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

Jeremy Schmutz1,2,17, Phillip E McClean3,17, Sujan Mamidi3, G Albert Wu1, Steven B Cannon4, Jane Grimwood2, Jerry Jenkins2, Shengqiang Shu1, Qijian Song5, Carolina Chavarro6, Mirayda Torres-Torres6, Valerie Geffroy7,8, Samira Mafi Moghaddam3, Dongying Gao6, Brian Abernathy6, Kerrie Barry1, Matthew Blair9, Mark A Brick10, Mansi Chovatia1, Paul Gepts11, David M Goodstein1, Michael Gonzales6, Uffe Hellsten1, David L Hyten5,16, Gaofeng Jia5, James D Kelly12, Dave Kudrna13, Rian Lee3, Manon M S Richard7, Phillip N Miklas14, Juan M Osorno3, Josiane Rodrigues5,16, Vincent Thareau7, Carlos A Urrea15, Mei Wang1, Yeisoo Yu13, Ming Zhang1, Rod A Wing12, Perry B Cregan5, Daniel S Rokhsar1 & Scott A Jackson6

Common bean (Phaseolus vulgaris L.) is the most important grain legume for human consumption and has a role in sustainable agriculture owing to its ability to fix atmospheric nitrogen. We assembled 473 Mb of the 587-Mb genome and genetically anchored 98% of this sequence in 11 chromosome-scale pseudomolecules. We compared the genome for the common bean against the soybean genome to find changes in soybean resulting from polyploidy. Using resequencing of 60 wild individuals and 100 landraces from the genetically differentiated Mesoamerican and Andean gene pools, we confirmed 2 independent domestications from genetic pools that diverged before human colonization. Less than 10% of the 74 Mb of sequence putatively involved in domestication was shared by the two domestication events. We identified a set of genes linked with increased leaf and seed size and combined these results with quantitative trait locus data from Mesoamerican cultivars. Genes affected by domestication may be useful for genomics-enabled crop improvement.

Common bean (P. vulgaris L.) is a crop of major societal importance and is a major 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 (see URLs). More than 200 million people in sub-Saharan Africa depend on the common bean as a primary staple. It has many health-beneficial1,2 nutrients whose concentrations are heritable3, and increasing the concentrations of these nutrients is a breeding objective worldwide4.

Multiple lines of evidence have shown that wild common bean is organized in two geographically isolated and genetically differentiated wild gene pools (Mesoamerican and Andean) that diverged from a common ancestral wild population more than 100,000 years ago5. From these wild gene pools, nearly 8,000 years ago, common bean was independently domesticated in what is now Mexico and in South America6–9, and these domestication events were followed by local adaptations resulting in landraces with distinct characteristics. In what is now 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 Americas10. Domestication led to morphological changes, including increased seed and leaf sizes, changes in growth habit and photoperiod responses11, and variation in seed coat color and pattern that distinguish culturally adapted classes of beans12.

Independent domestication events, starting from distinct gene pools of a single species, provide experimental replication not typically found in domestication or evolutionary studies. It is possible to deduce domestication history on 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, 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 southern range of the species in Argentina. In addition, comparative genomics with soybean (Glycine max), a closely related crop, identified effects of shared and lineage-dependent polyploidies on gene fractionation and recent transposable element expansion in the common bean.

RESULTS

Reference genome and analysis

To obtain a high-quality reference genome, we sequenced an inbred landrace line of P. vulgaris (G19833) derived from the Andean pool (Race Peru) using a whole-genome shotgun sequencing strategy that combined multiple linear libraries (18.6× assembled sequence coverage) and ten paired libraries of varying insert sizes (1.8× assembled) sequenced with the Roche 454 platform together with 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 (0.54× long-insert pairs) for a total assembled sequence coverage level of 21.0× (Supplementary Tables 1 and 2). 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 F2 lines from a StampeDe × Red Hawk cross and a similar set of Infinium markers and 261 SSRs (simple sequence repeats) typed on 88 F2-derived recombinant inbred lines (RILs) derived from the same cross (P.B.C. and Q.S., unpublished data). Additional refinements to the pseudomolecules were made on the basis of synteny with soybean (G. max), where available by allowed 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 included 468.2 Mb of mapped sequence in 240 scaffolds. The total release includes 472.5 Mb of the ~587-Mb genome (see URLs), with half of the assembled nucleotides in contigs longer than 39.5 kb (contig N50) (Supplementary Table 3). To annotate the chromosomal assembly, we combined Sanger-derived EST resources and a substantial amount of new RNA sequencing (RNA-seq) reads (727 million reads from 11 tissues and developmental stages; Supplementary Table 4) 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 5), an underestimate that will increase with additional transcriptomes and analyses. Most of these genes (91%) were retained in synteny blocks with G. max (Supplementary Note).

We identified recent transposable element activity and expansions of transposon numbers (Supplementary Figs. 1–3). Although recently diverged repeats could not be annotated directly from Roche 454 pyrosequencing data, extensive BAC-end 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. 4–14). Centromere and pericentromeric regions were primarily repetitive, and, similar to in other sequenced genomes13,14, these pericentromeric genomic regions were recombinationally inert (Supplementary Fig. 15 and Supplementary Table 6). Using a threshold of 2 Mb/cM to identify transitions into pericentromeric regions, pericentromeres spanned ~54% of the genome and had an average recombination rate of 4,350 kb/cM versus 220 kb/cM in the euchromatic arms (Supplementary Table 7). The pericentromeres were primarily repetitive but, owing to their size, still contained 26.5% of the genes.

The majority of the repetitive elements in the genome were long terminal repeat (LTR) retrotransposons, and we identified 2,668 complete LTR retrotransposons and classified them into 165 families, including 65 Ty1-copia, 78 Ty3-gypsy and 22 unclassified families (Supplementary Tables 8 and 9). Although there were ancient elements that inserted into the genome more than 10 million years ago, ~75% (2,011/2,668) of the LTR retroelements integrated into P. vulgaris within the last 2 million years (Supplementary Fig. 1). Notably, the insertion times of 20% (543/2,668) of the elements were more recent than 0.5 million years ago—this is likely an underestimate, as our sequencing approach is biased against the annotation of completely identical LTRs. These results were similar to those in soybean15 and suggest that LTR retrotransposons underwent recent amplification events in both legumes. The 165 LTR retrotransposon families varied in the copy number of complete elements: more than 78% (130/165) of the families had fewer than 10 complete retroelements, whereas 11 families had more than 50 complete elements and contained 63% (1,690/2,668) of the complete elements in the P. vulgaris genome. Some families showed extremely high copy numbers; for example, the pvRetroS2 family contained 446 complete elements (likely an underestimate, as some elements would not have been annotated uniquely).

We observed dense clusters of resistance-associated genes in the common bean genome. The majority of putative resistance-associated genes in plants encode nucleotide-binding and leucine-rich repeat domains and are collectively known as NB-LRR (NL) genes13. We identified 376 NL genes, of which 106 encoded an N-terminal Toll/interleukin-1 receptor (TIR)-like domain (TNLs) and 108 encoded an N-terminal coiled-coil domain (CNLs) (Supplementary Table 10). The majority of NL sequences were physically organized in complex clusters, often located at the ends of chromosomes (Supplementary Fig. 16). In particular, three large clusters were located at the ends of chromosomes Pv04, Pv10 and Pv11 and contained more than 40 NL genes that were enriched for CNL (Pv04 and Pv11) or TNL (Pv10) genes that colocalized with previously mapped genes related to disease resistance16–21. Local tandem duplications and ectopic recombination between clusters are involved in the evolution of these NL gene clusters22.

