Data in Brief 55 (2024) 110759

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

RADseq datasets of native beans from Mexico

Aragón-Magadán Marco Aurelio, Cruz-Cárdenas Carlos Iván, Calvillo-Aguilar Francisco Fabián, Pichardo-González Juan Manuel, Guzmán Luis Felipe*

National Genetic Resources Center, National Institute of Forestry, Agricultural and Livestock Researches, Jalisco, México

ARTICLE INFO

Article history:
Received 30 May 2024
Revised 11 July 2024
Accepted 12 July 2024
Available online 17 July 2024

Dataset link: Native bean species Raw sequence reads (Original data)
Dataset link: VCF file for 45 accessions of the genus Phaseolus: P. . acutifolius (14), P. coccineus (12), P. lunatus (8), P. dumosus (6), P. leptostachyus (2), P. filiformis (2) y P. vulgaris (1) (Original data)

Keywords:
Genotype-by-sequencing
Bioinformatics
Phaseolus
SNP
VCF

ABSTRACT

Forty-five accessions of the genus *Phaseolus* from the orthodox seed collection of the National Center for Genetic Resources (CNRG) of the National Institute of Forestry, Agricultural, and Livestock Research (INIFAP) of Mexico were sequenced using RADseq. The species utilized were: *P. acutifolius (14), P. coccineus (12), P. lunatus (8), P. dumosus (6), P. leptostachyus (2), P. filiformis (2),* and *P. vulgaris* (1). A variant call file (VCF) was generated using GATK with the *P. vulgaris* reference genome GCF_000499845.1, identifying 97,103 shared SNPs among the species. These data have the potential to be used for studies of genetic diversity intra and interspecies, phylogeny, evolution, genetic resource conservation, and agricultural improvement.

* Corresponding author.
E-mail address: guzman.luis@inifap.gob.mx (G.L. Felipe).

https://doi.org/10.1016/j.dib.2024.110759

Specifications Table

Subject	Agronomy and Crop Science
Specific subject area	Genotype-by-sequensing
Type of data	*Raw sequencing libraries, Raw VCF files*
Data collection	Raw RADseq libraries of *Phaseolus coccineus, P. lunatus, P. dumosus, P. acutifolius, P. leptostachyus, P. filiformis* and *P. vulgaris*. *RAW VCF file*
Data source location	Seeds of *Phaseolus* spp. were collected from 18 states in Mexico. They were germinated and samples were collected from the plants for sequencing at the Centro Altos Experimental Field of INIFAP, Jalisco, Mexico (20.8734816, -102.7127676). The plant material is stored at the National Center for Genetic Resources of INIFAP, Orthodox Seed Collection section, Jalisco, Mexico (20.8728617, -102.7099154)
Data accessibility	Repository name: National Center for Biotechnology Information Data identification number: BioProject PRJNA1117450, BioSamples SAMN41569804 to SAMN41569848 Repository name: Figshare Data identification number: 10.6084/m9.figshare.25917763 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1117450; 10.6084/m9.figshare.25917763.v2
Related research article	

1. Value of the Data

- The data are valuable because they consist of accessions that are part of the Orthodox Seed Collection of the CNRG-INIFAP for long-term conservation.
- The data are valuable for the long-term conservation of native *Phaseolus* species because it provides important information on genetic diversity, population structure, and potential specific markers. This information enables the development of conservation strategies that ensure the genetic stability of these genetic resources over time. Furthermore, the data could be used to create a panel of SNPs specific for the species reported, which would be helpful for genotyping.
- The data provide information on SNPs of native bean species present in Mexico.
- The data are useful for researchers conducting genetic diversity analyses in wild species.
- The data are useful because they allow the detection of SNP molecular markers shared among *Phaseolus* species.
- The data are valuable because they provide sequencing information on native bean varieties that are underrepresented in genomic databases.

2. Background

Mexico is the center of origin and domestication of beans (*Phaseolus* spp.), which are fundamental in the diet of its citizens and form part of the country's cultural and gastronomic heritage [1]. Within this genus, there are five domesticated species: *P. vulgaris* (common bean), *P. lunatus* (lima bean), *P. acutifolius* (tepary bean), *P. coccineus* ssp. coccineus (runner bean), and *P. dumosus* [1], all of which are important for long-term conservation at the National Center for Genetic Resources (CNRG).

Despite the importance of *Phaseolus* in Mexico, most efforts to study the genetic diversity of this genus have focused on the common bean [2–4]. Therefore, the primary motivation for this dataset is to generate information on SNP molecular markers for the study of intra- and interspecies genetic diversity of *P. coccineus, P. lunatus, P. dumosus, P. acutifolius, P. leptostachyus, P. filiformis,* and *P. vulgaris.*

3. Data Description

There are 45 raw RADseq sequencing libraries of bean accessions from seven species of the genus *Phaseolus*, preserved in the orthodox seed collection of CNRG-INIFAP. Of these, 14 are from *P. acutifolius* A., 12 *from P. coccineus L.,* eight from *P. lunatus L.*, six from *P. dumosus*, two from *P. filiformis*, two from *P. leptostachyus*, and one from *P. vulgaris* (Table 1).

