147**, 105-131 (2006).
23 David, P. *et al.* A Nomadic Subtelomeric Disease Resistance Gene Cluster in Common Bean. *Plant Physiology* **151**, 1048-1065, doi:10.1104/pp.109.142109 (2009).
24 Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Systematic Biology* **54**, 575-594 (2005).
25 Gill, N. *et al.* Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant physiology* **151**, 1167-1174 (2009).
26 McClean, P. E., Mamidi, S., McConnell, M., Chikara, S. & Lee, R. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC genomics* **11**, 184 (2010).
27 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet* **5**, e1000695, doi:10.1371/journal.pgen.1000695 (2009).
28 Chacon, S. M. I., Pickersgill, B. & Debouck, D. G. Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theoretical and applied genetics* **110**, 432-444 (2005).
29 Kwak, M. & Gepts, P. Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). *Theoretical and applied genetics* **118**, 979-992 (2009).

30 Rossi, M. *et al.* Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. *Evolutionary Applications* **2**, 504-522 (2009).
31 Rubin, C.-J. *et al.* Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences* **109**, 19529-19536 (2012).
32 Rubin, C.-J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-591 (2010).
33 Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat Genet* **44**, 808-811 (2012).
34 Geffroy, V. *et al.* Identification of an ancestral resistance gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen Colletotrichum lindemuthianum. *Molecular plant-microbe interactions : MPMI* **12**, 774-784 (1999).
35 Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309-1321 (2006).
36 Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497-501 (2012).
37 Fornara, F., de Montaigu, A. & Coupland, G. SnapShot: Control of flowering in Arabidopsis. *Cell* **141**, 550, 550 e551-552 (2010).
38 Baurle, I. & Dean, C. The timing of developmental transitions in plants. *Cell* **125**, 655-664 (2006).
39 Chen, H. *et al.* Arabidopsis CULLIN4-damaged DNA binding protein 1 interacts with CONSTITUTIVELY PHOTOMORPHOGENIC1-SUPPRESSOR OF PHYA complexes to regulate photomorphogenesis and flowering time. *Plant Cell* **22**, 108-123 (2010).
40 Gepts, P. Crop domestication as a long-term selection experiment. *Plant Breed Rev* **24**, 1-44 (2004).
41 Disch, S. *et al.* The E3 Ubiquitin Ligase BIG BROTHER Controls Arabidopsis Organ Size in a Dosage-Dependent Manner. *Current Biology* **16**, 272-279 (2006).
42 Breuer, C. *et al.* BIN4, a novel component of the plant DNA topoisomerase VI complex, is required for endoreduplication in Arabidopsis. *Plant Cell* **19**, 3655-3668 (2007).
43 Pérez-Vega, E. *et al.* Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.). *Theoretical and applied genetics* **120**, 1367-1380 (2010).
44 Koinange, E. M., Singh, S. P. & Gepts, P. Genetic control of the domestication syndrome in common bean. *Crop Science* **36**, 1037-1045 (1996).
45 Weeden, N. F. Genetic changes accompanying the domestication of Pisum sativum: is there a common genetic basis to the 'domestication syndrome'for legumes? *Annals of Botany* **100**, 1017-1025 (2007).
46 Van Daele, I. *et al.* A comparative study of seed yield parameters in *Arabidopsis thaliana* mutants and transgenics. *Plant biotechnology journal* **10**, 488-500 (2012).
47 Hwang, I., Sheen, J. & Muller, B. Cytokinin signaling networks. *Annu Rev Plant Biol* **63**, 353-380 (2012).
48 González, A. M., De la Fuente, M., De Ron, A. M. & Santalla, M. Protein markers and seed size variation in common bean segregating populations. *Molecular Breeding* **25**, 723-740 (2010).

49 JoinMap4. Software for the calculation of genetic linkage maps in experimental populations. (Kyazma BV, Wageningen, The Netherlands, 2006).

50 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome research* **13**, 91-96 (2003).

51 Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666 (2003).

52 Smit, A. F., Hubley, R. & Green, P. *RepeatMasker Open-3.0*, <http://www.repeatmasker.org> (1996-2010).

53 Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. *Genome Research* **10**, 516-522 (2000).

54 Yeh, R.-F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome research* **11**, 803-816 (2001).

55 Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-3646 (2004).

56 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

57 Mafi Moghaddam, S. *et al.* Developing market class specific InDel markers from next generation sequence data in *Phaseolus vulgaris* L. *Frontiers in Plant Science* **1**, 0 (2013).