Comparison of genome changes in sister legume species

P. vulgaris (common bean) and G. max (soybean) diverged ~19.2 million years ago but shared a whole-genome duplication (WGD) event ~56.5 million years ago23. G. max experienced an independent WGD ~10 million years ago14. These events were evident in plots of synonymous changes in coding sequences (Ks) between and within these genomes (Supplementary Fig. 17), which also showed that P. vulgaris has evolved more rapidly than G. max since they split from their last common ancestor. Assuming a divergence time of ~19.2 million years ago23, the Ks value (synonymous substitution rate) for P. vulgaris was 1.4 times that of G. max (8.46 × 10−9 versus 5.85 × 10−9 substitutions/year).

We identified orthologous P. vulgaris and G. max genes using synteny and Ks values as criteria (Supplementary Table 11). Consistent with earlier work, there was extensive synteny between P. vulgaris and G. max, except in pericentromeric regions, where microcollinearity was often stretched out and thinned owing to genomic expansion in one or both genomes. Typically, two chromosomal blocks in G. max mapped to a single region of P. vulgaris owing to the most recent WGD in G. max (Fig. 1)14,24,25. Most of the P. vulgaris genes (91%; 24,861) were in identifiable synteny blocks in G. max, and 57% were in synteny blocks in P. vulgaris itself—a result of the ancient WGD event 55 million
years ago. Within synteny blocks, the
G. max–G. max duplication had a mean of
33 genes/block, whereas the older, shared
P. vulgaris–G. max WGD event had an aver-
age of 14 genes/block.

Evolution of gene pools in common bean
Mesoamerica has been suggested to be the
center from which common bean originated,
ultimately forming the distinct modern wild
Andean and Mesoamerican gene pools. To
investigate the differentiation of these wild
populations, we performed pooled resequencing of 30 individuals each from
Mesoamerican and Andean wild populations
(Fig. 2 and Supplementary Table 12). Using
π (the average pairwise nucleotide differ-
ces in a sample) and θ (the proportion of
nucleotide polymorphisms in a sample), the
Mesoamerican wild population (π (per bp) =
0.0061; θ (per bp) = 0.0041) was more diverse than the Andean
wild population (π (per bp) = 0.0014; θ (per bp) = 0.0013). We used
~663,000 polymorphic sites (at least 5 kb from a gene and not in a repeat
sequence) to estimate demographic parameters using the joint allele
frequency spectrum (δαι)26 (Supplementary Note). The strong fixation
index FST of ~0.34 between these two wild populations indicates that they have substantial allelic differentiation from each other. We
estimated that divergence of the two wild pools occurred ~165,000 years
ago, with an ancestral effective population size of 168,000. This
date is earlier than a previous estimate of ~110,000 years ago but falls
within the 95% confidence interval of the previous estimate, which was based on 13 loci from 24 wild genotypes25, but it is later than other estimates of ~500,000 years ago27. The whole-genome analysis resulted in
a much tighter confidence interval of 146,000–184,000 years ago.

Demographic inference for the wild Andean gene pool suggested that it was derived from the wild Mesoamerican population with a founding population of only a few thousand individuals (Fig. 3a and
Supplementary Note). The wild Andean population showed no appreciable growth in effective population size for ~76,000 years
after founding, although there was continual asymmetric gene flow between the two wild populations, with a higher Mesoamerican-
to-Andean migration rate (Supplementary Table 13). The Andean
population then underwent an exponential growth phase that began
~90,000 years ago and has continued to the present. The strong pre-
domestication bottleneck in the Andean population has been observed in previous analyses28,29; in contrast, however, no detectable bot-
tleneck 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 4 pooled populations representing distinct Mesoamerican
landraces and 2 pooled populations representing distinct Andean
landraces (n = 7–26 landraces). These landraces represent subpopulations from Mexico, Central America and South America with low levels of admixture (Supplementary Fig. 18). Because the four Mesoamerican 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 Mesoamerican and a composite Andean landrace SNP data set, respectively, for further
analysis. This approach allowed us to distinguish selection from ran-
dom fixation across the genome30 and to search for signals associ-
ated with domestication events. The number of SNPs ranged from
8,890,318 for the wild Mesoamerican subpopulation to 1,397,405 SNPs
for the Andean landrace subpopulation from Peru (Supplementary
Table 14), and ~16% of these SNPs were within genes.

To characterize variation among the populations, we calculated
diversity (π) and population differentiation (FST) statistics using data
averaged over 10-kb windows with a 2-kb slide (10-kb/2-kb windows;
Supplementary Table 15). Whereas the Mesoamerican landraces were
less diverse than the wild Mesoamerican population, Andean landrace
populations were more diverse than the wild Andean population, possibly owing to admixture with Mesoamerican populations and/or
de novo mutation within the Andean gene pool. Diversity was further
reduced within the Mesoamerican Central American and southern
Andean landraces, suggesting that these subpopulations underwent additional selection that might correspond to local adaptation.

Multiple results point to independent domestication events in the Mesoamerican and Andean gene pools, a feature observed for only a few modern crops. We characterized domestication of common bean at the genomic level by comparing wild and landrace populations across 10-kb/2-kb sliding windows, selecting windows that met strict composite criteria that required they be in the top 90% of the population's empirical distribution for both $\pi_{\text{wild}}/\pi_{\text{landrace}}$ ratios and $F_{ST}$ values (Figs. 3b,c and 4). We observed 930 windows in Mesoamerican populations (totaling 74 Mb of sequence) with both low diversity and high differentiation. Because low diversity and high differentiation are two features of selection, we consider these to be selection windows. Of these windows, 209 that were longer than 100 kb accounted for 70.1% of the total selection distance. Among the 750 selection windows in Andean populations exhibiting low diversity and high differentiation, 172 that were longer than 100 kb covered 69.8% of the total selection distance (60 Mb). As expected for independent Mesoamerican and Andean domestication events, these selection regions were distinct. Within the Mesoamerican landrace population, chromosomes 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 event primarily involved chromosomes Pv01, Pv02 and Pv10 (Fig. 4). Interestingly, only 7.234 Mb of the regions predicted to be involved in domestication were shared by the two gene pools, suggesting different genetic routes to domestication.

We identified candidate genes associated with domestication using the same criteria applied to find selection windows (requiring that they be in the top 90% of the pool's empirical distribution for both $\pi_{\text{wild}}/\pi_{\text{landrace}}$ ratios and $F_{ST}$ values). We identified 1,835 Mesoamerican and 748 Andean candidate genes associated with domestication (Supplementary Tables 16 and 17), and all candidates had a negative Tajima's $D$ value, indicating positive selection. Most notably, only 59 of the candidate genes (3% of the Mesoamerican and 8% of the Andean candidates) were shared by the 2 landrace populations. For the 59 common candidates, the mean $F_{ST}$ value was 0.67, suggesting selection on different alleles or the appearance of unique mutations in the two gene pools. This finding is consistent with evidence at the PvTF1l1 QTL. For many loci, we calculated $D$, the number of unique alleles derived in each pool, and $H$, the number of alleles shared between the pools. In general, $D$ was lower (less than 0.05) than $H$ (less than 0.25). Interestingly, $D$ and $H$ were higher for some of the domestication loci compared to non-domestication loci.