The exploratory variant analysis performed with GATK revealed 97,103 variants (SNPs) shared among species. Further analysis with vcfR shows that the number of variants is distributed throughout the entire reference genome used (Fig. 1). The read depth is below 10 (Fig. 2), while the mapping quality is close to the maximum value of 60 in most positions (Fig. 1). These val-

Table 1
Samples sequenced by RADseq.

Sample name	Accession	Species
G40018	SAMN41569804	*Phaseolus acutifolius*
G40082	SAMN41569805	*Phaseolus acutifolius*
G40105	SAMN41569806	*Phaseolus acutifolius*
G40045	SAMN41569807	*Phaseolus acutifolius*
G40083	SAMN41569808	*Phaseolus acutifolius*
G40020	SAMN41569809	*Phaseolus acutifolius*
G40034	SAMN41569810	*Phaseolus acutifolius*
G40052	SAMN41569811	*Phaseolus acutifolius*
G40107	SAMN41569812	*Phaseolus acutifolius*
G40119	SAMN41569813	*Phaseolus acutifolius*
G40127	SAMN41569814	*Phaseolus acutifolius*
G40006B	SAMN41569815	*Phaseolus acutifolius*
G40125	SAMN41569816	*Phaseolus acutifolius*
G40035	SAMN41569817	*Phaseolus acutifolius*
G35055	SAMN41569818	*Phaseolus coccineus*
G35097	SAMN41569819	*Phaseolus coccineus*
G35072	SAMN41569820	*Phaseolus coccineus*
G35005	SAMN41569821	*Phaseolus coccineus*
G35831A	SAMN41569822	*Phaseolus coccineus*
G35011	SAMN41569823	*Phaseolus coccineus*
G35052	SAMN41569824	*Phaseolus coccineus*
G35009	SAMN41569825	*Phaseolus coccineus*
G35106	SAMN41569826	*Phaseolus coccineus*
G35814	SAMN41569827	*Phaseolus coccineus*
G35053	SAMN41569828	*Phaseolus coccineus*
G35713	SAMN41569829	*Phaseolus coccineus*
G35071A	SAMN41569830	*Phaseolus dumosus*
G35789	SAMN41569831	*Phaseolus dumosus*
G35002	SAMN41569832	*Phaseolus dumosus*
G35054	SAMN41569833	*Phaseolus dumosus*
G35788	SAMN41569834	*Phaseolus dumosus*
G35004	SAMN41569835	*Phaseolus dumosus*
G40690	SAMN41569836	*Phaseolus filiformis*
G40687	SAMN41569837	*Phaseolus filiformis*
G40615	SAMN41569838	*Phaseolus leptostachyus*
G40566	SAMN41569839	*Phaseolus leptostachyus*
G25713C	SAMN41569840	*Phaseolus lunatus*
G25291	SAMN41569841	*Phaseolus lunatus*
G25230	SAMN41569842	*Phaseolus lunatus*
G25708A	SAMN41569843	*Phaseolus lunatus*
G25288	SAMN41569844	*Phaseolus lunatus*
G26174	SAMN41569845	*Phaseolus lunatus*
G25232	SAMN41569846	*Phaseolus lunatus*
G25229	SAMN41569847	*Phaseolus lunatus*
G22229	SAMN41569848	*Phaseolus vulgaris*

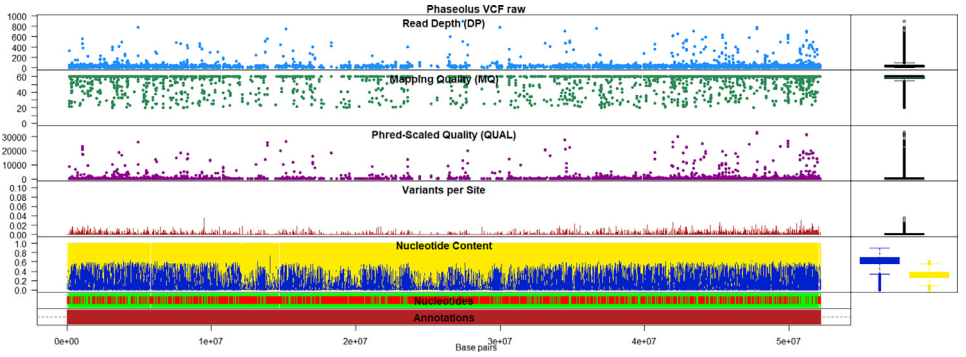

Fig. 1. Basic statistics in vcfR of the variant call.