58 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

59 Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437-460 (1983).

60 Watterson, G. On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7**, 256-276 (1975).

61 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

Figure legends

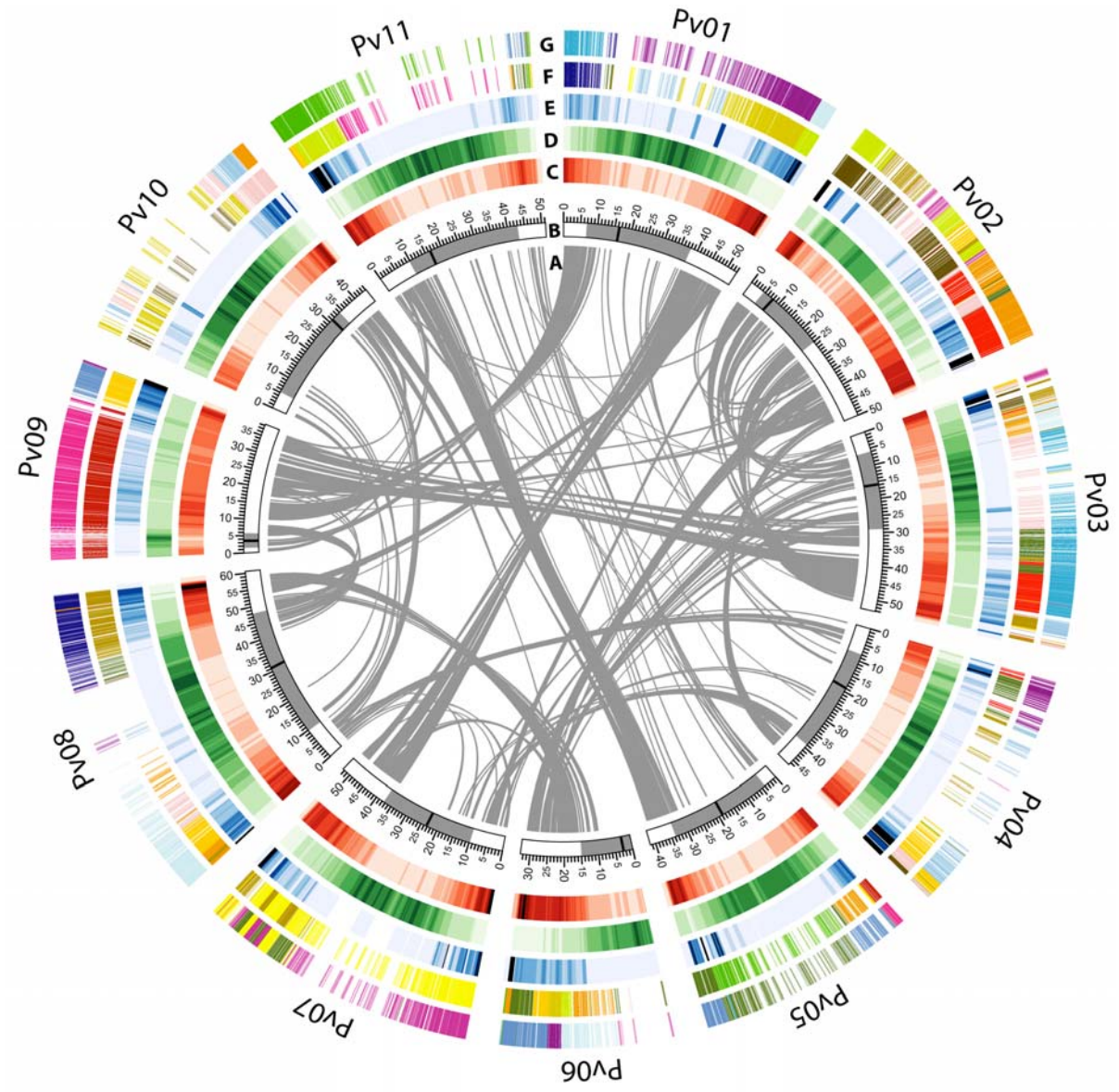
Figure 1. **Structure of the *Phaseolus* genome and synteny with *Glycine max*. A -** Grey lines connect duplicated genes. **B** – Chromosome structure with centromeric and pericentromeric regions in black and grey, respectively (scale is Mbp). **C** – Gene density in sliding windows of 1Mb at 200Kb intervals. **D** – Repeat density in sliding windows of 1Mb at 200Kb intervals. **E** – Recombination rate in cM/Mb based on the genetic and physical mapping of 6945 SNPs and SSRs. **F-G** – Syntenic regions in *Glycine max* which has a lineage-specific duplication resulting in two chromosome segments for every one in *Phaseolus*.