Figure 2 Geographic distribution of sampled genotypes.

Figure 3 Evolution and domestication of common bean. (a) Divergence of the wild Mesoamerican and Andean common bean pools. The wild Andean gene pool diverged from the wild Mesoamerican gene pool ~165,000 years ago, with a small founding population and a strong bottleneck that lasted ~76,000 years. The bottleneck was followed by an exponential growth phase extending to the present day. Asymmetric gene flow between the two pools had a key role in maintaining genetic diversity, especially in the Andean population, with average migration rates $M_{21} = 0.135$ (wild Mesoamerican to wild Andean) and $M_{12} = 0.087$ (wild Andean to wild Mesoamerican). This scenario conforms to the Mesoamerican origin model of the common bean, with an Andean bottleneck that predated domestication. ($n_{anc}$, size of ancestral population; $t_{bottleneck}$, start of bottleneck; $n_{bottleneck}$, size of bottleneck population; $L_{bottleneck}$, length of bottleneck)

(b) Population genomic analysis based on SNP data from the resequencing of DNA pools for common bean. The size of the circle for each pool is proportional to the $\pi$ value for the pool. For a reference, $\pi = 0.0061$ for the wild Mesoamerican (MA) pool. $F_{ST}$ statistics, representing the differentiation of any two pools, are noted on the lines (not proportional) connecting pools. Data are average statistics across all 10-kb/2-kb sliding windows with <50% called bases. Land, landrace; N, north; S, south; C, central.

(c) Variation in seed size in common bean. The seeds of wild Mesoamerican and Andean beans (two each) are smaller than the seeds corresponding to the reference genotype (G19833) and the multiple market classes of common beans grown in the United States (navy to light red kidney).
in the Mesoamerican Mexican landraces. The fact that only 33 of these genes were shared by these 2 subpopulations indicates unique evolutionary trajectories among subpopulations of the Mesoamerican gene pool. Within the Andean gene pool, none of the candidate genes from the northern and southern Andean landrace populations were shared. These results demonstrate that the sexually compatible Mesoamerican and Andean lineages with similar morphologies and life cycles underwent independent selection upon distinct sets of genes. This is in contrast to the situation in rice, where many major domestication genes were shared by gene flow between the indica and japonica types.

Domestication had distinct effects on genes involved in flowering and the vernalization pathway. Orthologs of VRN1 (Phvul.003G033400) and VRN2 (Phvul.002G000500) were Mesoamerican candidate genes, and orthologs of FRL1 (Phvul.006G053200) and TFL2 (Phvul.009G117500) were Andean candidate genes. COP1 encodes a photoperiod pathway regulator that controls FT through CO. The Mesoamerican ortholog of COP1 was a candidate domestication gene, and Phvul.006G165300, a CUL4 ortholog that encodes a protein that is part of a complex that along with COP1 regulates CO, was an Andean candidate gene for domestication. This finding demonstrates independent selection on genes encoding different members of the same protein complex. The only shared domestication candidates were Phvul.007065600, an ortholog of AGL42, which regulates flowering through the gibberellin pathway, and Phvul.009G203400, an ortholog of FUL, which regulates SOC1.

Increased plant size is typically associated with plant domestication and multiple Mesoamerican candidate genes influence this trait. Phvul.011G213300 is an ortholog of Arabidopsis thaliana BB, a component of the ubiquitin ligase degradation pathway that controls flower and stem size, and Phvul.009G040200 is an ortholog of BIN4, which regulates cell expansion and final plant size. Multiple candidate genes for domestication were also components of nitrogen metabolism pathways, which directly affect plant size. The Mesoamerican candidate gene Phvul.008G168000 encodes nitrate reductase, a critical element for plant and seed growth, which genetically maps to the SW8.2 quantitative trait locus (QTL) for seed weight. Other candidate genes for domestication involved in nitrogen metabolism included the Mesoamerican (Phvul.005G132200) and Andean (Phvul.002G242900) nitrogen transporters and the Mesoamerican asparagine synthase (Phvul.006G069300).

Increased seed size is a major phenotypic shift associated with the domestication of the common bean and other legumes and was a candidate domestication gene, and orthologs of FRL1 (Phvul.006G053200) and TFL2 (Phvul.009G117500) were Andean candidate genes. COP1 encodes a photoperiod pathway regulator that controls FT through CO. The Mesoamerican ortholog of COP1 was a candidate domestication gene, and Phvul.006G165300, a CUL4 ortholog that encodes a protein that is part of a complex that along with COP1 regulates CO, was an Andean candidate gene for domestication. This finding demonstrates independent selection on genes encoding different members of the same protein complex. The only shared domestication candidates were Phvul.007065600, an ortholog of AGL42, which regulates flowering through the gibberellin pathway, and Phvul.009G203400, an ortholog of FUL, which regulates SOC1.

Increased plant size is typically associated with plant domestication and multiple Mesoamerican candidate genes influence this trait. Phvul.011G213300 is an ortholog of Arabidopsis thaliana BB, a component of the ubiquitin ligase degradation pathway that controls flower and stem size, and Phvul.009G040200 is an ortholog of BIN4, which regulates cell expansion and final plant size. Multiple candidate genes for domestication were also components of nitrogen metabolism pathways, which directly affect plant size. The Mesoamerican candidate gene Phvul.008G168000 encodes nitrate reductase, a critical element for plant and seed growth, which genetically maps to the SW8.2 quantitative trait locus (QTL) for seed weight. Other candidate genes for domestication involved in nitrogen metabolism included the Mesoamerican (Phvul.005G132200) and Andean (Phvul.002G242900) nitrogen transporters and the Mesoamerican asparagine synthase (Phvul.006G069300).

Increased seed size is a major phenotypic shift associated with the domestication of the common bean and other legumes and was a candidate domestication gene, and orthologs of FRL1 (Phvul.006G053200) and TFL2 (Phvul.009G117500) were Andean candidate genes. COP1 encodes a photoperiod pathway regulator that controls FT through CO. The Mesoamerican ortholog of COP1 was a candidate domestication gene, and Phvul.006G165300, a CUL4 ortholog that encodes a protein that is part of a complex that along with COP1 regulates CO, was an Andean candidate gene for domestication. This finding demonstrates independent selection on genes encoding different members of the same protein complex. The only shared domestication candidates were Phvul.007065600, an ortholog of AGL42, which regulates flowering through the gibberellin pathway, and Phvul.009G203400, an ortholog of FUL, which regulates SOC1.

