Title

A high-quality genome assembly and integrative data portal (*Phabase*) for the Mesoamerican black bean (*Phaseolus vulgaris* cv. Negro Jamapa)

Author list

Turgut Y. Akyol[2*], Eber D. Villa-Rodriguez[2*], Heladia Salgado[1], Emanuel Pacheco[1], Nancy Trujillo[1], Lavinia I. Fechete[2], Stig U. Andersen[2], Damien Formey[1], and Jesús Montiel[1]

[*]These authors contributed equally to the study.

Author affiliations

[1]: Department of Molecular Biology and Genetics, Aarhus University, 8000, Denmark

[2]: Centro de Ciencias Genómicas. UNAM. Cuernavaca, Mexico

Author emails

Turgut Y. Akyol: tya@mbg.au.dk

Eber D. Villa-Rodriguez: eber.villa@mbg.au.dk

Heladia Salgado: heladia@ccg.unam.mx

Emanuel Pacheco: emanuelpachecoalberto@gmail.com

Nancy Trujillo: nancy.trujillo@ibt.unam.mx

Lavinia I. Fechete: lif@mbg.au.dk

Stig U. Andersen: sua@mbg.au.dk

Damien Formey: formey@ccg.unam.mx (correspondence)

Jesús Montiel: jmontie@ccg.unam.mx (correspondence)

Short running head

Genome & data portal for Negro Jamapa bean

Abstract

Common bean (*Phaseolus vulgaris* L.) is one of the major legume crops worldwide, constituting a fundamental part of the diet in several countries. In central and south Mexico, the Mesoamerican Negro Jamapa variety is one of the most widely consumed cultivars, and it has notably served as a working model for functional genomics to address the molecular responses to stress and during the nitrogen-fixing symbiosis with rhizobia. Despite this, the genome sequence of this landrace is not yet available, and the transcriptomic data generated by different research groups on this species are scattered. This prompted us to carry out a *de novo* sequencing, assembly and gene prediction for Negro Jamapa genome, using Hi-Fi PacBio technology and advanced genomic tools, respectively. Herein, we present a high-quality genome assembly for Negro Jamapa at chromosome level. The assembled genome is 522 Mb length, with an N50 of 45Mb, 89X coverage and BUSCO completeness of 98.4%, parameters that are considerably better than the available Mesoamerican reference genome. In parallel, we built an expression atlas of *P. vulgaris* from various tissues and growth conditions, by re-mapping publicly available RNA-seq and small RNA-seq data produced in recent years by distinct research groups. This analysis is available for the scientific community in the new user-friendly portal developed in this study, *Phabase* (https://*phabase*.ccg.unam.mx/). This portal includes several tools such as BLAST, gene and microRNA expression atlas, JBrowse, download section and integration with other *Phaseolus* spp. genomes.

Introduction

Legumes are essential sources of protein and micronutrients for human nutrition. Among all, the common bean (*Phaseolus vulgaris*) is the most important crop across Latin America, with two major domesticated gene pools: Mesoamerican and Andean [1,2]. Common beans diverged into these two principal gene pools, approximately 0.1 to 0.2 million years ago [3]. The corresponding species were domesticated independently around 8,000 years ago [2,3] and show strong differences in their seeds, their adaptation range, and their resistance to abiotic and biotic stresses. The consequence of this dual domestication is an extensive genomic differentiation, such as lineage-specific alleles and structural variants, responsible for key agronomic traits [3,4].

The Negro Jamapa cultivar, a Mesoamerican variety, was developed in 1957 by the Mexican Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP) through a structured plant breeding program, becoming a staple food across diverse regions of Mexico and serving as a principal source of protein and dietary fiber [5]. The resilience, adaptability, and combined nutritional and ecological qualities of Negro Jamapa have made it a key focus in agricultural research and breeding programs directed at improving yield stability, disease resistance, and nutritional quality [6,7], and also a major contributor to food security and cultural identity in the region.

The chromosome-scale reference genome for common bean was first generated for the Andean landrace G19833 [2]. This was followed by the assembly of the BAT93 Mesoamerican cultivar genome [8], unfortunately still not at the chromosome level to date, with detailed transcriptomic profiles from multiple tissues and developmental stages, regrettably not easily accessible by the bioinformatic skill-lacking common bean community members. In addition to the common bean genotypes originating from Mesoamerican or Andean gene pools, whole-genome assembly of a European *P. vulgaris* accession was released [9]. More recently, to identify genomic regions that contribute to flowering responses in European cultivars, whole-genome resequencing was performed on 232 *P. vulgaris* accessions, mainly from Europe [10]. Furthermore, the first telomere-to-telomere (T2T) genome assembly of the YP4 cultivar has been completed using PacBio High-Fidelity reads, ONT ultra-long sequencing, and Hi-C technologies, resulting in an assembly with exceptional completeness (BUSCO score: 99.5%) [11]. Despite these advancements, existing genomic data remain fragmented or limited in annotation consistency. The genomic diversity of elite cultivars, particularly those widely used in breeding programs or as model genotypes in physiological studies, like Negro Jamapa, remains underrepresented.

Model legumes such as soybean, *Medicago truncatula*, and *Lotus japonicus* benefit from dedicated, integrative web portals, which unify genome assemblies, annotations, mutant resources, and expression data within a single platform [12,13]. Among others, these portals provide several tools such as genome browsers for visualizing genomic features at multiple scales, BLAST and sequence search tools for similarity analysis, synteny viewers for comparative genomics, gene expression atlases showing tissue-specific and condition-specific patterns. For common bean, there is still no equivalent, cultivar-aware environment that links a high-quality Mesoamerican genome with a comprehensive, re-analysed expression atlas and modern visualization tools. At the same time, the growing volume of transcriptomic, proteomic, and epigenomic data [14,15,16] has created an urgent need for integrative bioinformatic platforms that allow comprehensive exploration of gene expression and regulatory networks in *P. vulgaris*. This gap is particularly limiting for Negro Jamapa, cultivar from which much of the functional genomics work and the RNA-seq data production for common bean has been performed. However, the absence of a tailored reference genome forces researchers to project data onto non-Mesoamerican assemblies, leading to less accurate and efficient sequence-based analysis experiments.

Previous efforts, such as gene expression atlases, provided valuable insight but were constrained by old sequencing technologies, lack of biological replicates, and outdated genome annotations [17]. A unified, user friendly, and dynamically updated resource is essential to fully exploit the wealth of omics information and to support data-driven hypothesis generation in this crop. To address this gap, here, we present the first high-quality genome assembly of the Negro Jamapa cultivar using long-read

sequencing, complemented with a comprehensive genome annotation. Further, we developed an integrative online portal, *Phabase* which hosts the corresponding assembled genome, gene annotations, and associated expression data. The portal includes interactive tools such as a genome browser, BLAST search, and an expression atlas of genes and microRNAs. Our genome assembly is based on PacBio HiFi sequencing, which enables highly contiguous and accurate assemblies. Leveraging this technology for Negro Jamapa will offer the opportunity for precise research in the crop's comparative and functional genomics, genomic features relevant to symbiosis, biotic and abiotic stress responses, and yield-related traits.

Results

Assembly and quality metrics

The raw data of the whole genome sequencing of common bean Negro Jamapa consisted of 4,487,410 HiFi circular consensus sequencing (ccs) DNA reads with an average length of 21,232 bp and 90.77% of the bases had a quality score greater than 30. The size of the assembled genome is 522,239,992 bp with a GC content of 36.31%, 89X mean coverage and 98.4% completeness according to BUSCO. In addition to the 11 chromosomes and the mitochondrial and chloroplast genomes **(Table 1)**, the assembly includes 1,407 unplaced contigs collectively spanning 131,292,346 bp.

Table 1. *P. vulgaris* cv. Negro Jamapa genome assembly and annotation statistics.

Identifier	Molecule	Size	GC%	Transcripts
PvNJ1.1_chr1	Chromosome	56,338,280	35.85	2780
PvNJ1.1_chr2	Chromosome	49,825,282	33.77	3392
PvNJ1.1_chr3	Chromosome	50,421,212	33.92	3052
PvNJ1.1_chr4	Chromosome	48,781,969	36.48	1718
PvNJ1.1_chr5	Chromosome	40,291,510	36.52	1724
PvNJ1.1_chr6	Chromosome	35,918,269	34.50	2383
PvNJ1.1_chr7	Chromosome	41,572,620	33.50	2944
PvNJ1.1_chr8	Chromosome	56,415,169	35.89	2904
PvNJ1.1_chr9	Chromosome	40,546,796	32.48	2778
PvNJ1.1_chr10	Chromosome	44,913,661	37.53	1689

PvNJ1.1_chr11	Chromosome	56,669,149	37.48	2047
PvNJ1.1_mito	Mitochondrial	395,789	45.11	80
PvNJ1.1_chloro	Chloroplastic	150,286	35.44	170
	Whole genome	522,239,992	36.31	30731

Comparative Genome Assembly and Annotation Metrics

The genome assembly of *P. vulgaris* cv. Negro Jamapa showed important improvement in contiguity and completeness compared with the reference genotype G19833 [2]. The total size of the assembly was 522 Mb, which is close to the 531 Mb of G19833, but with a significantly higher N50 of 45 Mb instead of only 2 Mb in G19833. Sequencing coverage was also slightly higher in Negro Jamapa (89x) compared to G19833 (83x), supporting the high quality of the assembly. This improvement was mainly due to the use of long-read sequencing and better assembly strategies, allowing to reach almost chromosome-scale continuity.

We identified 27,635 protein-coding genes in the annotation of Negro Jamapa, similar to the 27,433 genes in G19833. BUSCO analysis showed completeness values of 98.4% for Negro Jamapa, slightly improved compared to the 97.7% estimated in G19833, confirming that both assemblies are almost complete and of excellent quality. In addition, the new annotation of Negro Jamapa contained more microRNA precursors annotated (270 vs. 226), which suggests a larger repertoire of small RNAs or a consequence of the higher number of publicly available sRNA libraries generated from Negro Jamapa samples. Together, these data indicate that the Negro Jamapa genome is an improved resource, with high contiguity, completeness, and annotation depth, useful for comparative and functional genomics studies in *P. vulgaris*.

Genomic landscape and structural features

To better understand how the Negro Jamapa genome is organized, we analyzed the distribution of multiple genomic features across its eleven chromosomes **(Figure 1)**. The distribution of repetitive elements was uneven: certain chromosomes showed clear concentrations near their pericentromeric regions, consistent with the heterochromatic nature of these areas. The arrangement of annotated protein-coding genes revealed alternating regions rich and poor in genes, a pattern typical of the euchromatic and heterochromatic domains found in other legume genomes. When examining potential paralogous relationships and segmental duplications, we found that these events were particularly frequent among chromosomes 1, 8, and 11, hosts of more than 38% of the cumulated

duplication length. This observation is consistent with our structural variation analysis, which identified several large translocations and highly diverged duplicated regions.