Figure 2. **Geographic distribution of sampled genotypes**.

Figure 3. (a) Divergence of the wild Mesoamerican and Andean common bean pools. The wild Andean diverged from wild Mesoamerican gene pool ~ 165 kya, with a small founding population and strong bottleneck that lasted ~ 76 ky. This was followed by an exponential growth phase till present. Asymmetric gene flow between the two pools played a key role in maintaining genetic diversity, especially in the Andean population, with average migration rates $M21=2*N_{anc}*m21 = 0.135$ (wild Mesoamerican to wild Andean) and M12=0.087 (Andean to Mesoamerican). This scenario conforms to the Mesoamerican origin model of the common bean, with an Andean bottleneck that pre-dated domestication. **(b) Population genomic analysis based on SNP data from resequencing of common bean DNA pools.** The size of the circle for each pool is proportional to the π value for the pool. For a reference, π=0.0061 for the wild Mesoamerican pool. The F_{ST} statistic, representing the differentiation of any two pools, is noted on lines (not proportional) connecting pools. Arrows indicate that evolutionary pathway for each pool. Lines without arrowheads are presented for comparative purposes. The data is the average statistic across all 10kb/2kb sliding windows, discarding windows with <50% called bases.

Figure 4. Differentiation and reduction of diversity during common bean domestication. Genome-wide view in 10kb/2kb sliding windows of differentiation (F_{ST}) and reduction in diversity (π ratio) statistics associated with domestication within the common bean Mesoamerican (**a**) and Andean (**b**) gene pools. Log_{10} π ratio values less than zero are not shown. Lines represent the 90%, 95%, and 99% tail for the empirical distribution for each statistic.

Figure 5. Genome-wide association analysis of seed weight. (a) A 280 member panel of Mesomerican cultivars was grown in four USA locations. The phenotypic data were coupled with 34,799 SNP markers and analyzed using a mixed model analysis that controlled for population structure and genotype relatedness. **(b)** A close-up view of the seed weight GWAS and linkage disequilibrium results around a 1.23 Mb Mesoamerican sweep window on Pv07. The positions of domestication candidates are noted by asterisks above the GWAS display. The candidates range from *Phvul.007G094299* to *Phvul.007G.99700* (see **Supplementary material –Mesoamerican Domestication Candidates** for descriptions).