Increased plant size is typically associated with plant domestication and multiple Mesoamerican candidate genes influence this trait. Phvul.011G213300 is an ortholog of Arabidopsis thaliana BB, a component of the ubiquitin ligase degradation pathway that controls flower and stem size, and Phvul.009G040200 is an ortholog of BIN4, which regulates cell expansion and final plant size. Multiple candidate genes for domestication were also components of nitrogen metabolism pathways, which directly affect plant size. The Mesoamerican candidate gene Phvul.008G168000 encodes nitrate reductase, a critical element for plant and seed growth, which genetically maps to the SW8.2 quantitative trait locus (QTL) for seed weight. Other candidate genes for domestication involved in nitrogen metabolism included the Mesoamerican (Phvul.005G132200) and Andean (Phvul.002G242900) nitrogen transporters and the Mesoamerican asparagine synthase (Phvul.006G069300).

Increased seed size is a major phenotypic shift associated with the domestication of the common bean and other legumes and was a candidate domestication gene, and orthologs of FRL1 (Phvul.006G053200) and TFL2 (Phvul.009G117500) were Andean candidate genes. COP1 encodes a photoperiod pathway regulator that controls FT through CO. The Mesoamerican ortholog of COP1 was a candidate domestication gene, and Phvul.006G165300, a CUL4 ortholog that encodes a protein that is part of a complex that along with COP1 regulates CO, was an Andean candidate gene for domestication. This finding demonstrates independent selection on genes encoding different members of the same protein complex. The only shared domestication candidates were Phvul.007065600, an ortholog of AGL42, which regulates flowering through the gibberellin pathway, and Phvul.009G203400, an ortholog of FUL, which regulates SOC1.

Increased plant size is typically associated with plant domestication and multiple Mesoamerican candidate genes influence this trait. Phvul.011G213300 is an ortholog of Arabidopsis thaliana BB, a component of the ubiquitin ligase degradation pathway that controls flower and stem size, and Phvul.009G040200 is an ortholog of BIN4, which regulates cell expansion and final plant size. Multiple candidate genes for domestication were also components of nitrogen metabolism pathways, which directly affect plant size. The Mesoamerican candidate gene Phvul.008G168000 encodes nitrate reductase, a critical element for plant and seed growth, which genetically maps to the SW8.2 quantitative trait locus (QTL) for seed weight. Other candidate genes for domestication involved in nitrogen metabolism included the Mesoamerican (Phvul.005G132200) and Andean (Phvul.002G242900) nitrogen transporters and the Mesoamerican asparagine synthase (Phvul.006G069300).

Increased seed size is a major phenotypic shift associated with the domestication of the common bean and other legumes and was a candidate domestication gene, and orthologs of FRL1 (Phvul.006G053200) and TFL2 (Phvul.009G117500) were Andean candidate genes. COP1 encodes a photoperiod pathway regulator that controls FT through CO. The Mesoamerican ortholog of COP1 was a candidate domestication gene, and Phvul.006G165300, a CUL4 ortholog that encodes a protein that is part of a complex that along with COP1 regulates CO, was an Andean candidate gene for domestication. This finding demonstrates independent selection on genes encoding different members of the same protein complex. The only shared domestication candidates were Phvul.007065600, an ortholog of AGL42, which regulates flowering through the gibberellin pathway, and Phvul.009G203400, an ortholog of FUL, which regulates SOC1.
distinguishes the many types of beans that humans consume. We surveyed the Mesoamerican domestication candidates for genes previously shown to be associated with seed weight and used the whole-genome sequence for a genome-wide association study (GWAS; Fig. 5a) to understand the genetic architecture of seed weight in modern Mesoamerican cultivars. We found 15 candidate genes previously shown to be involved in seed weight (Supplementary Table 19). Among these are nearly all the components of the cytokinin synthesis and multiple-component phosphorylation regulatory system (Supplementary Fig. 19). Included are Phvul.002G082400, which encodes a protein that transmits the phosphosignal in response to 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. An additional candidate gene, Phvul.01G038800, has orthologs that encode cytokinin oxidase/dehydrogenase proteins, which regulate the pathway by degrading active cytokinin. The relevance of these genes as candidate loci associated with seed weight is supported by work in Arabidopsis, where orthologs of the candidate genes in the cytokinin pathway have been shown in transgenic studies to regulate seed size and/or weight. In contrast, however, none of these genes were Andean domestication candidates.

GWAS analysis for seed weight confirmed three of these domestication candidates. It was not possible to confirm the other 12 candidates by GWAS because Mesoamerican domestication reduced diversity to near homozygosity, such that associations could not be found (Supplementary Table 20). GWAS analysis was able to place 75 domestication candidate genes within 50 kb of a SNP significantly associated with seed weight, and a significantly associated SNP was found within eight candidate genes (Supplementary Table 21). One sweep window on Pv07 (9.662–10.662 Mb) contained 33 domestication candidates and was located in a GWAS peak that exhibited extensive linkage disequilibrium (Fig. 5b). By GWAS, we also detected candidate genes for seed weight that resulted from modern breeding of the common bean. These included 15 improvement-related genes previously shown to be associated with seed weight, 5 of which function in the cytokinin regulation/degradation pathway (Supplementary Table 22). Finally, three genes in complete linkage disequilibrium with equally significant association ($P = 6.3 \times 10^{-6}$) were located in a Pv07 QTL for seed weight that has been replicated in many experiments.

DISCUSSION

Common bean is the most important grain legume for human consumption and is an especially 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 abiotic and biotic stresses. The clustering of resistance-associated genes in a few genomic locations suggests that stacking resistances between clusters should be relatively easy but that stacking multiple resistance genes located within a single physical cluster and then combining these traits by breeding may prove more challenging. The observation that the dual domestication events for common bean had few selective sweeps in common leads us to posit that domestication, previously thought to typically be associated with selection at a few major loci, can also be achieved via multiple genetic pathways resulting in similar or the same phenotypes (for example, seed size). In addition, the lack of correspondence between selective sweeps in domestication and genetic bottlenecks imposed by breeding indicates that domestication-derived traits were fixed early and that subsequent selection was likely on traits for local adaptation and desired seed and plant traits. Together, these findings provide information on regions of the genome that have undergone intense selection, 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

Methods and any associated references are available in the online version of the paper.

Accession codes. Assembly and annotation are available at http://www.phytozone.net/commonbean.php and have been deposited in GenBank under accession ANNZ01000000.

METHODS

Methods and any associated references are available in the online version of the paper.

ACKNOWLEDGMENTS

The work conducted by the US Department of Energy joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract DE-AC02-05CH11231. This research was funded by grants from the US Department of Agriculture–National Institute for Food and Agriculture (2006-35300-17266) and the National Science Foundation (DBI 0822258) to S.A.J. and from the US Department of Agriculture–Cooperative State Research, Education and Extension Service (2009-01860 and 2009-01929) to S.A.J. and from the US Department of Agriculture–National Institute for Food and Agriculture.

AUTHOR CONTRIBUTIONS

J.S., P.E.M., D.S.R. and S.A.J. conceived the study and jointly wrote the manuscript with S.B.C. Genomic clones and DNA were provided by R.A.W., Y.Y., D.K., R.L. and M.B. The following analyses were performed by the indicated investigators: repeat annotation, D.G.; identification of 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.J.; sequencing, assembly and/or annotation, J.G., J.J., S.S., K.B., M.C., D.M.G., U.H., M.W. and M.Z.; comparative, population and/or evolutionary analyses, S.M., G.A.W., S.B.C., C.C., S.M.M., B.A., M.T.-T. and M.G.; and GWAS, S.M.M., M.A.B., P.G., J.D.K., P.N.M., J.M.O. and C.A.U.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/.