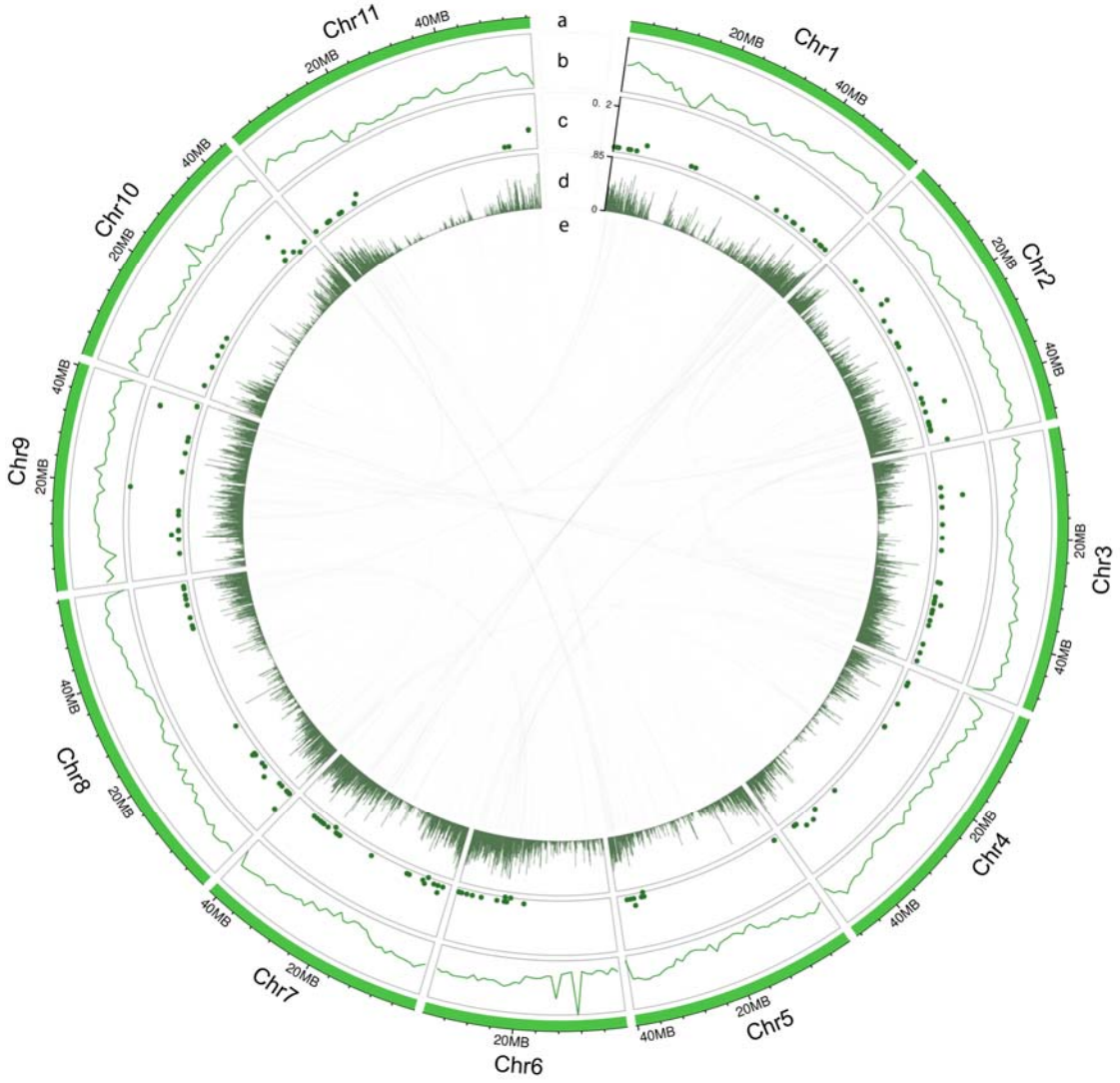

Figure 1. Circos plot illustrating the genomic landscape and structural features of the *P. vulgaris* Negro Jamapa chromosomes. Track (a) displays chromosomal ideograms, arranged sequentially and scaled according to chromosome length. Track (b) represents the distribution of repetitive elements (green lines). Track (c) shows the genomic positions of micro RNA (miRNA) precursor genes (dark green points). Track (d) corresponds to annotated protein-coding genes density (vertical green lines). (e) Colored inter-chromosomal links connect pairs of potentially duplicated genes, representing putative segmental duplications or paralogous relationships.

Overall, these results offer a comprehensive view of the Negro Jamapa genome and illustrate how repetitive sequences, gene organization, and large-scale duplications have shaped the evolution of this *P. vulgaris* cultivar.

Structural Variation Between Negro Jamapa and G19833

To contextualize the quality of the Negro Jamapa assembly and identify cultivar-specific structural features relevant for downstream functional genomics, we performed a comparative analysis against the widely used G19833 reference genome. Whole-genome alignment revealed high levels of structural and sequence variation, showing the deep divergence between the two main gene pools of common beans **(Figure 2A)**. In the aligned regions, the SNP rate was estimated at 7.6 per kilobase, which agrees with previous reports about strong nucleotide differentiation between Mesoamerican and Andean lineages. However, structural variants represented the majority of the genomic differences, with several categories detected **(Figure 2B)**. Among them, highly diverged regions represented the largest proportion (23.4%) of the polymorphic DNA, followed by unaligned regions in the query genome (23.2%) and inter-chromosomal translocations (20.6%). Inversions accounted for 12.7%, while copy number variants, including copy gains (6.2%) and segmental duplications (5.4%), represented together more than 11%. Smaller fractions were related to insertions (5.2%) and tandem repeats (0.6%).

In general, these results indicate that large structural changes and highly diverged genomic regions are the main components of divergence between Negro Jamapa and G19833, more than single-nucleotide changes. Such extensive structural differences probably contribute to reproductive isolation, local adaptation, and the phenotypic differences between these two main *P. vulgaris* gene pools.

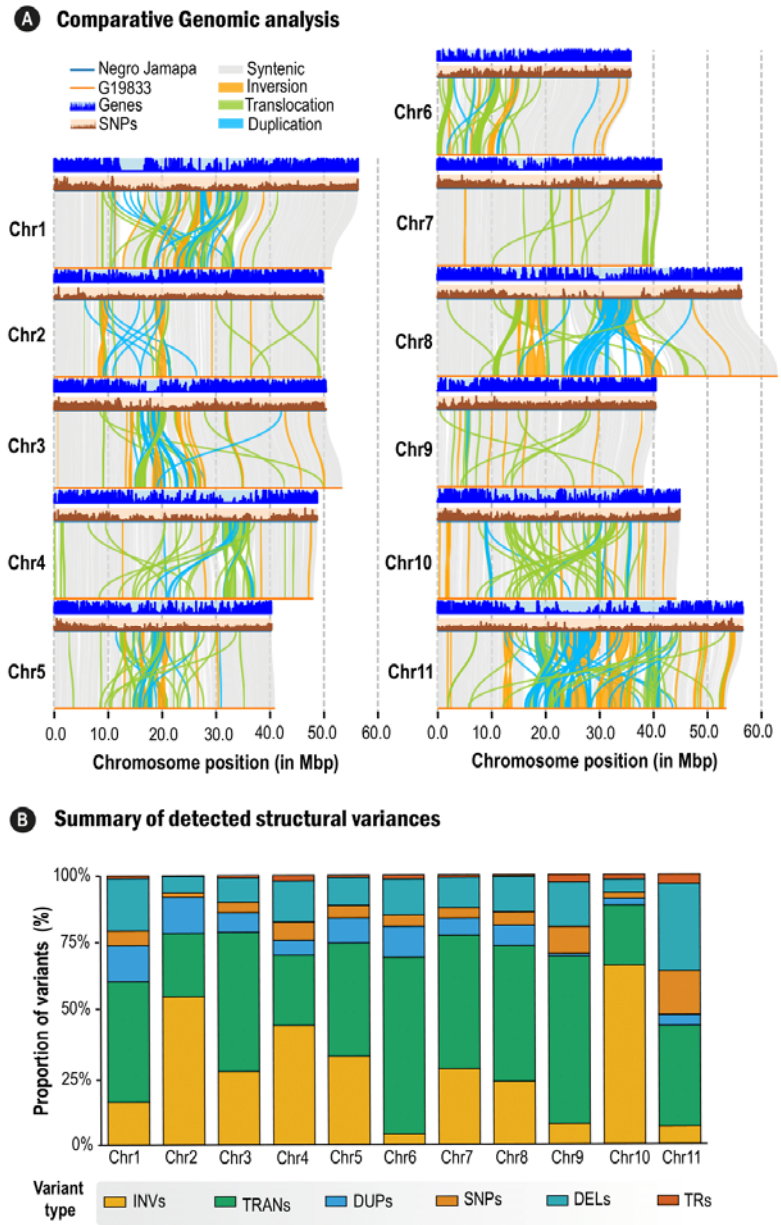

Figure 2. Genome-wide distribution of structural variants between *P. vulgaris* Negro Jamapa and G19833 genomes. (A) Comparative genomic analysis. For each chromosome, the upper track displays the distribution of SNPs (blue) and the lower track represents annotated genes (brown). Vertical links between chromosomes indicate structural variants identified by SyRI and visualized with plotsr, including inversions (yellow), translocations (green), and duplications (light blue). Syntenic regions are represented in grey. Vertical dashed lines indicate the relative chromosome position (Megabases). **(B)** Summary of detected structural variants. Chromosomal representation of the cumulated length proportion of different classes of structural variants identified across the 11 chromosomes of *P. vulgaris* Negro Jamapa. Each bar indicates the relative abundance (0–100%) of variant types per chromosome. INVs: inversions; TRANs: translocations; DUPs: duplications; SNPs: single nucleotide polymorphisms; DELs: deletions; TRs: tandem repeats. Structural variants were detected using SyRI (Synteny and Rearrangement Identifier) based on pairwise genome alignment.

Phabase: an integrative database for Phaseolus spp.

Phabase was implemented as a modular web system designed to integrate genomic and transcriptomic information from *P. vulgaris*. The architecture follows a three-tier model comprising the database layer, web services layer, and presentation layer, each deployed in isolated Docker containers to ensure scalability and maintainability **(Figure 3)**. Overall, *Phabase* brings together genomic and transcriptomic tools within a streamlined, user-friendly workflow **(Figure 4)**. The platform provides open access to multiple datasets, including *P. vulgaris* G19833 v2.1 [2], *P. vulgaris* BAT93 v1.0 [8], and the newly assembled *P. vulgaris* cv. Negro Jamapa v1 generated in this study (**Figure 4A**). It is an intuitive platform for exploring genes and retrieving both nucleotide and amino acid sequences from any of the available genomes **(Figure 4B)**. Additionally, *Phabase* integrates the Basic Local Alignment Search Tool (BLAST), enabling users to identify orthologous sequences across the available genomes using either nucleotide or amino acid queries **(Figure 4C)**.