C - Gene Density 1 64 genes/Mb

D - Repeat Density 0 100 %

E - Recombination rate 0 10 cM/Mb

F / **G** - Soybean chromosomes

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

Figure 1

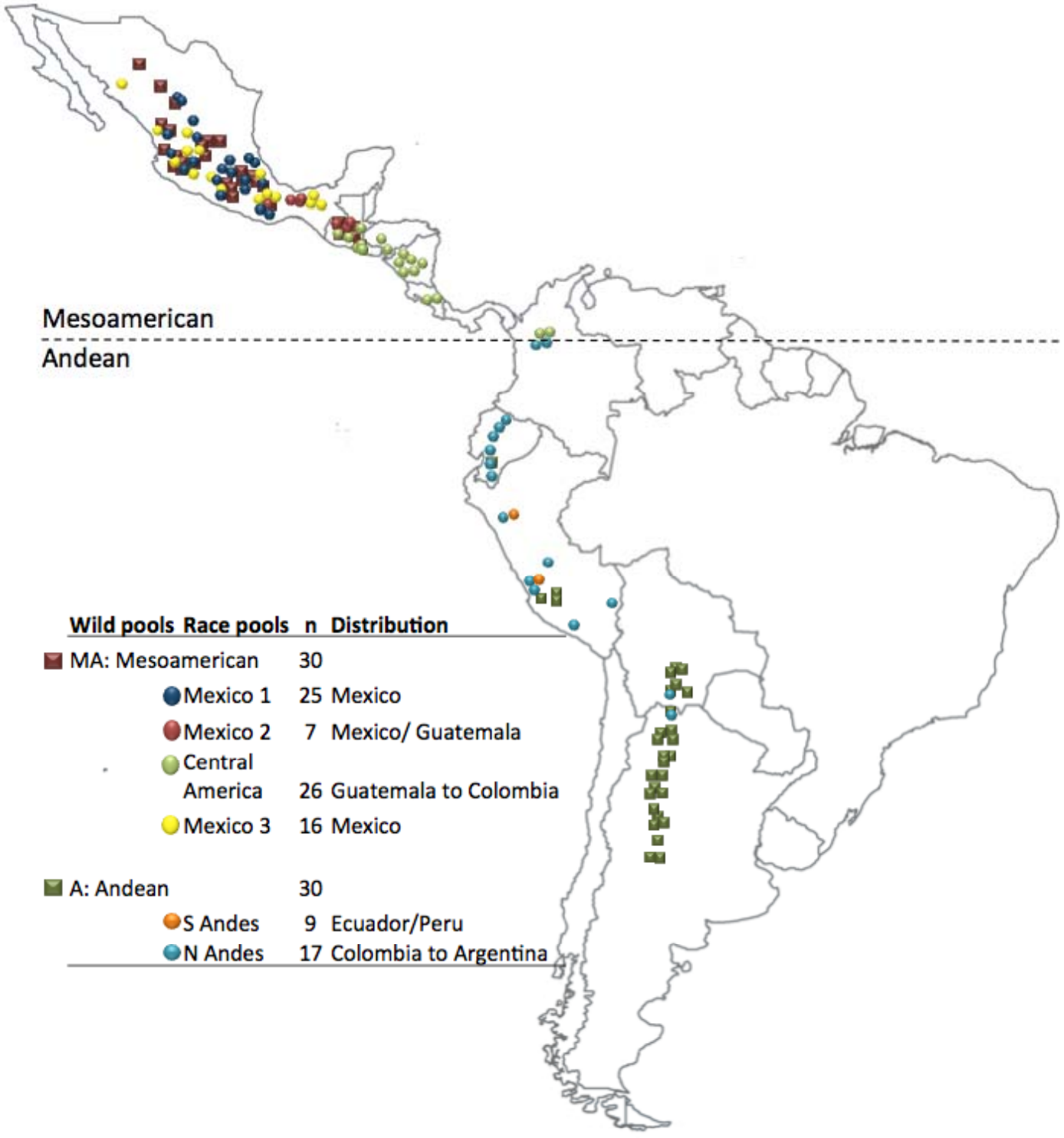

Mesoamerican
Andean
Wild pools Race pools n Distribution
MA: Mesoamerican 30
Mexico 1 25 Mexico
Mexico 2 7 Mexico/ Guatemala
Central America 26 Guatemala to Colombia
Mexico 3 16 Mexico
A: Andean 30
S Andes 9 Ecuador/Peru
N Andes 17 Colombia to Argentina

Figure 2

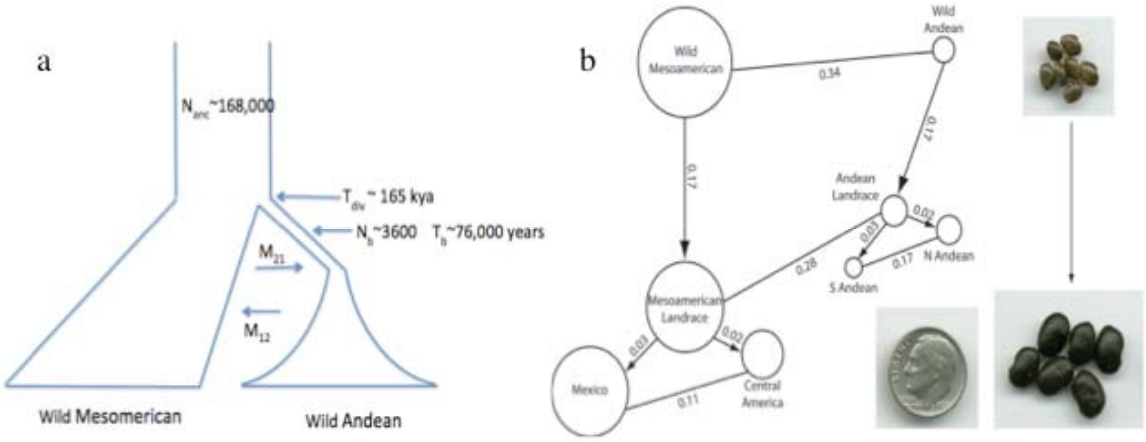

a
N_{anc}~168,000
T_{div} ~ 165 kya
N_b~3600 T_b~76,000 years
M_{21}
M_{12}
Wild Mesomerican
Wild Andean
b
Wild Mesoamerican
Wild Andean
0.34
0.17
0.17
Andean Landrace
0.02
0.03
N Andean
0.17
S Andean
0.28
Mesoamerican Landrace
0.03
0.02
Mexico
Central America
0.11