1. Anderson, J.W. et al. Hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men. Am. J. Clin. Nutr. 40, 1146–1155 (1984).
2. Geil, P. & Anderson, J. Nutrition and health implications of dry beans: a review. J. Am. Coll. Nutr. 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 Sci. 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 110, 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 Phytol. 197, 300–313 (2013).
7. Bitocchi, E. et al. Mesoamerican origin of the common bean (Phaseolus vulgaris L.) is revealed by sequence data. Proc. Natl. Acad. Sci. USA 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. *Econ. Bot.* 40, 451–468 (1986).

9. Mamidi, S. et al. Investigation of the domestication of common bean (Phaseolus vulgaris) using multilocus sequence data. *Func. Plant Biol.* 38, 953–967 (2011).

10. Zúñiga-Villarreal, D. & Colunga-GarcíaMarín, P. Origin of agriculture and plant domestication in West Mesoamerica. *Genet. Resour. Crop Evol.* 57, 813–829 (2010).

11. Singh, S.P., Gepts, P. & Debouck, D.G. Races of common bean (Phaseolus vulgaris, Fabaceae). *Econ. Bot.* 45, 379–396 (1991).

12. McClean, P.E., 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.). *J. Hered.* 93, 148–152 (2002).

13. Paterson, A.H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551–556 (2009).

14. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183 (2010).

15. Meyers, B.C., Kaushik, S. & Nandety, R.S. Evolving disease resistance genes. *Curr. Opin. Plant Biol.* 9, 129–134 (2005).

16. Geffroy, V. 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. *Theor. Appl. Genet.* 120, 108–123 (2010).

17. Gepts, P. Crop domestication as a long-term selection experiment. *Plant Breed. Rev.* 24, 1–44 (2004).

18. 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 Physiol.* 148, 1740–1759 (2008).

19. 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 Phytol.* 187, 941–956 (2010).

20. 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. *Theor. Appl. Genet.* 96, 494–502 (1998).

21. 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, 109–131 (2006).

22. Davd, P. et al. A nomadic subtelomeric disease resistance gene cluster in common bean. *Plant Physiol.* 151, 1048–1065 (2009).

23. Lavin, M., Herendeen, P.S. & Wojciechowski, M.F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst. Biol.* 54, 575–594 (2005).

24. Gill, N. et al. Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* 151, 1167–1174 (2009).

25. 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).

26. 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 (2009).

27. Chacón S.M., Pickersgill, B. & Debouck, D.G. Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theor. Appl. Genet.* 118, 979–992 (2009).

28. Rossi, M. et al. Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. *Evol. Appl.* 2, 504–522 (2009).

29. Rubin, D.-J. et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591 (2010).

30. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* 127, 1309–1321 (2006).

31. Repinski, S.L., Kwak, M. & Gepts, P. The common bean growth habit gene PvTFL1y is a functional homolog of *Arabidopsis* TFL1. *Theor. Appl. Genet.* 124, 1539–1547 (2012).

32. Sweeney, M.T. et al. Global dissemination of a single mutation conferring white pericarp in rice. *PloS Genet.* 3, e113 (2007).

33. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501 (2012).

34. 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 Physiol.* 148, 1740–1759 (2008).

35. Chen, H. et al. *Arabidopsis* CULLIN4-damaged DNA binding protein 1 interacts with CONSTITUTIVE PHOTOMORPHOGENIC1–SUPPRESSOR OF PHYA complexes to regulate photomorphogenesis and flowering time. *Plant Cell* 22, 108–123 (2010).

36. Gepts, P. Crop domestication as a long-term selection experiment. *Plant Breed. Rev.* 24, 1–44 (2004).

37. Disch, S. et al. The E3 ubiquitin ligase BIG BROTHER controls *Arabidopsis* organ size in a dosage-dependent manner. *Curr. Biol.* 16, 272–279 (2006).

38. Breuer, I. et al. BINA, a novel component of the plant DNA topoisomerase VI complex, is required for endoreduplication in *Arabidopsis*. *Plant Cell* 19, 3655–3668 (2007).

39. 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.). *Theor. Appl. Genet.* 120, 1367–1380 (2010).

40. Koinange, E.M., Singh, S.P. & Gepts, P. Genetic control of the domestication syndrome in common bean. *Crop Sci.* 36, 1037–1045 (1996).

41. Gepts, P. Genetic changes accompanying the domestication of *Pisum sativum* is there a common genetic basis to the ‘domestication syndrome’for legumes? *Ann. Bot.* 100, 1017–1025 (2007).

42. van Dalee, I. et al. A comparative study of seed yield parameters in *Arabidopsis* thaliana mutants and transgenics. *Plant Biotechnol.* J. 10, 488–500 (2012).

43. Hwang, I., Sheen, J. & Muller, B. Cytokinin signaling networks. *Annu. Rev. Plant Biol.* 63, 353–380 (2012).

44. González, A.M., De la Fuente, M., De Ron, A.M. & Santalla, M. Protein markers and seed size variation in common bean segregating populations. *Mol. Breed.* 25, 723–740 (2010).
ONLINE METHODS

Sequencing. The majority of de novo genome sequencing reads were collected with standard sequencing protocols provided by the manufacturer on Roche 454 XLR and Illumina HiSeq 2000 machines at the Department of Energy Joint Genome Institute in Walnut Creek, California. Two types of linear 454 data were collected, standard XLR data (31 runs; 10.7 Gb) and FLX+ data (8.5 runs; 5.615 Gb). Six different paired 454 libraries were created, three libraries with average insert sizes of 2.8–4.8 kb, 1 library with average insert size of 8.0 kb, 1 library with average insert size of 9.2 kb, 1 library with average insert size of 11.9 kb and 1 library with average insert size of 12.2 kb, and were sequenced by standard XLR (26.5 runs; 6.282 Gb of useable data). Two standard 400-bp fragment libraries were sequenced at 2 × 101 bp (four channels; 135.8 Gb) on an Illumina HiSeq 2000. Two fosmid libraries (328,704 reads; 223.9 Mb) with 35.0-kb and 36.0-kb insert sizes and 3 BAC libraries (89,017 reads; 55.1 Mb) with 127.0-kb, 92,160 reads; 65.9 Mb), 133.5-kb (81,408 reads; 57.6 Mb) and 122.0-kb average insert sizes were sequenced on both ends with Sanger sequencing for a total of 591,289 Sanger reads of 402.5 Mb of high-quality sequence. Fosmid-end and BAC-end sequence data were collected using standard protocols at the HudsonAlpha Institute in Huntsville, Alabama, and at the Arizona Genomics Institute in Tucson, Arizona. Sixty P. vulgaris genotypes representing 30 wild Mesoamerican and 30 wild Andean individuals were pooled into 2 sequencing libraries, and 5×v and 4×v genome equivalences were collected on a HiSeq 2000 with unamplified libraries. Similarly, 100 genotypes from 6 individual landrace classes, selected from a structure analysis, were pooled into 6 libraries, and sequencing details from 3.4 to 7.1× were achieved.