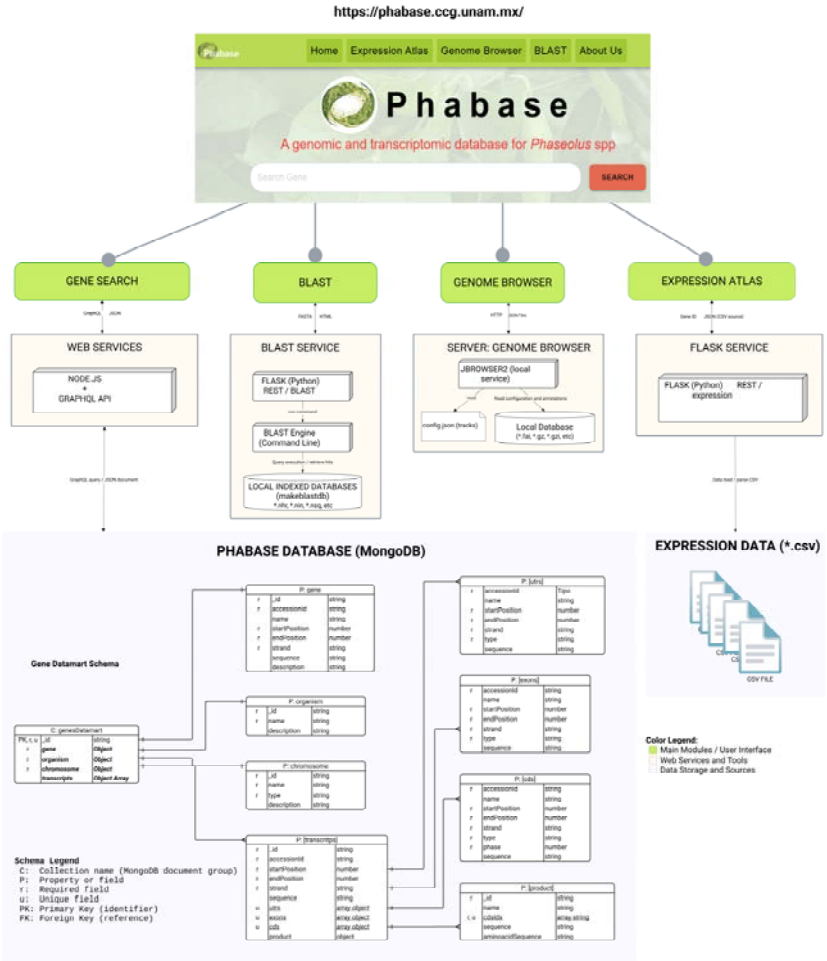

Figure 3. System architecture of *Phabase*. The diagram illustrates the modular architecture of the *Phabase* platform, showing the interaction between the main user modules, web services, and data sources. Green boxes represent the user interface modules, orange boxes correspond to web services and computational tools, and blue boxes denote data storage and sources.

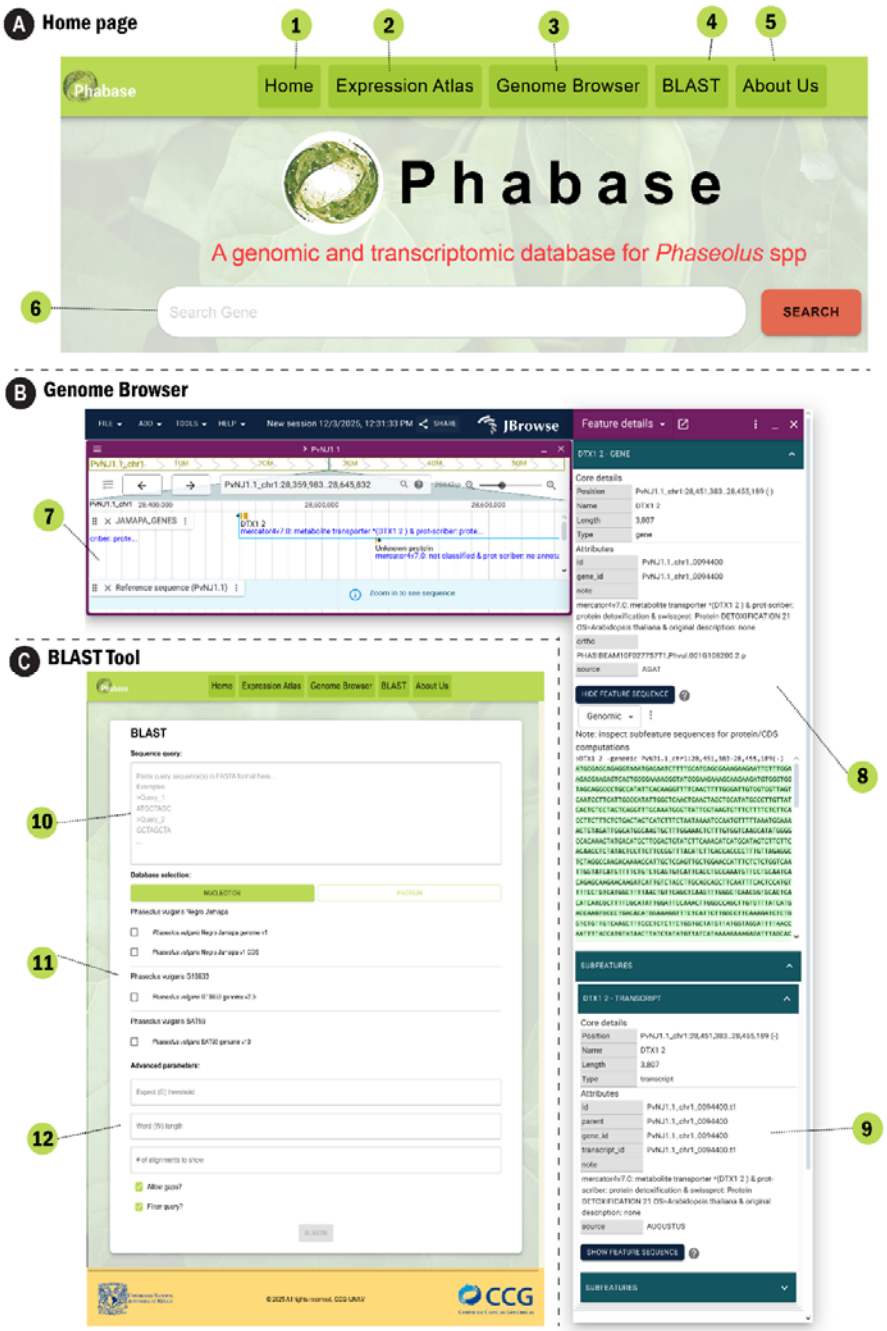

Figure 4. Overview of *Phabase* tools and interface. (A) Home page of *Phabase* showing the main navigation menu. *Home* (1), *Gene Expression* (2), *Genome Browser* (3), *BLAST* (4), and *About Us* (5) sections. The search bar (6) allows users to query genes of interest directly. **(B)** Genome Browser interface illustrates the visualization of genomic regions. Genes are represented as annotated tracks and users can inspect gene models, annotations, and neighboring genomic elements (7). Selecting a feature opens a detailed information panel (8), which includes gene coordinates, structure, functional annotation, and nucleotide sequences. Additional transcript-level annotations and predicted subfeatures (9) can be explored through collapsible sections. **(C)** BLAST interface allows users to submit a nucleotide or protein query sequence in FASTA format (10) and search against multiple *Phaseolus* reference genomes selectable from the database list (11). Optional advanced parameters (12), including E-value threshold, word size, number of alignments to display, and filtering options, enable customized similarity searches.

For the microRNA expression atlas, 284 small RNAs have been selected. Almost all of them are microRNAs (274), including 154 non-common bean-specifics and 120 only identified in *P. vulgaris* genome [15,18]. The other ten small RNAs are tRFs, transfer RNA (tRNA)-derived small RNA fragments from rhizobia [19]. In this atlas, we gather data from 30 experimental conditions. The most common topics were focused on plant-microbe interactions, particularly on symbiotic interactions (30%) and pathogen responses (20%) **(Figure 5B)**. In this case, the datasets come from 11 cultivars, including Negro Jamapa, Flavert, and Pinto Villa (each one representing around 13% of all samples) **(Figure 5B)**. Expression profiles were obtained from eight organs or tissues, mainly leaves (27%), seeds (27%), and roots (20%) **(Figure 5B)**. These studies were produced by projects led by research groups from seven countries, with an important contribution from Mexico (33%), reflecting a leading role in small RNA studies in beans **(Figure 5D)**. Altogether, this atlas represents the most complete and diverse collection of *P. vulgaris* expression data so far. It offers a strong base for comparative studies, helps to identify regulatory networks, and facilitates the analysis of gene/microRNA interactions across different cultivars, organs, and environmental conditions.