Figure 3

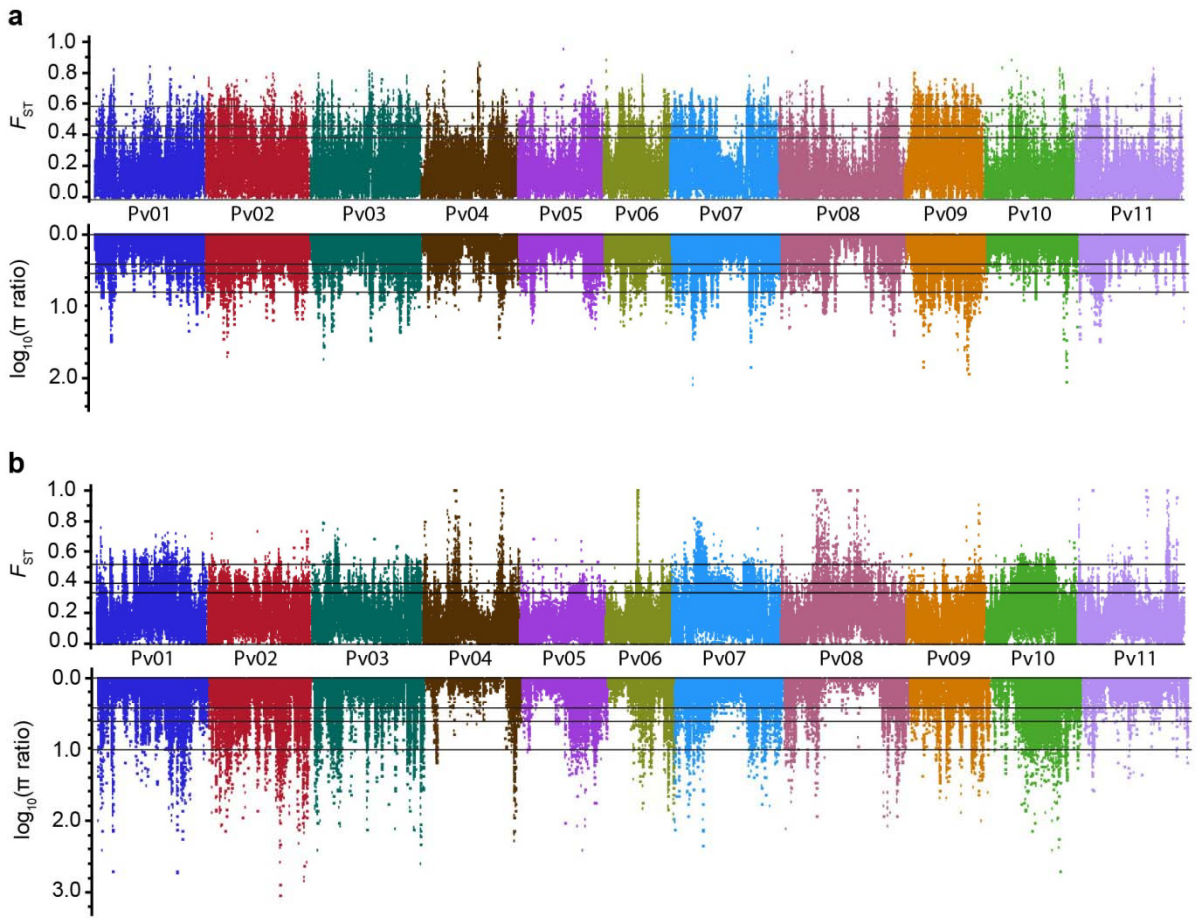

a
1.0
0.8
0.6
0.4
0.2
0.0
F_{ST}
Pv01
Pv02
Pv03
Pv04
Pv05
Pv06
Pv07
Pv08
Pv09
Pv10
Pv11
0.0
1.0
2.0
$\log_{10}(\pi$ ratio)
b
1.0
0.8
0.6
0.4
0.2
0.0
F_{ST}
Pv01
Pv02
Pv03
Pv04
Pv05
Pv06
Pv07
Pv08
Pv09
Pv10
Pv11
0.0
1.0
2.0
3.0
$\log_{10}(\pi$ ratio)