Construction of the genetic map. We obtained 19,619 Mb of 121-bp paired-end Illumina Genome Analyzer IIx short reads from a diverse set of genotypes for common bean. Reads were aligned to the genome reference sequences for common bean with 14× coverage, and SNPs were called using CASAVA1.7 software (Illumina, 2010) with the default settings. After filtering out A/T or G/C SNPs, SNPs with Ns in the 60 nt of flanking sequence and SNPs residing within 25 nt of another SNP, a total of 992,682 SNPs remained. Using these SNPs, an Illumina Infinium BeadChip (BARCBEAN6K_1 with 5,232 SNPs) was designed. The SNPs for BARCBEAN6K_1 were selected to optimize polymorphism among the various common bean market classes, and, when possible, SNPs were targeted to sequence scaffolds (>10 kb) in an early P. vulgaris assembly. A mapping population of 267 F2 progeny from a cross of the common bean cultivars Stamped and Red Hawk developed at North Dakota State University was genotyped with the BARCBEAN6K_1 BeadChip. An additional BeadChip (BARCBEAN6K_2 with 5,514 SNPs) was designed using the same steps as with the P. vulgaris v0.9 assembly, with markers selected to anchor and orient additional scaffold sequences and used to type the same population. Both BeadChips and 261 SSR markers were also used to genotype 88 F2-derived RILs from the cross of the Stamped and Red Hawk cultivars. SSSRs were selected from sequence scaffolds in the P. vulgaris 8× assembly, PCR markers described and fragment length polymorphism were assessed as described in Song et al.66. Linkage maps were constructed using JoinMap 4.0 (ref. 47) software on the basis of 6,531 polymorphic SNPs from these 2 BeadChips and 484 SNP loci that were genotyped with the Illumina GoldenGate assay at the US Department of Agriculture–Agricultural Research Service in Beltsville, Maryland48, as well as 261 SSR markers and 25 framework markers. The final map contained 7,276 SSR and SNP markers arranged in 11 linkage groups via framework markers.

Genome assembly and construction of pseudomolecule chromosomes. Before assembly, reads corresponding to organelle DNA were removed by screening against identified fragments of mitochondria, chloroplast and rDNA. For Roche 454 linear reads, any read <200 bp in length was discarded. Roche 454 paired reads were split into pairs, and any pair with a read shorter than 50 bp was discarded. An additional deduplication step was applied to the 454 paired libraries that identified and retained only one copy of each PCR duplicate. All remaining 454 reads were compared against 24.1 Gb of trimmed HiSeq 2000 V3 reads from two separate libraries, and any insertion-deletions in the 454 reads were corrected to match the Illumina alignments. Before assembly, 454 reads that contained >80% 24-mers that occurred ≥40 times in the data set were removed to reduce improper assembly of transposon sequences. Sequence reads were assembled using our modified version of Arachne v.20071016 (ref. 49) with parameters maxclq1 = 250 and BINGE_AND_PURGE = True, best = False BINGE_AND_PURGE = True lap_ratio = 0.8 max_bad_lock = 2000 (note: Arachne error correction was on). An additional filtering step to remove contigs of <300 bp in length or with fewer than four reads was applied. This produced 1,627 scaffold sequences, with a scaffold L50 value of 6.0 Mb; 171 scaffolds were greater than 100 kb in length, and the total genome size was 474.3 Mb (Supplementary Table 2). Scaffolds were screened against bacterial proteins, organelle sequences and the GenBank nr database and were removed if found to be a contaminant. Additional scaffolds were removed if they (i) consisted of >95% 24-mers that occurred four other times in scaffolds greater than 50 kb in length, (ii) contained only unanchored RNA sequences or (iii) were less than 1 kb in length.

The 7,015 markers from the genetic map were aligned to the assembly using BLAT50 (parameters: -t = dna -q = dna -minScore = 200 –extendThroughN). Positions of SSR markers were determined using E-PCR51. Scaffolds were broken if they contained linkage group or syntenic discontinuity coincident with an area of low BAC or fosmid coverage. A total of 71 breaks were executed and 284 joins were made to form the final assembly consisting of 11 pseudomolecule chromosomes. Each chromosome join was padded with 10,000 Ns to indicate unsized map joins. The final assembly contained 708 scaffolds (41,391 contigs) that cover 472.5 Mb of the genome with a contig N50 value of 39.5 kb and a scaffold N50 value of 50.4 Mb.

Completeness of the eukaryotic portion of the genome assembly was assessed using 108,874 P. vulgaris EST sequences obtained from GenBank. These sequences were aligned to the assembly to estimate completeness using BLAT (parameters: -t = dna -q = dna –extendThroughN). Alignments that comprised ≥90% base-pair identity and ≥85% EST coverage were retained. The screened alignments indicated that 102,254 of the 108,874 CDNAS (93.92%) aligned to the assembly. At least 30% of the ESTs that did not align were bacterial or fungal contaminants. In addition, BAC clones from eukaryotic regions and moderately to highly repetitive regions were sequenced and compared to the assembly (Supplementary Figs. 19–23).

Annotation. We constructed 43,627 transcript assemblies from about 727 million reads of paired-end Illumina RNA-seq data. These transcript assemblies were constructed using PERTRAN (S.S., unpublished data). We built 47,464 transcript assemblies using PASA52 from 79,630 P. vulgaris Sanger ESTs and the RNA-seq transcript assemblies. Loci were identified by transcript assembly alignments and/or EXONERATE alignments of peptides from Arabidopsis, poplar, Medicago truncatula, grape (Vitis vinifera) and rice (Oryza sativa) peptides to the repeat-soft-masked genome using RepeatMasker53 on the basis of a transposon database developed as part of this project (see URLs) with up to 2,000-bp extension on both ends, unless they extended into another locus on the same strand. Gene models were predicted by the homology-based predictors FGENESH+ (ref. 53), FGENESH, EST (similar to FGENESH+; EST as splice-site and intron input instead of peptide/translated ORF) and GenomeScan54. The highest scoring predictions for each locus were selected using multiple positive factors, including EST and peptide support, and one negative factor—overlap with repeats. Selected gene predictions were improved by PASA, including by adding UTRs, correcting splicing and adding alternative transcripts. PASA-improved gene model peptides were subjected to peptide homology analysis with the above-mentioned proteomes to obtain Cscore values and peptide coverage. Cscore is the ratio of the peptide BLASTP score to the mutual best hit BLASTP score, and peptide coverage is the highest percentage of peptide aligned to the best homolog. A transcript was selected if its Cscore value was greater than or equal to 0.5 and its peptide coverage was greater than or equal to 0.5 or if it had EST coverage but the proportion of its coding sequence overlapping repeats was less than 20%. For gene models where greater than 20% of the coding sequence overlapped with repeats, the Cscore value was required to be at least 0.9 and homology coverage was required to be at least 70% to be selected. Selected gene models were subjected to Pfam analysis, and gene models whose encoded peptide contained more than 30% Pfam transposon element domains were removed. The final gene set consisted of 27,197 protein-coding genes and 31,638 protein-coding transcripts.
Repeat analysis. In addition to the genome sequence, 15 publicly available BAC sequences for common bean were also downloaded from GenBank for a total of 2.2 Mb of sequence, including from accession DQ205649, DQ323045, FJ817289–FJ817291 and GU315957–GU315966. Transposon annotation was conducted using different methods according to the sequence structures and transposases of various transposons. To annotate LTR retrotransposons, the genome sequence was screened with LTR_Finder using default parameters, except that we set a 50-bp minimum LTR length and 50-bp minimum distance between LTRs. All predicted LTR retrotransposons were manually inspected to eliminate incorrectly predicted sequences, including tandem repeats, nested transposons, incomplete DNA transposons and other sequences. 