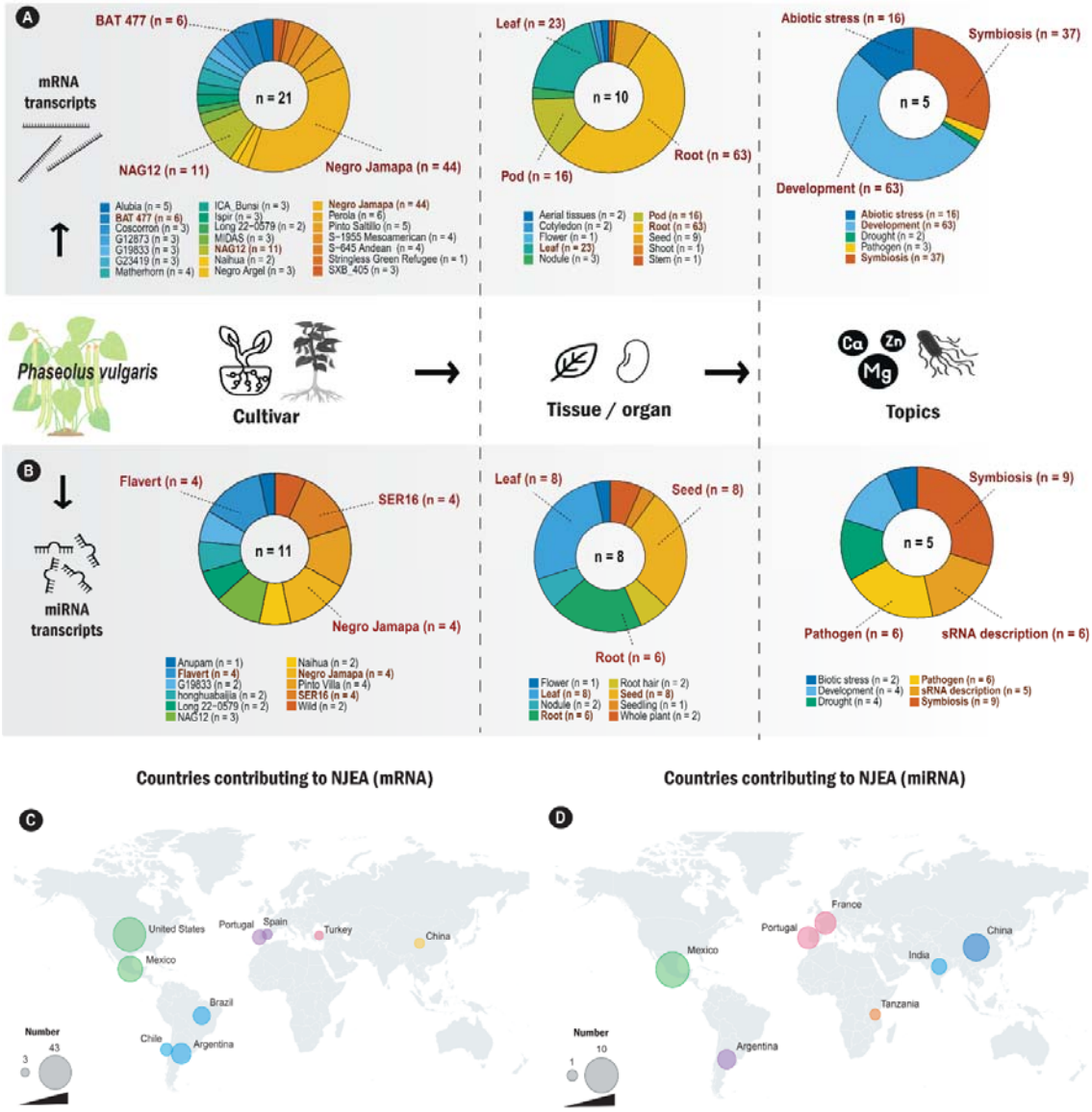

Figure 5. Overview of the Expression Atlas module in *Phabase*. Summary of the datasets integrated into *Phabase* for **(A)** Gene and **(B)** microRNA Expression interface. Charts display the number of samples categorized by cultivar, tissue/organ, and topic. The number at the center of each chart indicates the total number of elements represented in the corresponding chart. Geographic distribution of the research groups that generate datasets that were processed to obtain the **(C)** Gene and **(D)** microRNA Expression module. The diameter of a circle is proportional to the number of conditions produced in the corresponding country. Colors differentiate regions of the world.

Exploring MYB36 orthologs in P. vulgaris Negro Jamapa using Phabase

To illustrate how *Phabase* can be applied by typical end users to explore gene function and generate data-driven hypotheses, we selected MYB36 as a case study (**Figure 6A**). This is a transcription

factor recently characterized for its role in Casparian strip formation and its indirect influence on microbial colonization dynamics in *Arabidopsis thaliana* [20].

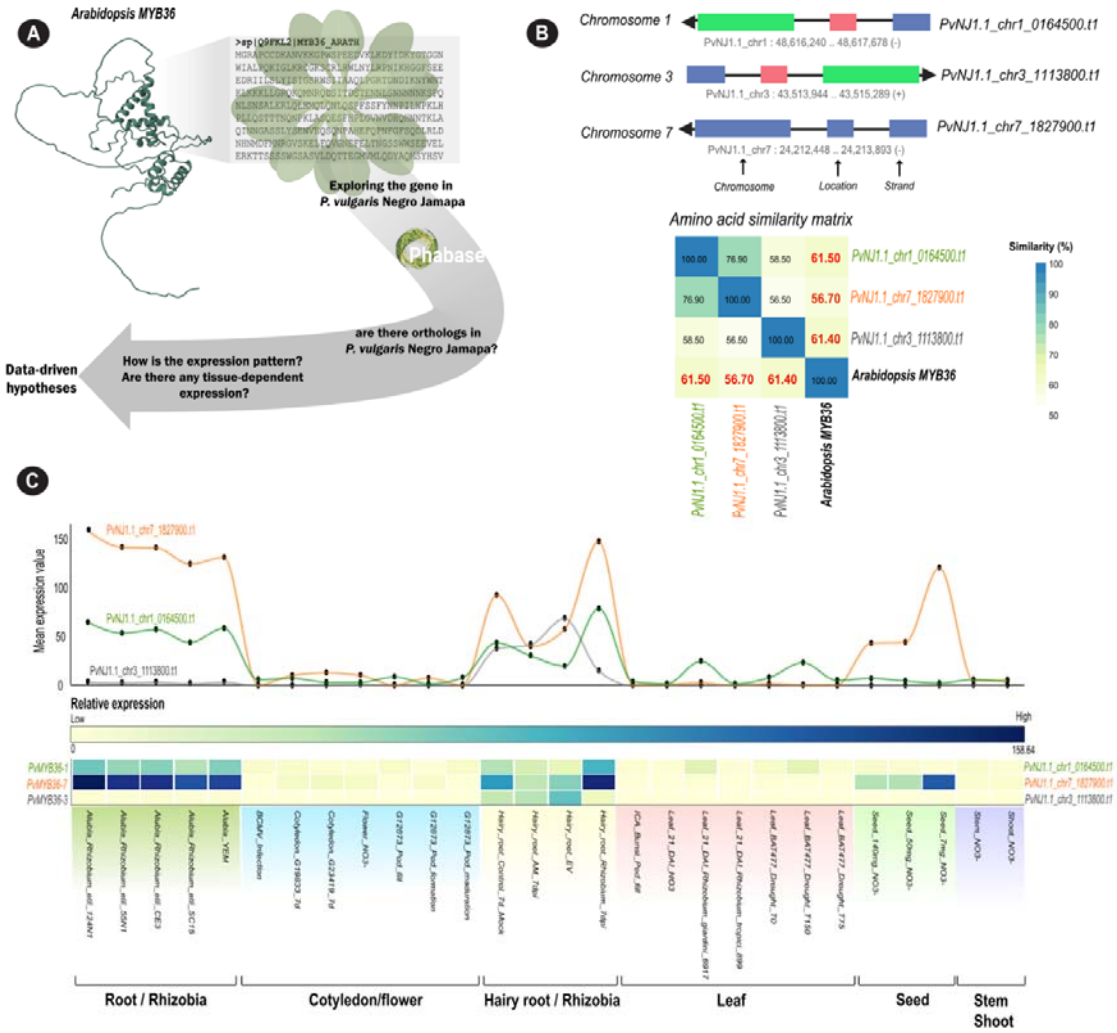

Figure 6. *Phabase*-enabled identification of MYB36 orthologs in *P. vulgaris* cv. Negro Jamapa. (A) Workflow illustrating how *Phabase* was used to identify MYB36 orthologs in *P. vulgaris* cv. Negro Jamapa. **(B)** Genomic locations and amino acid similarity matrix of the putative MYB36 orthologs. **(C)** Expression profiles of the identifiedMYB36 orthologs across different tissues and conditions.

We first investigated whether orthologous sequences of *AtMYB36* are present in the *P. vulgaris* cv. Negro Jamapa genome. To do so, we retrieved the *A. thaliana* MYB36 amino acid sequence from the UniProt database and used it as a query in the *Phabase* BLASTP tool against the *P. vulgaris* Negro Jamapa v1 protein dataset. This search identified three candidate sequences with E-values below 4e-31: PvNJ1.1_chr7_1827900.t1 (Identities = 55/60, 92%; Positives = 59/60, 98%; Gaps = 0%),

PvNJ1.1_chr3_1113800.t1 (Identities = 54/60, 90%; Positives = 57/60, 95%; Gaps = 0%), and PvNJ1.1_chr1_0164500.t1 (Identities = 55/60, 92%; Positives = 59/60, 98%; Gaps = 0%), located on chromosomes 7, 3, and 1, respectively (**Figure 6B**). Overall, these three putative orthologs shared more than 56% sequence identity with *A. thaliana* MYB36 (**Figure 6B)** and were annotated as MYB transcription factors according to Mercator v4 pipeline [21].

In *A. thaliana*, *MYB36* (AT5G57620) is primarily expressed in roots and seeds during germination [22,23]. We hypothesized that the identified orthologs in *P. vulgaris* would exhibit a similar expression pattern. To investigate this, we analyzed the expression profiles of the putative MYB36 orthologs (*PvNJ1.1_chr7_1827900.t1* = *Pv*MYB36-7, *PvNJ1.1_chr3_1113800.t1* = *Pv*MYB36-3, and *PvNJ1.1_chr1_0164500.t1* = *Pv*MYB36-1) across various tissues and experimental conditions using the NJEA module in *Phabase* (**Figure 6C)**. Overall, the three orthologs were predominantly expressed in roots, with *PvMYB36-7* showing the highest transcript abundance. Notably, *PvMYB36-7* was also the only ortholog detected in seeds. Its expression profile, characterized by strong root-specific expression and moderate activity in developing seeds (**Figure 6C)**, closely mirrors that of *MYB36* in *A. thaliana*, which is known to regulate endodermal differentiation and Casparian strip formation. This similarity in spatial expression patterns suggests that *Pv*MYB36-7 likely represents the functional MYB36 ortholog in *P. vulgaris*, potentially playing a conserved role in root barrier formation and developmental regulation.

Together, this quick analysis illustrates the utility of *Phabase* for integrating sequence-based ortholog identification with expression profiling to rapidly generate data-driven hypotheses.

Discussion

Negro Jamapa is one of the technological and agronomic milestones in Latin America common bean history. It is not only part of the national food security pilar of Mexico but also represents a key model for functional genomics analyses and the generation of expression data. The assembly and annotation of its genome, one of the aims of this project, give a strong base to study the genomic structure and functional diversity of Mesoamerican common beans. This new genome helps to close a gap in bean genomics, since accessible genomic analysis tools for the common bean community are lacking and only available for the G19833 genome reference, from the Andean gene pool. With this high-quality assembly and the creation of a unified data portal, *Phabase*, our work improves the access and integration of genomic, transcriptomic, and small RNA data, from an agronomically important Mesoamerican cultivar, for the bean research community.

Comparative Genome Architecture and Evolutionary Insights

The Negro Jamapa genome shows robust quality and completeness (N50 = 45 Mb; 98.4%), which are higher than those of the reference genome G19833 (N50 = 2 Mb; 97.7%). This difference reflects how long-read sequencing and assembly technologies allow the generation and democratization of high-quality genomes, especially for complex plant genomes. Other studies show that structural variations contribute to reproductive isolation, local adaptation, and domestication traits in *P. vulgaris* [2,24, 25, 26]. As expected, Negro Jamapa and G19833 belonging to different gene pools, their genomes present clear divergence, mainly explained by large structural variants. The facilitating access to this information allows the community to explore the genomic details of this divergence and reveal the phenotypic and physiological characteristics specific to Negro Jamapa and Mesoamerican common beans.

An Integrative Atlas for Transcriptomic and Small RNA Data

To extend the utility of the Negro Jamapa genome for functional and comparative studies, we developed *Phabase*, an open and integrative data portal designed for the *P.* research community. *Phabase* offers a complete expression atlas that integrates both coding and non-coding genes, connecting genomic information with transcriptomic and small RNA data.