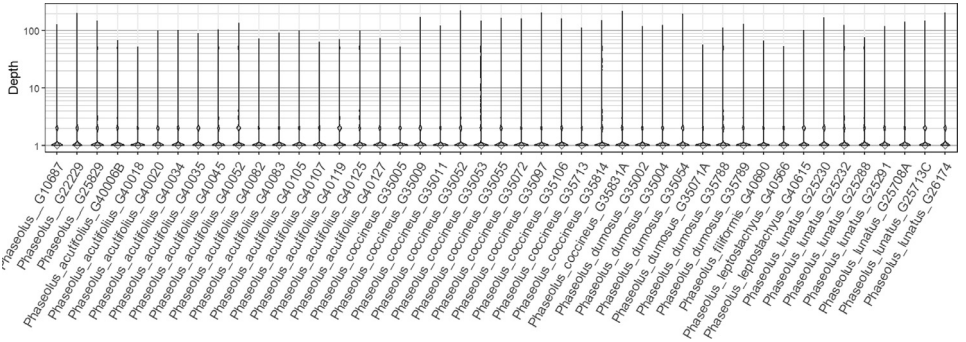

Fig. 2. Read depth of variants per accession.

ues are explained by the interspecific nature of the data, where the mapped regions are shared among the different accessions, resulting in a lower read depth.

When mapping sequencing reads to a reference genome, it is common for only highly conserved regions among different species to align effectively [5]. This results in lower coverage in specific regions of the genome due to genetic differences between species (Fig. 1). Although single-species analyses are the most common, SNP analyses among species of the same genus allow for the detection of variants in conserved regions [5,6]. This can help identify shared loci, useful for studies of genetic diversity, phylogeny, evolution, genetic resource conservation, and agricultural improvement [6]. These analyses provide crucial information on genetic variability and evolutionary relationships, facilitating the development of conservation strategies and genetic improvement programs [7].

4. Experimental Design, Materials and Methods

Forty-five bean accessions from seven species of the genus *Phaseolus*, preserved in the orthodox seed bank of CNRG-INIFAP, were included in this study. Of these, 14 were from *P. acutifolius* A., 12 from *P. coccineus L.*, eight from *P. lunatus* L., six from *P. dumosus*, two from *P. filiformis*, two from *P. leptostachyus*, and one from *P. vulgaris*. These accessions were collected from 18 states in Mexico.

The seeds were put in vinyl germination boxes with peat moss in a germination room at a temperature of 25 to 27°C, 70% relative humidity ± 10%, with 16 hours of light and 8 hours of

darkness for 15 days at the Centro Altos Experimental Field of INIFAP, Jalisco, Mexico. At least three plants per accession were obtained by planting three to five seeds per accession.

Leaves were collected from healthy, clean seedlings that were visually free of dryness or pests and were placed in 2 mL microcentrifuge tubes, frozen at -80°C for 48 hours, lyophilized at a temperature of -84°C and a minimum pressure of 0.120 mBar for five days using a Labconco Freezone 12 plus lyophilizer, and mechanically pulverized using a Tissue Lyser II.

DNA extraction was performed using the MagMax commercial kit (ThermoFisher Scientific) according to the manufacturer's recommendations. The DNA concentration was determined by fluorometry using a Qubit 2.0 instrument (Thermo Fisher Scientific) with the commercial Qubit dsDNA HS Assay kit from Invitrogen. Samples were standardized to a concentration of 100 ng/µL, and dilution plates were prepared at 20 ng/µL with a volume of 50 µL per sample.

DNA integrity was verified by electrophoresis at 80 V for 30 minutes on a 1% agarose gel stained with GelRed. Ten µL of λ DNA [10 ng/µL] uncut was used as a molecular weight marker. Ten µL of sample DNA plus 3 µL of loading buffer were applied. Bands were visualized with UV light using a transilluminator, and images were captured with a Carestream photodocumentation system.

DNA sequencing were performed on an Illumina HiSeq2000 sequencer by the service provider Floragenex. Data preprocessing included adapter trimming and selection of sequences with quality ≥Q25 using fastp [8].

For variant calling, an in-house developed script (https://github.com/Marco-CNRG/beans.git) [9] was used with the following workflow: sequence alignment with BWA-mem2 [10] against the common bean reference genome GCF_000499845.1, duplicate marking, and variant calling with GATK [11]. Finally, VCF file was analyzed in R 4.4.0 using the vcfR package [12].

Limitations

The main limitation stems from the variant calling due to the lack of published annotated genomes for the seven bean species used in this dataset. As of the publication date, genomes are only available for *P. vulgaris, P. lunatus*, and *P. coccineus*. This deficiency complicates variant calling and specific genetic diversity calculations for each species. For this reason, it was decided to use *P. vulgaris* as a reference, as it is one of the few available genomes and provides a suitable basis for mapping and analyzing the other species in the genus.