The internal sequences of LTR retrotransposons were used to perform BLASTX and/or BLASTP searches to define superfamilies: Ty1-copia, Ty3-gypsy or other. LINEs (long interspersed elements) were predicted on the basis of the non-LTR retrotransposase and polyA sequences. SINEs (short interspersed elements) were annotated with the polyA structure feature and combined with BLAST searches. To find DNA transposons, conserved domains for transposases from different reported superfamilies were used as queries to search the common bean genome. The matching sequences and flanking sequence (10 kb on each side) were extracted to conduct BLASTN searches to identify complete DNA transposons by terminal inverted repeats (TIRs) and target size duplication (TSD). Furthermore, MITEs-Hunter software was also used to identify DNA elements. The annotated transposons and two reported LTR retrotransposons, pva1-118424-re-5 (FJ102927) and Tpv2-6 (AJ085762), were combined and used as a transposon library to screen the genome using RepeatMasker with default settings except that we used the ‘nolow’ option to avoid masking low-complexity DNA or simple repeats. Transposons were summarized according to names, subclasses and classes, and overlapping regions in the RepeatMasker output file were counted on (Supplementary Table 9).

To estimate the insertion times of LTR retrotransposons, the 5’ and 3’ LTRs for each full-length LTR retroelement were aligned and used to calculate the nucleotide divergence rate with the Kimura-2 parameter using MEGA 4. The insertion date \( T \) was estimated with the formula \( T = K/2r \), where \( K \) is the average number of substitutions per aligned site and \( r \) is an average substitution rate. We used the average substitution rate of 1.3 \( \times 10^{-9} \) substitutions per synonymous site per year to calibrate the insertion times.

Identification of disease resistance genes. NL proteins were identified in an iterative process. First, an HMM (Hidden Markov model) search of the predicted protein sequences identified sequences containing the NB-ARC domain. The ‘trusted cutoff’ of the NB-ARC domain HMM (PF00931) established by Pfam was used as the threshold for detecting NBs domains. We identified 399 predicted proteins corresponding to 342 annotated genes that encoded homologs of NL proteins. To identify diverse homologs, all the NL predicted protein sequences were used as queries for TBLASTN against the entire genome. All resulting sequences (\( E \) value < 1 \( \times 10^{-10} \)) were manually inspected using Artemis. This procedure identified an additional 38 putative NL genes that were not part of the genome annotation. A new identifier was created for each missing gene (with last digits set as 50). NL genes were assessed manually in Artemis software for the presence of sequences encoding TIR (PF01582), NB-ARC (PF00931) and LRR (PF00560, PF07273, PF07725, PF12799, PF13306, PF13516, PF13504 and PF13855) domains with HMMer using the trusted cutoffs defined in Pfam. Coiled-coil domains were identified using Coils with a 14-amino-acid search window and a cutoff score of 2.9. Artemis was used for further manual analysis. Gene models with stop codons and/or frameshifts were classified as pseudogenes.

Development of wild and landrace pools for sequencing of common bean. Initially, 126 wild and 179 landrace genotypes, collected from the full geographic range of the species, were scored with 22 indel markers distributed throughout the genome. A Bayesian analysis was performed on the nucleotide data within each of the two groups using STRUCUTURE software with the parameters outlined previously. For the wild genotypes where \( k \) is the number of populations, \( k = 2 \) best fit the data, and, for the landraces, \( k = 6 \) defined 3 Mexican subpopulations, 1 Central American subpopulations and 2 Andean subpopulations. A genotype was assigned to a subpopulation if its subpopulation parentage was >70%. DNA pools for resequencing were created by selecting individuals with high subpopulation membership (>98% for wild subpopulations and >90% for landrace subpopulations; Supplementary Fig. 18). In adopting other approaches, several individual-pool SNP data were combined with other pool SNP data to create a pool SNP data set representing a putative ancestral state.

Pooled DNA sequencing and SNP identification. DNA from each of these pools was sequenced to ~4× depth using Illumina technology (Supplementary Table 12). Each read was mapped to the v1.0 version of the assembled reference genome using Burrows-Wheeler Aligner (BWA) with the maximum edit distance set to 8. All reads with a mapping quality score of less than 25 were discarded. An mpileup file was created for each sequenced pool using SAMtools with the –BA options. VarScan 2.2.10 (ref. 66) used the mpileup file for SNP calling with the following parameters: minimum coverage = 5, minimum consensus quality = 25 and minimum variant frequency = 0.01. To further reduce SNP call quality, SNPs were discarded (i) if the reference or variant allele was an N; (ii) if more than one variant allele was observed; and (iii) if the variant allele was a single-nucleotide indel. The minimum number of reads required for the reference or variant allele was three. The number of SNPs ranged from 8,690,318 for the wild Mesoamerican pool to 1,397,405 for the Peru landrace pool (Supplementary Table 14). Among wild genotypes, 10,158,326 SNPs were observed, whereas the Mesoamerican landrace genotypes contained 9,661,807 SNPs and the Andean landrace genotypes contained 3,154,648 SNPs. For individual and combined pools, the proportion of SNPs found within genes was ~16%, indicating that genes were not disproportionately prone to more (or less) variation.

Demographic modeling. To minimize bias in demographic inferences due to selection, we used neutral sites defined to be at least 5 kb away from a gene (as annotated in the gff3 file v1.0) and not located in repetitive regions. The number of different haplotypes for each pooled sample was close to 30. Data were thus down-sampled to 25 haplotypes for each pool via hypergeometric projection (random sampling of 25 alleles without replacement), from which the joint allele frequency spectrum (\( jAFS \)) was derived. To eliminate spurious singleton, we excluded sites appearing as singletons in either of the two pools, resulting in a total of 663,000 polymorphic sites for \( jAFS \).

We compared different demographic models on the basis of the relative log likelihoods of the models given the observed site frequency spectrum. Asymmetric migration rates were assumed in the model (Fig. 1). To infer model parameters, we ran \( \delta \) simulations with different starting points in an eight-dimensional parameter space until convergence was achieved. Parameter values for the best-fit model are listed in Supplementary Table 13, using a base substitution rate \( \mu = 8.46 \times 10^{-9} \) substitutions/bp/year (S.B.C., unpublished data) derived from silent sites. To estimate parameter uncertainties, we divided the genome into 10-cM segments and performed 100 bootstraps on the chromosome segments. Confidence intervals were derived on the basis of simulation results for the bootstrapped samples (Supplementary Table 13) as were comparisons between model prediction and observed data (Supplementary Figs. 24 and 25).