The expression section currently includes 121 experimental conditions representing diverse biological contexts, such as plant development, biotic and abiotic stresses, and symbiotic interactions, across 21 cultivars, increasing five times the quantity of information compared to the original gene atlas [17]. On average, these experiments have also more than two replicates, increasing the robustness of the data, compared to the one-replicate experiments available originally [17]. Root samples dominate the collection (around 51%), highlighting the strong focus of the bean community on root biology, nodulation, and rhizobial symbiosis. Remarkably, Negro Jamapa contributes to more than one-third of the total data, reinforcing the importance of this cultivar in the community and the need to generate robust genomic data for this cultivar.

For small RNAs, *Phabase* compiles data of 284 small RNAs, from 30 experimental conditions involving 11 cultivars, with samples from different organs and stress environments. To our knowledge, this is the first publicly available plant sRNA expression atlas website and also the unique to include exogenous sRNA identified as host regulators by croos-kingdom RNAi [18]. Together, these efforts make *Phabase* the most comprehensive and diverse transcriptomic resource currently available for *Phaseolus*.

Phabase is not only as a data repository but also a platform for functional exploration. We included tools for visualization of expression profiling and comparative analyses, allowing users to easily explore the data and test hypotheses without bioinformatics skills. *Phabase* also helps to reduce bean

genomic data fragmentation by integrating and homogenizing the publicly available data that were previously dispersed across different sources. We believe that it creates a unified and accessible space for large-scale comparative studies and meta-analyses in bean genomics.

As proof of concept and illustrate how *Phabase* can help when a user is interested in the analysis of one specific protein from another organism, we looked at AtMYB36, a transcription factor known for its role in the Casparian strip and endodermis formation in *Arabidopsis thaliana*. When we searched for similar sequences in the Negro Jamapa genome, we found three possible candidates: *PvMYB36-7, PvMYB36-3, and PvMYB36-1*, each one on a different chromosome. Investigating its expression profile, we confirmed that *PvMYB36-7* could be the functional orthologs of *AtMYB36*. These two genes display similar expressions in roots, with some signal in developing seeds. This small test shows that, thanks to *Phabase*, the community can quickly check homology, look at expression across several conditions and cultivars, and start forming hypotheses without bioinformatic skills.

The Negro Jamapa genome and *Phabase* together fill a gap for the *Phaseolus* community: a Mesoamerican reference genome with high contiguity and a platform gathering different types of genomic data. Compare Mesoamerican and Andean gene pools and explore questions related to domestication, stress responses, or plant-microbe interaction is now more accessible. More generally, these tools reduce the fragmentation of bean genomic data and make it easier for researchers, especially those without deep bioinformatics experience, to work with large-scale datasets.

Materials and methods

Biological material and sequencing

Commercial seeds of *P. vulgaris* var. Negro Jamapa were disinfected with 3% sodium hypochlorite for 10 minutes, followed by three washes with sterile distilled water to eliminate residual chlorine. The sterilized seeds were sown in pots containing Peat Moss and Leica substrate (ratio 3:1) and cultivated under greenhouse conditions for two weeks. Young leaves from 15 individual plants were harvested, frozen, and subsequently macerated in liquid nitrogen. High molecular weight DNA was extracted using the NucleoBond HMW DNA kit (Macherey Nagel), in accordance with the manufacturer's protocol. The resulting HMW DNA was submitted to Novogene for sequencing via Hi-Fi PacBio technology, yielding 95.3 Gb sequencing data with the following parameters: Maximum read length, 64 kb; Mean read length, 21 kb; N50 read length, 21 kb.

Assembly pipeline

HiFi PacBio reads were assembled to contigs using hifiasm version 0.20.0-r639 with the default parameters [27]. Organellar genomes were assembled using Oatk version 1.0 with -c parameter (minimum k-mer coverage = 200; k-mer length = 1001) was set to 200 [28]. We then performed homology searches for hifiasm contigs over the organellar genomes using BLAST. Short hifiasm contigs (< 1,000,000 nt) with high similarity (BLAST search filters: e-value < 1E-06, % identity > 99%, match length > 1000) to the organellar genomes were removed, leaving the contigs pertaining to the nuclear genome. These nuclear contigs then were subjected to the correcting and scaffolding steps of the RagTag pipeline [29]. Genome assembly *P. vulgaris* G19833 v2.0 was implemented in these steps as the reference genome [2]. The final assembly, including all 11 chromosomes that were placed to the reference genome, the unplaced contigs, and the organellar genomes were then subjected to gene prediction.

Genome annotation

The repeat sequences in the genome were identified using Repeatmodeler (v2.0.5) [30] and were masked using RepeatMasker (v4.1.5) [31]. Previously published RNAseq [17] data were mapped using STAR (v2.7.11b) [32]. The Viridiplantae OrthoDB (v10.1) [33], merged with the proteins from the reference genome of *P. vulgaris* were also used in the pipeline. The annotation was done using Braker3 (v3.0.8) [34]. The completeness of the annotations was assessed using BUSCO (v5.8.3) [35]. The protein sequences were functionally annotated using Mercator4 v7.0 [21]. The UTRs for the genes were added using Stringtie based on the RNAseq data mapping [36]. miRNA genes were predicted thanks to miRDeep-P2 [37] and Shortstack [38] ; and validated by miRScore [39], using the dataset displays in the miRNA Gene Atlas as sRNA library inputs (Table S1).

Overview and technologies Phabase

The database layer is based on MongoDB, which stores the integrated data model (genesDatamart) derived from multiple collections containing gene, transcript, and product annotations. This hierarchical structure enables fast retrieval of gene-centric information including sequence data, genomic coordinates, and functional annotations. The web services layer is implemented in Node.js and exposes a GraphQL API for handling data queries and gene search operations. This layer also communicates with specialized Flask (Python) services responsible for processing computational requests. The presentation layer is built with Next.js, providing an interactive user interface named JAMAPA Browser. This interface integrates four functional modules: (i) Gene Search, (ii) Expression Atlas, (iii) Genome Browser, and (iv) BLAST. All client requests are routed through a Nginx reverse proxy, which manages HTTPS communication between the user and the web services. The modular

design allows each component to operate independently, facilitating system updates and integration of new datasets or analysis tools. All components communicate through REST or GraphQL interfaces, ensuring reproducibility, interoperability, and transparent data access for both visualization and downstream analysis.

The Genome Browser module uses JBrowse 2, deployed as a separate service configured via a JSON file with references to genome FASTA and GFF3 annotation files. In addition, *Phabase* integrates the Basic Local Alignment Search Tool (BLAST), enabling users to identify orthologous sequences across the available genomes using either nucleotide or amino acid queries. The BLAST service runs locally on a dedicated server, where the Flask API executes the BLAST command-line tool against locally indexed databases generated with makeblastdb. Search results are returned in JSON or tabular format. For more precise control, advanced parameters are enabled, allowing users to adjust their searches according to specific needs.

Expression Atlas

The Expression Atlas service accesses precomputed RNA-seq data stored in CSV files. Flask reads and parses these datasets, serving the results as JSON objects that are visualized dynamically in the front-end using D3.js. The gene and microRNA expression atlas encompass 261 libraries from 121 experimental conditions, and 67 libraries from 30 experimental conditions, respectively, including responses to abiotic stress, symbiotic interactions, and various developmental stages (Table S1 and S2). NJEA allows users to visualize and compare gene expression patterns across diverse conditions, facilitating the identification of context- and tissue-specific gene regulation. To build the NJEA, raw RNA-seq reads were retrieved from NCBI using SRA-toolkit [40] and low-quality bases and adapter sequences were removed with fastp [41]. The resulting high-quality reads were aligned to the *P. vulgaris* Negro Jamapa reference genome using the splice-aware aligner STAR [32]. Transcript-level expression values were quantified with featureCounts [42] and read counts were subsequently normalized using edgeR [43]. For miRNA, reads where collapsed and counted using FASTQ/A Collapser from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and annotated with compiled data from previous studies [15,18,19].

Acknowledgements

The authors thank the Bioinformatics Analysis Unit (UAB) and the Technology and Information Management Unit (UATI) for their invaluable assistance with genome assembly and website management. We also thank Andrés G. López-Almazo, Felipe Betancourt-Figueroa, and Gabriel

Alarcón-Carranza for their technical expertise, assistance with installation, and continuous support in bringing the *Phabase* database online.

Funding

This work was supported by the Mexican Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI, grant CBF2023-2024-834 to J. M.) and Dirección General de Asuntos del Personal Académico (DGAPA)-Universidad Nacional Autónoma de México (UNAM) – Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT, grant IA200125 to J. M. and IN203925 to D.F.).

Contributions

J.M. and D.F. planned and designed the research. S.U.A. and D.F. analyzed the data. J.M. obtained the Negro Jamapa HMW-DNA. H.S. and E.P. developed Phabase. T.Y.A., E.V.R., N.T., and L.I.F. performed the genome assembly, annotation and, in silico analyses. T.Y.A., D.F., and E.V.R. wrote the manuscript. J.M., D.F., and S.U.A. revised the manuscript.

Data availability statement

Phabase and corresponding data can be accessed at the following URL: https://phabase.ccg.unam.mx/ and repository : https://github.com/tyakyol/Phabase.

Conflict of interests

The authors declare no competing interests.

Supplementary information

Table S1 and Table S2.

References

1 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 of the United States of America* 2012; **109**: E788–E796.

2 Schmutz J *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* 2014; **46**: 707–713.

3 Bitocchi E *et al.* Molecular analysis of the parallel domestication of the common bean (Phaseolus vulgaris) in Mesoamerica and the Andes. *New Phytologist* 2013; **197**: 300–313.

4 Gepts P. Origin and evolution of common bean: Past events and recent trends. *HortScience* 1998; **33**: 1124–1130.

5 Chavez-Mendoza C, S\'anchez E. Bioactive compounds from Mexican varieties of the common bean (Phaseolus vulgaris): Implications for health. *Molecules* 2017; **22**: 1360.

6 Espinosa-Alonso LG, Lygin A, Widholm JM, Valverde ME, Paredes-Lopez O. Polyphenols in wild and weedy Mexican common beans (Phaseolus vulgaris L.). *Journal of agricultural and food chemistry* 2006; **54**: 4436–4444.

7 Palacio-Marquez A, Ojeda-Barrios D, Jimenez-Castro J, Preciado-Rangel P, Hernandez-Rodriguez OA, Sanchez E. Biofortification potential in common bean (Phaseolus vulgaris L.): bioactive compounds, antioxidant capacity and physicochemical properties of 155 varieties grown in M\'exico. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* 2021; **49**: 12123–12123.

8 Vlasova A *et al.* Genome and transcriptome analysis of the Mesoamerican common bean cultivar Negro Jamapa. *Genome Biology* 2016; **17**: 32.

9 Carr\`ere S, Mayjonade B, Lalanne D, Gaillard S, Verdier J, Chen NWG. First whole genome assembly and annotation of a European common bean cultivar using PacBio HiFi and Iso-Seq data. *Data in Brief* 2023; **48**: 109182.