Figure 4

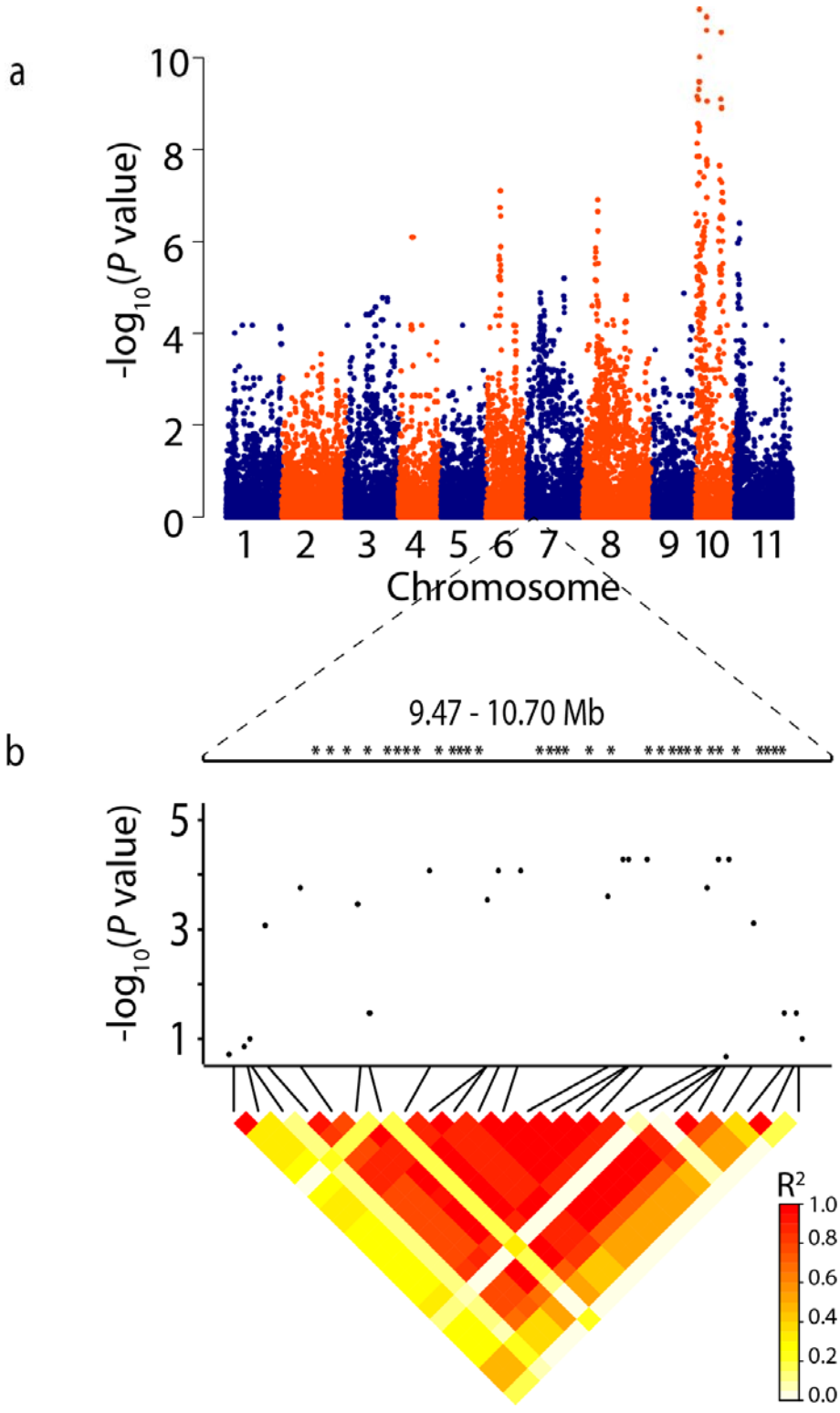

a
10
8
6
4
2
0
-log10(P value)
1 2 3 4 5 6 7 8 9 10 11
Chromosome
9.47 - 10.70 Mb
b
5
3
1
-log10(P value)
R2
1.0
0.8
0.6
0.4
0.2
0.0

Figure 5