Population genetics statistics. Several population genetics statistics were calculated in 100-kb/10-kb and 10-kb/2-kb sliding windows and each gene in each DNA pool. Any window or gene with >50% Ns was excluded, and all statistics were based on the number of non-N nucleotides in the window. Nucleotide diversity (\( \pi \)), the average number of nucleotide differences per site between two DNA sequences chosen randomly from the sample population; ref. 67 was calculated using the following formula:

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i x_j}{2n^2} \delta_{ij}$$

Here \( x_i \) and \( x_j \) are the respective frequencies of the \( i \)th and \( j \)th sequences, \( \pi \) is the number of nucleotide differences per nucleotide site between the \( i \)th and \( j \)th sequences, and \( n \) is the number of sequences in the sample. The Watterson estimate (\( \Theta_w \); ref. 68), which is an estimation of population
 mutations rate, was calculated on the basis of the number of segregating sites using the formula
\[ \theta_S = \frac{S}{a_n} \]
where \( S \) is the number of segregating sites and
\[ a_n = \sum_{i=1}^{n-1} i \]
Tajima's \( D \), calculated as described in ref. 69. \( F_{ST} \) (ref. 70) is a measure of population differentiation estimated from the average pairwise differences between chromosomes in each analysis panel compared to the combined samples as described in ref. 71
\[ F_{ST} = 1 - \frac{2 \eta j \sum n_j \eta j - 1 x_j (1 - x_j)}{\sum j (\eta j / 2)} - \frac{\sum 2 \eta j \eta j - 1 x_j (1 - x_j)}{\sum j (\eta j / 2)} \]
where \( x_j \) is the estimated frequency of the minor allele at SNP \( j \) in population \( j \), \( n_j \) is the number of genotyped chromosomes at that position and \( 1 \) is the number of chromosomes analyzed in that population. The lack of the \( j \) subscript in the denominator indicates that statistics \( n_j \) and \( x_j \) are calculated across the combined data sets.

The relative diversity among two pooled samples was compared by a nucleotide diversity \( \pi \), using the formula
\[ \pi = \frac{4N_e \theta}{2N_e + 1} \]
where \( \theta = \sum x_j \sum n_j \eta_j - 1 x_j (1 - x_j) / \sum j (\eta_j / 2) \).

Identifying selected windows and genes and defining sweep windows. A composite scoring system was used to determine whether a 10-kb/2-kb sliding window or gene window was under selection. This approach is similar to the one applied for silk moth where a reduction in nucleotide diversity and Tajima's \( D \) was applied to discover domestication-related genes\(^72\). Here a 10-kb/2-kb window or a gene was considered a selection window or domestication candidate gene if it was in the upper 90% of the pool's empirical distribution for the \( \pi \) or \( F_{ST} \) statistics. The cutoff values for various comparisons can be found in Supplementary Table 3. All 10-kb/2-kb selection windows within 40 kb of each other were merged in a 'sweep window'. The numbers of domestication candidates and total genes were calculated for each sweep window.

Annotating candidates for seed weight and size in common bean. We used the Arabidopsis protein sequence for all genes found to be associated with seed weight\(^43\) as queries for a BLASTP analysis of a database of the common bean proteins. We identified 141 common bean gene models with 50% identity and 80% coverage that matched 70% of the query length, and these inherited the Arabidopsis names for the gene associated with seed weight.

Association mapping. In total, 271 diverse modern common bean varieties from the Mesoamerican and Andean gene pools were grown in replicated field trials by the Nebraska and Colorado State University bean breeding programs. Each variety from the Mesoamerican gene pool were grown in replicated field trials by the Nebraska and Colorado State University bean breeding programs. Each variety was genotyped with 34,799 SNPs. Missing data were imputed in fastPHASE 1.3.

The linkage disequilibrium \( r^2 \) between all marker loci was calculated in PLINK\(^9\) using a minor allele frequency of 0.1. The EMMA kinship matrix and the GWAS were calculated in the genome association and prediction integrated tool (GAPIT) package in R\(^79\), without P3D and compression. Only markers with minor allele frequency of 0.1 or greater were considered in the GWAS results. Protein sequences for Arabidopsis genes associated with seed weight\(^73\) were used as queries for a BLASTP analysis against a database of common bean proteins. We identified 141 common bean gene models with 50% identity and 80% coverage that matched 70% of the query length, and these inherited the Arabidopsis gene names.

46. Song, Q. et al. Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCOSYTRSR 1.0) in soybean. Crop Sci. 50, 1950–1960 (2010).
47. Van Ooijen, J. JoinMap 4. Software for the Calculation of Genetic Linkage Maps in Experimental Populations (Kyazma, Wageningen, The Netherlands, 2006).
48. Hyten, D.L. et al. High-throughput SNP discovery and assay development in common bean. BMC Genomics 11, 475 (2010).
49. D.B. et al. Whole genome assembly for mammalian genomes: Arachne 2. Genome Res. 13, 91–96 (2003).
50. Kent, W.J. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).
51. Schulier, G.D. Sequence mapping by electronic PCR. Genome Res. 7, 541–550 (1997).
52. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignement assemblies. Nucleic Acids Res. 31, 5654–5666 (2003).
53. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in Drasophila genome. Nucleic Acids Res. 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 Res. 11, 803–816 (2001).
55. Ma, J. & Bennett, J.L. Rapid recent growth and divergence of rice nuclear genomes. Proc. Natl. Acad. Sci. USA 101, 12404–12410 (2004).
56. Finn, R.D. et al. The Pfam protein families database. Nucleic Acids Res. 38, D211–D220 (2010).
57. Allschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997).
58. Rutherford, K. et al. Artemis: sequence visualization and annotation. Bioinformatics 16, 944–945 (2000).
59. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. Science 252, 1162–1164 (1991).
60. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587 (2003).
61. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945–959 (2000).
62. McLean, P.E. et al. Population structure and genetic differentiation among the USDA common bean (Phaseolus vulgaris L.) core collection. Genet. Resour. Crop Evol. 59, 499–515 (2012).
63. Eswaran, G., Begnout, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611–2620 (2005).
64. Li, H., Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
65. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
66. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012).
67. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595 (1989).
68. Watterson, G.A. On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. 7, 256–276 (1975).
69. International HapMap Consortium. A haplotype map of the human genome. Nature 437, 1299–1305 (2005).
70. Xia, Q. et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). Science 326, 433–436 (2009).
71. Kesanav, M., Song, J.T. & Seo, H.S. Seed size: a priority trait in cereal crops. Physiol. Plant. 147, 113–120 (2013).
72. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phases. Am. J. Hum. Genet. 78, 629–644 (2006).
73. SAS Institute, Inc. SAS 9.3 Language Reference: Concepts, Second Edition (SAS Institute Inc., Cary, NC, 2012).
74. Yang, J., Lee, M.-J. et al. Unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Natl. Genet. 38, 203–208 (2006).
75. Wang, H.M. et al. Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723 (2008).
76. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
77. Lipka, A.E. et al. GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397–2399 (2012).