10 Rend\'on-Anaya M, Yu L, Yang X-F, Ingvarsson PK. Species-specific genetic routes to a shared phenotype: convergent adaptation of beans to European long days. *bioRxiv* 2025; : 2025–09.

11 Wang Y *et al.* Telomere-to-telomere genome of common bean (Phaseolus vulgaris L., YP4). *GigaScience* 2025; **14**: giaf001.

12 Mun J-H, Bachmann A, Gupta V, Stougaard J, Andersen SU. Lotus Base: An integrated information portal for the model legume Lotus japonicus. *Scientific Reports* 2016; **6**: 39447.

13 Pecrix Y *et al.* Whole-genome landscape of Medicago truncatula symbiotic genes. *Nature Plants* 2018; **4**: 1017–1025.

14 Kim KJ *et al.* A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. *Plant Physiology* 2015; **168**: 1433–1447.

15 Formey D *et al.* Genome-wide identification of the Phaseolus vulgaris sRNAome using small RNA and degradome sequencing. *BMC Genomics* 2015; **16**: 423.

16 Subramani M *et al.* Comprehensive proteomic analysis of common bean (Phaseolus vulgaris L.) seeds reveal shared and unique proteins involved in terminal drought stress response in tolerant and sensitive genotypes. *Biomolecules* 2024; **14**: 109.

17 O'Rourke JA *et al.* An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics* 2014; **15**: 866.

18 Formey D *et al.* Regulation of Small RNAs and Corresponding Targets in Nod Factor-Induced Phaseolus vulgaris Root Hair Cells. *International Journal of Molecular Sciences* 2016; **17**: 887.

19 Ren B, Wang X, Duan J, Ma J. Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. *Science* 2019; **365**: 919–922.

20 Tsai H-H *et al.* Localized glutamine leakage drives the spatial structure of root microbial colonization. *Science* 2025; **390**: eadu4235.

21 Bolger M, Schwacke R, Usadel B. MapMan visualization of RNA-Seq data using Mercator4 functional annotations Solanum tuberosum. In: Dobnik D, Gruden K, Ram\vsak \vZiva, Coll A (eds). *2354*. Humana Press: New York, NY, 2021, pp 87–113.

22 Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *The Plant Journal* 2016; **88**: 1058–1070.

23 Reiser L, et al. The Arabidopsis Information Resource: making and mining the ``gold standard'' annotated reference plant genome. *Nucleic Acids Research* 2022; **51**: D??

24 Ambachew D, Londo\~no JM, Rodriguez Castillo N, Asfaw A, Blair MW. Genetic diversity, linkage disequilibrium, and population structure in a common bean reference collection. *Agronomy* 2024; **14**: 985.

25 MacQueen AH *et al.* Local to continental-scale variation in fitness and heritability in common bean (Phaseolus vulgaris). *Crop Science* 2022; **62**: 767–779.

26 Rend\'on-Anaya M *et al.* Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biology* 2017; **18**: 60.

27 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 2021; **18**: 170–175.

28 Zhou C, Brown M, Blaxter M, Darwin T of LPC, McCarthy SA, Durbin R. Oatk: a de novo assembly tool for complex plant organelle genomes. *Genome Biology* 2025; **26**: 235.

29 Alonge M *et al.* Automated assembly scaffolding using RagTag elevates and validates de novo plant genome assemblies. *Genome Biology* 2022; **23**: 258.

30 Flynn JM *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* 2020; **117**: 9451–9457.

31 Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* 2009; **25**: 4.10.1-4.10.14.

32 Dobin A *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**: 15–21.

33 Kriventseva EV *et al.* OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* 2019; **47**: D807–D811.

34 Gabriel L *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome research* 2024; **34**: 769–777.

35 Manni M, Berkeley MR, Seppey M, Sim\~ao FA, Zdobnov EM. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 2021; **38**: 4647–4654.

36 Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLOS Computational Biology* 2022; **18**: e1009730.

37 Kuang Z, Wang Y, Li L, Yang X. miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics* 2019; **35**: 2521–2522.

38 Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. *Rna* 2013; **19**: 740–751.

39 Vanek A, Griffiths-Jones S, Meyers BC, Shahid S, Axtell MJ. miRScore: A rapid and precise microRNA validation tool. *PLOS Computational Biology* 2025; **21**: e1013663.

40 Leinonen R, Sugawara H, Shumway M, International NSDC. The Sequence Read Archive. *Nucleic Acids Research* 2011; **39**: D19–D21.

41 Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018; **34**: i884–i890.

42 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014; **30**: 923–930.

43 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**: 139–140.

Figure legends

Figure 1. Circos plot illustrating the genomic landscape and structural features of the *P. vulgaris* Negro Jamapa chromosomes. Track (a) displays chromosomal ideograms, arranged sequentially and scaled according to chromosome length. Track (b) represents the distribution of repetitive elements (green lines). Track (c) shows the genomic positions of micro RNA (miRNA) precursor genes (dark green points). Track (d) corresponds to annotated protein-coding genes density (vertical green lines). (e) Colored inter-chromosomal links connect pairs of potentially duplicated genes, representing putative segmental duplications or paralogous relationships.

Figure 2. Genome-wide distribution of structural variants between *P. vulgaris* Negro Jamapa and G19833 genomes. (A) Comparative genomic analysis. For each chromosome, the upper track displays the distribution of SNPs (blue) and the lower track represents annotated genes (brown). Vertical links between chromosomes indicate structural variants identified by SyRI and visualized

with plotsr, including inversions (yellow), translocations (green), and duplications (light blue). Syntenic regions are represented in grey. Vertical dashed lines indicate the relative chromosome position (Megabases). **(B)** Summary of detected structural variants. Chromosomal representation of the cumulated length proportion of different classes of structural variants identified across the 11 chromosomes of *P. vulgaris* Negro Jamapa. Each bar indicates the relative abundance (0–100%) of variant types per chromosome. INVs: inversions; TRANs: translocations; DUPs: duplications; SNPs: single nucleotide polymorphisms; DELs: deletions; TRs: tandem repeats. Structural variants were detected using SyRI (Synteny and Rearrangement Identifier) based on pairwise genome alignment.

Figure 3. System architecture of *Phabase*. The diagram illustrates the modular architecture of the *Phabase* platform, showing the interaction between the main user modules, web services, and data sources. Green boxes represent the user interface modules, orange boxes correspond to web services and computational tools, and blue boxes denote data storage and sources.

Figure 4. Overview of *Phabase* tools and interface. (A) Home page of *Phabase* showing the main navigation menu. *Home* (1), *Gene Expression* (2), *Genome Browser* (3), *BLAST* (4), and *About Us* (5) sections. The search bar (6) allows users to query genes of interest directly. **(B)** Genome Browser interface illustrates the visualization of genomic regions. Genes are represented as annotated tracks and users can inspect gene models, annotations, and neighboring genomic elements (7). Selecting a feature opens a detailed information panel (8), which includes gene coordinates, structure, functional annotation, and nucleotide sequences. Additional transcript-level annotations and predicted subfeatures (9) can be explored through collapsible sections. **(C)** BLAST interface allows users to submit a nucleotide or protein query sequence in FASTA format (10) and search against multiple *Phaseolus* reference genomes selectable from the database list (11). Optional advanced parameters (12), including E-value threshold, word size, number of alignments to display, and filtering options, enable customized similarity searches.

Figure 5. Overview of the Expression Atlas module in *Phabase*. Summary of the datasets integrated into *Phabase* for **(A)** Gene and **(B)** microRNA Expression interface. Charts display the number of samples categorized by cultivar, tissue/organ, and topic. The number at the center of each chart indicates the total number of elements represented in the corresponding chart. Geographic distribution of the research groups that generate datasets that were processed to obtain the **(C)** Gene and **(D)** microRNA Expression module. The diameter of a circle is proportional to the number of conditions produced in the corresponding country. Colors differentiate regions of the world.

Figure 6. *Phabase*-enabled identification of MYB36 orthologs in *P. vulgaris* cv. Negro Jamapa. **(A)** Workflow illustrating how *Phabase* was used to identify MYB36 orthologs in *P. vulgaris* cv. Negro Jamapa. **(B)** Genomic locations and amino acid similarity matrix of the putative MYB36 orthologs. **(C)** Expression profiles of the identifiedMYB36 orthologs across different tissues and conditions.

Tables and figures

Table 1. *P. vulgaris* cv. Negro Jamapa genome assembly and annotation statistics.

Identifier	Molecule	Size	GC%	Transcripts
PvNJ1.1_chr1	Chromosome	56,338,280	35.85	2780
PvNJ1.1_chr2	Chromosome	49,825,282	33.77	3392
PvNJ1.1_chr3	Chromosome	50,421,212	33.92	3052
PvNJ1.1_chr4	Chromosome	48,781,969	36.48	1718
PvNJ1.1_chr5	Chromosome	40,291,510	36.52	1724
PvNJ1.1_chr6	Chromosome	35,918,269	34.50	2383
PvNJ1.1_chr7	Chromosome	41,572,620	33.50	2944
PvNJ1.1_chr8	Chromosome	56,415,169	35.89	2904
PvNJ1.1_chr9	Chromosome	40,546,796	32.48	2778
PvNJ1.1_chr10	Chromosome	44,913,661	37.53	1689
PvNJ1.1_chr11	Chromosome	56,669,149	37.48	2047
PvNJ1.1_mito	Mitochondrial	395,789	45.11	80
PvNJ1.1_chloro	Chloroplastic	150,286	35.44	170
	Whole genome	522,239,992	36.31	30731

Figure 1. Circos plot illustrating the genomic landscape and structural features of the *P. vulgaris* Negro Jamapa chromosomes.