Ethics Statement

The data used for this protocol were provided by the Orthodox Seed laboratory and DNA and Genomics Laboratory of the National Center for Genetic Resources of INIFAP as a result of the project named "Regeneration, characterization and conservation to long term of common bean (*Phaseolus vulgaris* L.) in INIFAP," No. 11584934760.

Data Availability

Native bean species Raw sequence reads (Original data) (NCBI)

VCF file for 45 accessions of the genus Phaseolus: P. . acutifolius (14), P. coccineus (12), P. lunatus (8), P. dumosus (6), P. leptostachyus (2), P. filiformis (2) y P. vulgaris (1) (Original data) (Figshare)

CRediT Author Statement

Aragón-Magadán Marco Aurelio: Software, Conceptualization; **Cruz-Cárdenas Carlos Iván:** Writing – review & editing; **Calvillo-Aguilar Francisco Fabián:** Writing – review & editing;

Pichardo-González Juan Manuel: Writing – review & editing; **Guzmán Luis Felipe:** Data curation, Conceptualization, Writing – review & editing.

Acknowledgements

To the National Forestry, Agriculture and Livestock Researches Institute (INIFAP), for allocating the necessary funds from their budget to facilitate the successful completion of this research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] V.M. Hernández-López, M.L.P. Vargas-Vázquez, J.S. Muruaga-Martínez, S. Hernández-Delgado, N. Mayek-Pérez, Origen, domesticación y diversificación del frijol común: Avances y perspectivas, Rev. Fitotec. Mex. 36 (2013) 95–104.
[2] G. Freytag, D. Debouck, Taxonomy, distribution and ecology of the genus Phaseolus (Leguminosae-Papilionoideae) in North America, Mexico and Central America /George F. Freytag & Daniel G. Debouck, SERBIULA Sist. Libr. 20 (23) (2002).
[3] J. Long, J. Zhang, X. Zhang, J. Wu, H. Chen, P. Wang, Q. Wang, C. Du, Genetic diversity of common bean (Phaseolus vulgaris L.) germplasm resources in chongqing, evidenced by morphological characterization, Front. Genet. 11 (2020) 697, doi:10.3389/fgene.2020.00697.
[4] S.G. Tigist, R. Melis, J. Sibiya, B.A. Amelework, G. Keneni, A. Tegene, Genetic diversity analysis of common bean (Phaseolus vulgaris L.) genotypes for resistance to Mexican bean weevil (Zabrotes subfasciatus), using single nucleotide polymorphism and phenotypic markers, Acta Agric. Scand. Sect. B Soil Plant Sci. 70 (2020) 495–506, doi:10.1080/09064710.2020.1779339.
[5] S. Montanari, C. Deng, E. Koot, N.V. Bassil, J.D. Zurn, P. Morrison-Whittle, M.L. Worthington, R. Aryal, H. Ashrafi, J. Pradelles, M. Wellenreuther, D. Chagné, A multiplexed plant-animal SNP array for selective breeding and species conservation applications, (2022) 2022.09.07.507051. 10.1101/2022.09.07.507051.
[6] S.J. Helyar, J. Hemmer-Hansen, D. Bekkevold, M.I. Taylor, R. Ogden, M.T. Limborg, A. Cariani, G.E. Maes, E. Diopere, G.R. Carvalho, E.E. Nielsen, Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges, Mol. Ecol. Resour. 11 (2011) 123–136, doi:10.1111/j.1755-0998.2010.02943.x.
[7] I.S. Elbasyoni, A.J. Lorenz, M. Guttieri, K. Frels, P.S. Baenziger, J. Poland, E. Akhunov, A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat, Plant Sci. 270 (2018) 123–130, doi:10.1016/j.plantsci.2018.02.019.
[8] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890, doi:10.1093/bioinformatics/bty560.
[9] M. Aragón-Magadán, Marco-CNRG/beans, (2024). https://github.com/Marco-CNRG/beans (accessed May 28, 2024).
[10] Md. Vasimuddin, S. Misra, H. Li, S. Aluru, Efficient architecture-aware acceleration of BWA-MEM for multicore systems, in: Proceedings of the 2019 IEEE International Symposium on Parallel and Distributed Processing IPDPS, 2019, pp. 314–324, doi:10.1109/IPDPS.2019.00041.
[11] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297, doi:10.1101/GR.107524.110.
[12] B.J. Knaus, N.J. Grünwald, vcfr: a package to manipulate and visualize variant call format data in R, Mol. Ecol. Resour. 17 (2017) 44–53, doi:10.1111/1755-0998.12549.