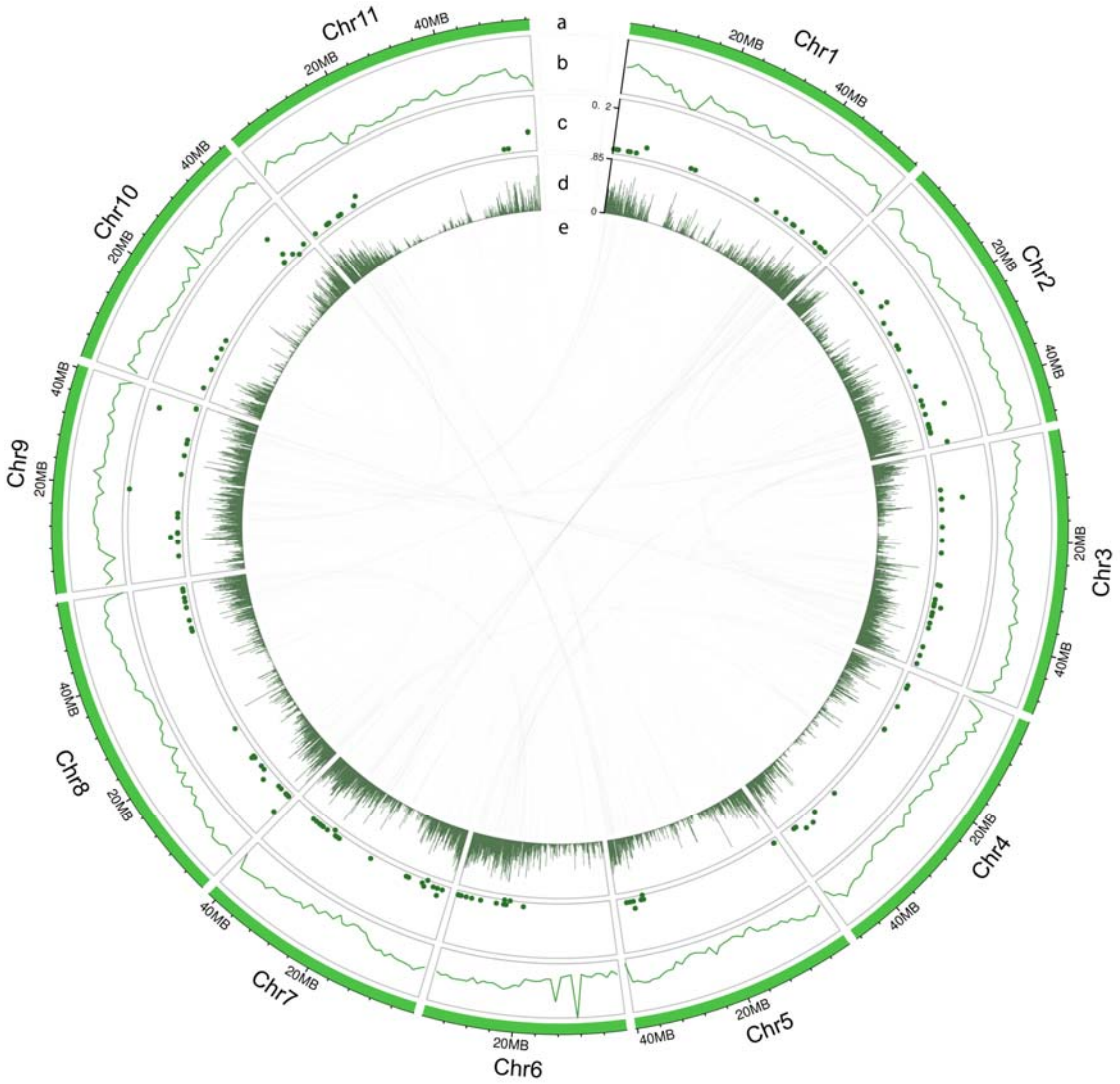

Figure 2. Genome-wide distribution of structural variants between *P. vulgaris* Negro Jamapa and G19833 genomes.

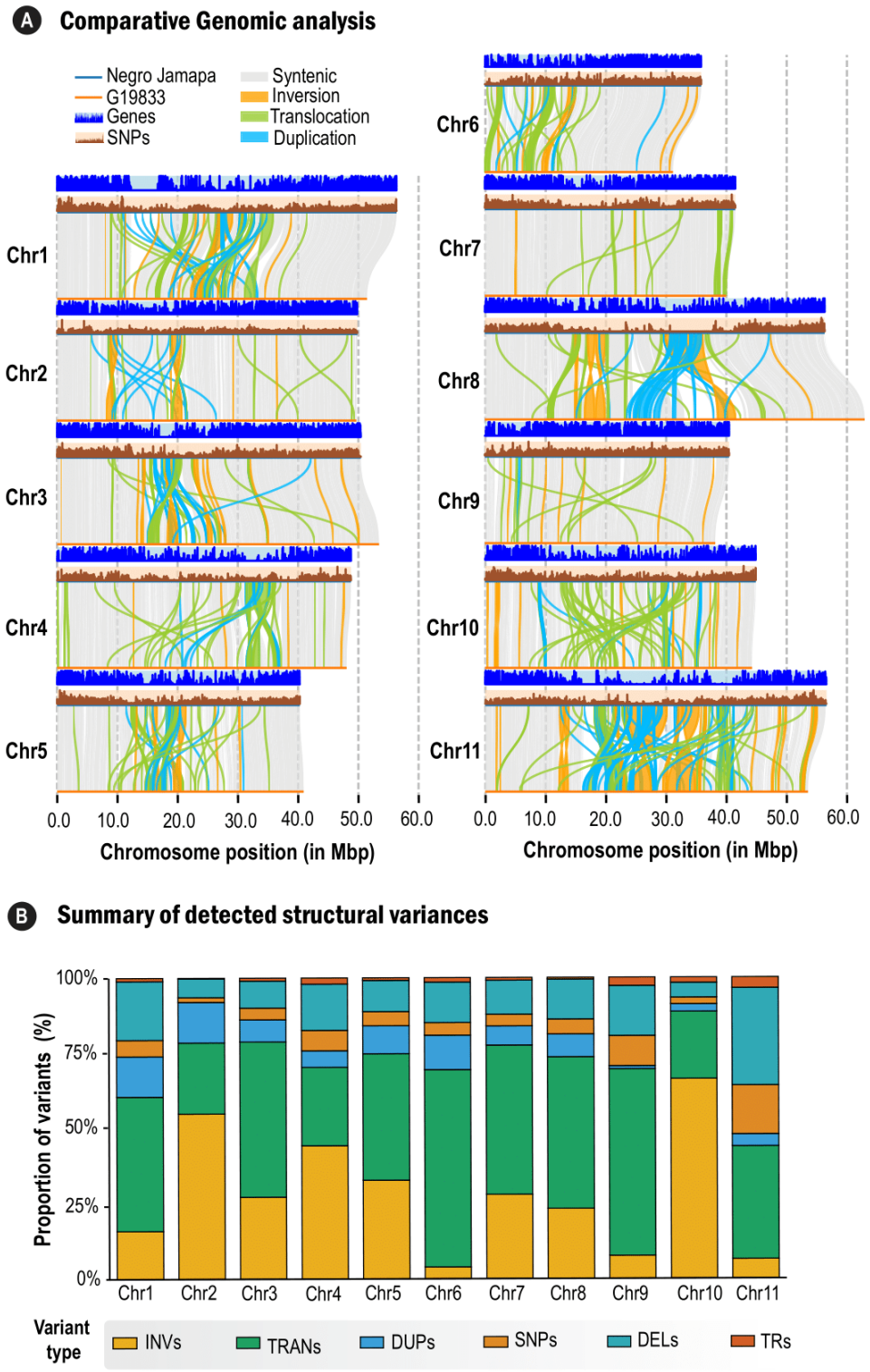

Figure 3. System architecture of *Phabase*.

https://phabase.ccg.unam.mx/

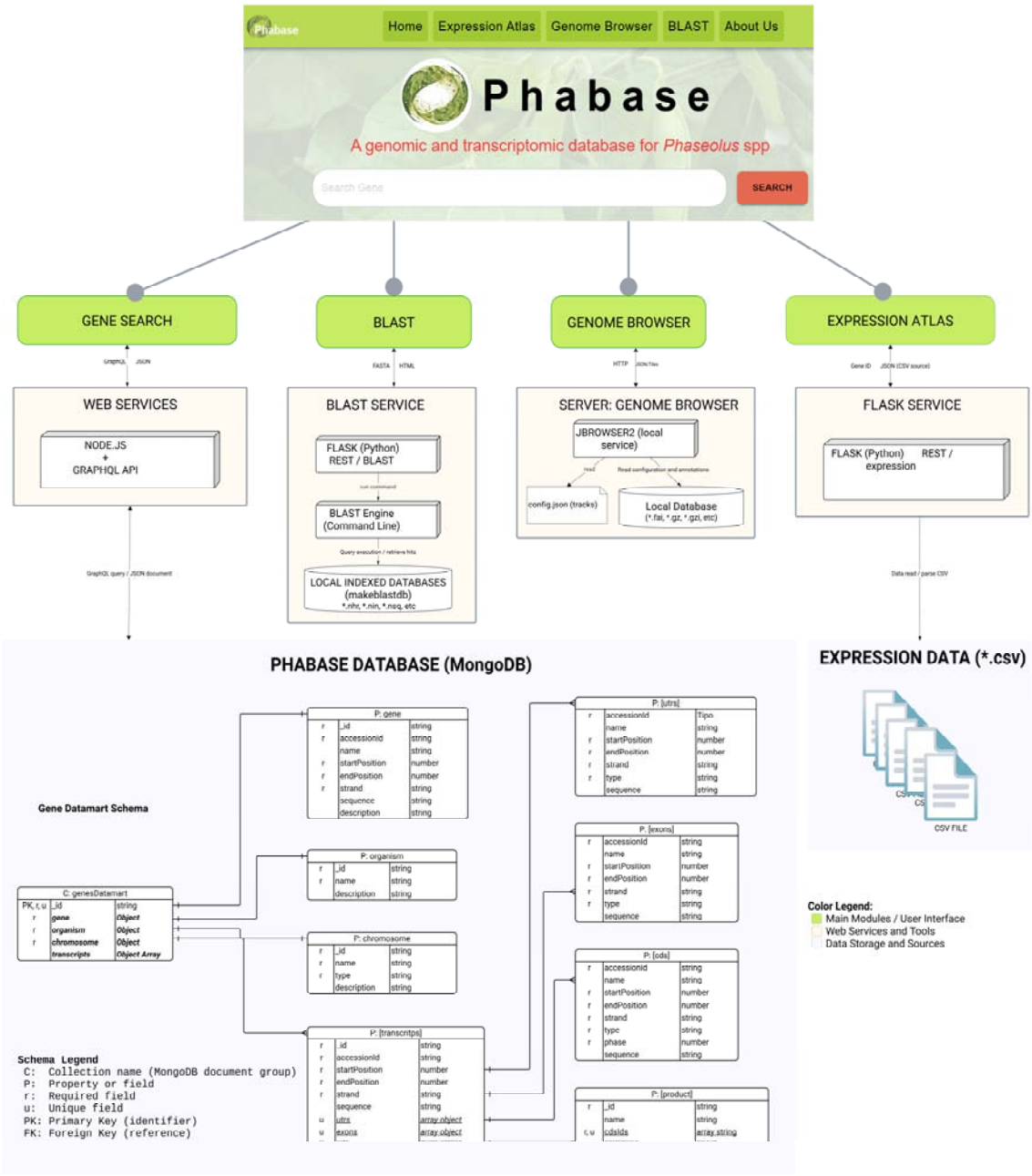

Figure 4. Overview of *Phabase* tools and interface.

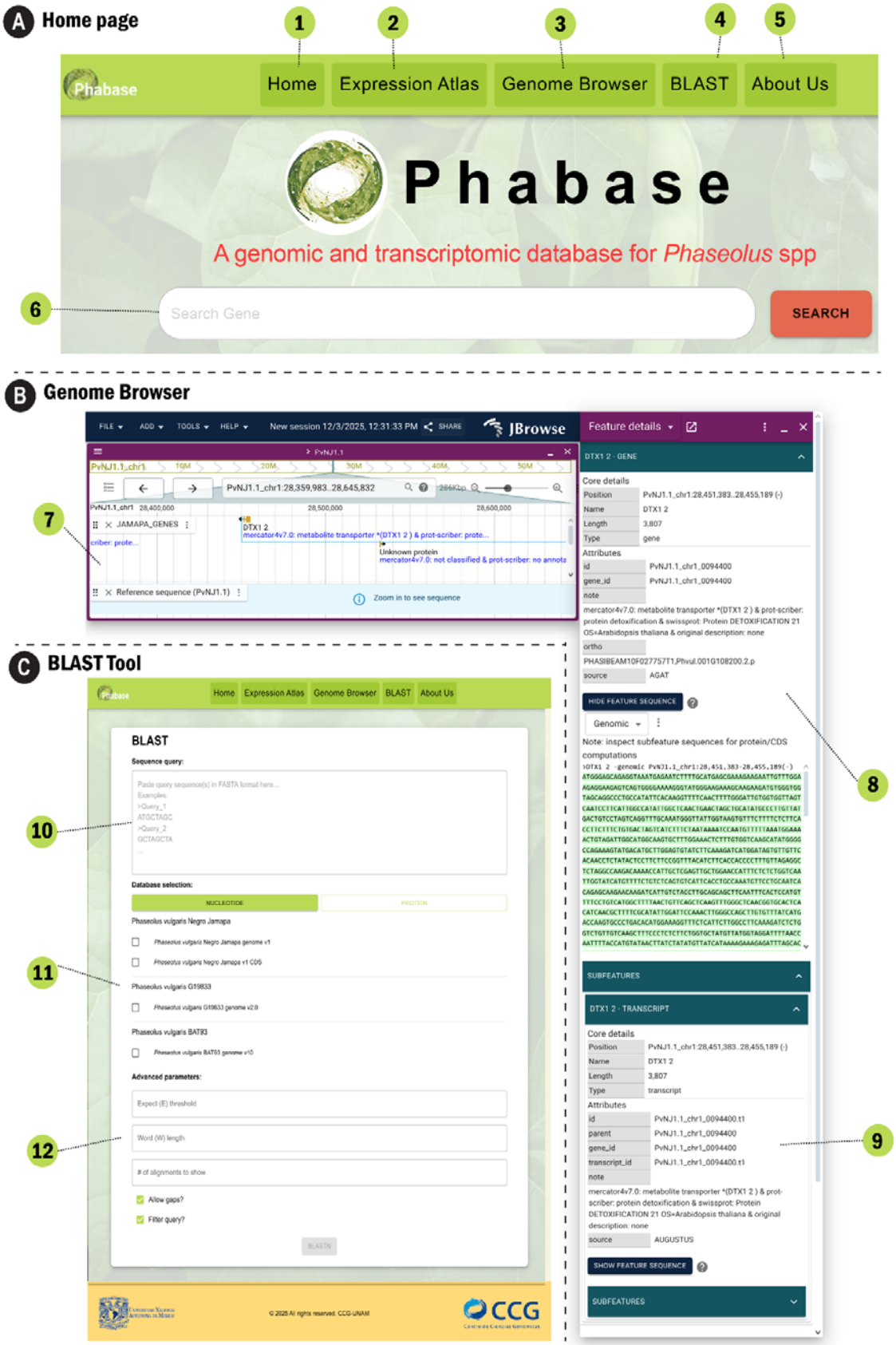

Figure 5. Overview of the Expression Atlas module in *Phabase*.

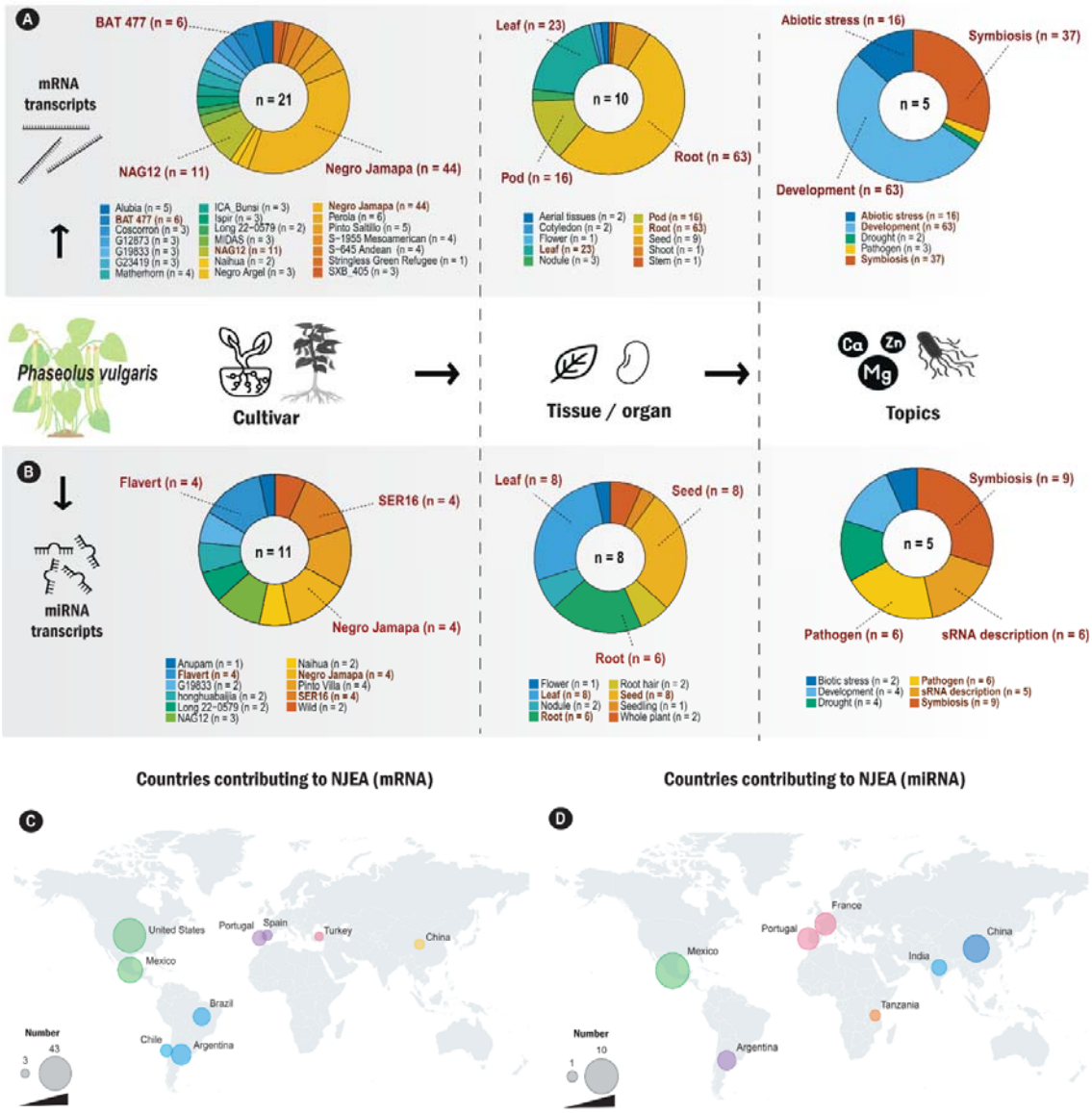
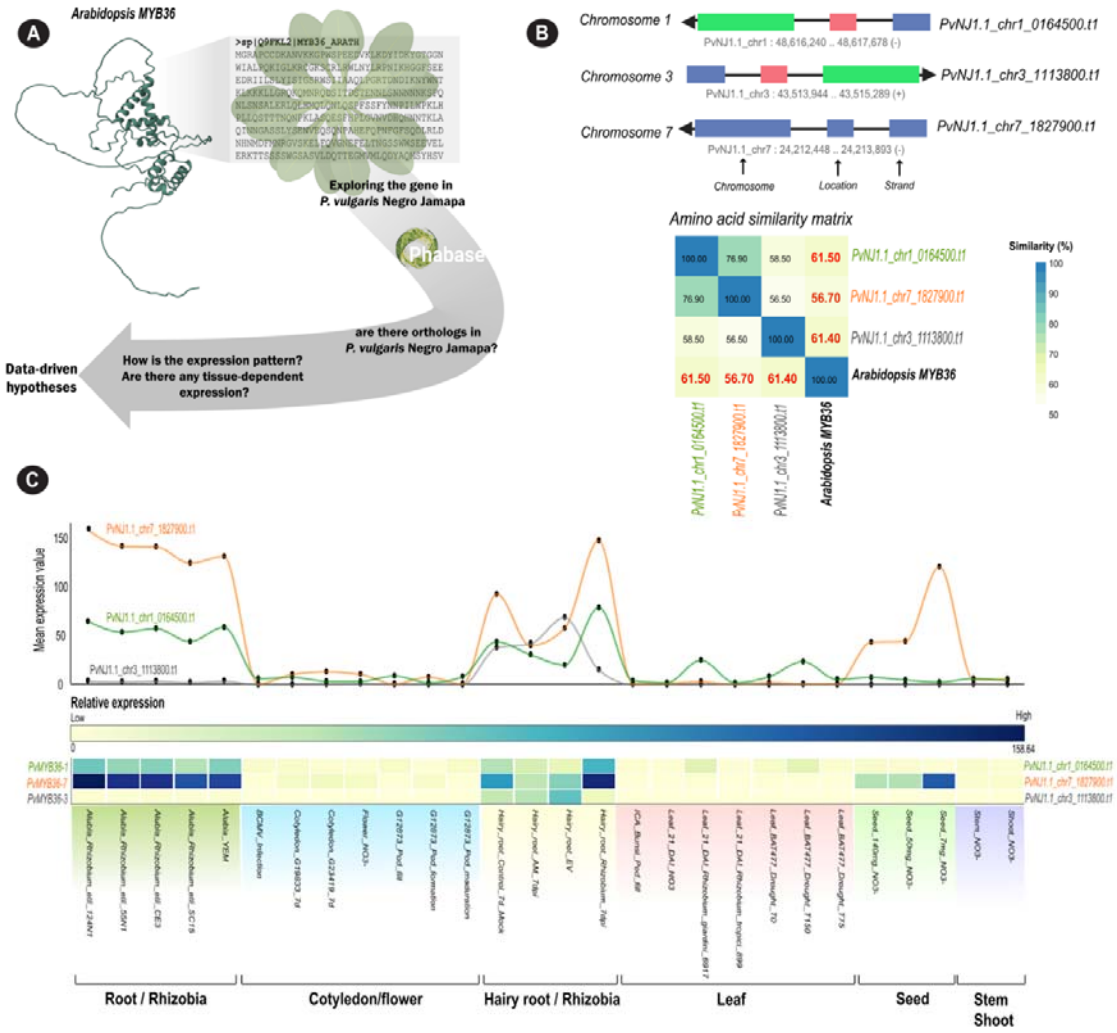
Figure 6. *Phabase*-enabled identification of MYB36 orthologs in *P. vulgaris* cv. Negro Jamapa.