First whole genome assembly and annotation of a European common bean cultivar using PacBio HiFi and Iso-Seq data

Sébastien Carrère, Baptiste Mayjonade, David Lalanne, Sylvain Gaillard, Jérôme Verdier, Nicolas W.G. Chen

Data in Brief 48 (2023) 109182

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

First whole genome assembly and annotation of a European common bean cultivar using PacBio HiFi and Iso-Seq data

Sébastien Carrère[a], Baptiste Mayjonade[a], David Lalanne[b], Sylvain Gaillard[b], Jérôme Verdier[b], Nicolas W.G. Chen[b,*]

[a] *Université de Toulouse, INRAE, CNRS, Laboratoire des Interactions Plantes Micro-organismes Environnement (LIPME), 31326 Castanet-Tolosan, France*
[b] *Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France*

ARTICLE INFO

Article history:
Received 29 March 2023
Accepted 19 April 2023
Available online 26 April 2023

Dataset link: DNA raw reads (SRR23332461) (Original data)
Dataset link: RNA raw reads (SRR23332460) (Original data)
Dataset link: P. vulgaris cv. Flavert WGS (Original data)
Dataset link: P. vulgaris cv. Flavert Annotation (Original data)

Keywords:
Whole genome sequencing
Phaseolus
Legume
PacBio
Annotation
Flavert

ABSTRACT

Common bean (*Phaseolus vulgaris* L.) is the most important grain legume for direct human consumption worldwide. Flageolet bean originates from France and presents typical organoleptic properties, including the remarkable feature of having small pale green colored seeds. Here, we report the whole-genome data, assembly and annotation of the flageolet bean accession 'Flavert'. High molecular weight DNA and RNA were extracted and subjected to long-read sequencing using PacBio Sequel II platform. The genome consisted of 566,238,753 bp assembled in 13 molecules, including 11 chromosomes plus the mitochondrial and chloroplastic genomes. Annotation predicted 29,549 protein coding genes and 6,958 non-coding RNA. This high-quality genome (99.2% BUSCO completeness) represents a valuable data set for further genomic and genetic studies on common bean and more generally on legumes. To our knowledge, this is the first whole-genome sequence of a common bean accession originating from Europe.

* Corresponding author.
E-mail address: nicolas.chen@agrocampus-ouest.fr (N.W.G. Chen).

https://doi.org/10.1016/j.dib.2023.109182

Specifications Table

Subject	Biology
Specific subject area	Agricultural and Biological Sciences, Genomics
Type of data	Whole-genome and transcriptome sequence data
How the data were acquired	Both genomic and transcriptomic data were sequenced on a PacBio Sequel II system
Data format	Raw circular consensus sequencing DNA reads (bam) Cleaned circular consensus sequencing RNA reads (fastq) *de novo* assembled genomic sequences (fasta) Annotation data (GFF3)
Description of data collection	All data come from *Phaseolus vulgaris* cultivar 'Flavert'. Genomic DNA was isolated from a pool of young leaves from 2- to 3-week-old plants. Total RNA was isolated from different tissues (*i.e.* first leaves, trifoliate leaves, roots, seedlings, flowers, young pods, developing seeds and mature seeds), then equimolarly pooled before sequencing.
Data source location	*Phaseolus vulgaris* cultivar 'Flavert' is marketed by Vilmorin-Mikado (Limagrain group, France). Plants were grown at LIPME, 31326 Castanet-Tolosan, France for genomic DNA, or IRHS, F-49000 Angers, France for RNA.
Data accessibility	All data were deposited at the National Center for Biotechnology Information (NCBI) under Bioproject PRJNA931244. Raw data were deposited at the NCBI Sequence Read Archive (SRA) under accession numbers SRR23332460 for Iso-Seq RNA reads and SRR23332461 for HiFi DNA reads. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JARGYP000000000. The version described in this paper is version JARGYP010000000. Annotation is available here: https://iris.angers.inrae.fr/bean-flavert/Annotation-EGN-EP-2.0.2.gff.gz. Assembly and annotation can be visualized using the following genome browser: https://iris.angers.inrae.fr/bean-flavert/.

Value of the Data

- This whole annotated genome assembly provides high-quality sequence data from *Phaseolus vulgaris* cv. 'Flavert'. To our knowledge, this is the first whole-genome data for a common bean of European origin. Thus, these data have the potential to reveal new genetic and evolutionary traits specific for European common beans.
- The data will be useful to both fundamental research (e.g. genomics, transcriptomics, functional genetics…) and breeding programs to improve the common bean crop (e.g. tolerance to biotic and abiotic stresses, nutritional value…).
- The data can be used to search for sequences of interest, regardless of the field of research and/or development.

1. Objective

Common bean (*Phaseolus vulgaris* L.) is a legume crop of major importance for human nutrition [1]. As a legume, it is able to perform biological nitrogen fixation through symbiosis with *Rhizobium* bacteria, thereby reducing the need of nitrogen fertilizers. Common bean yield is influenced by agronomic practices, of which breeding is of primary importance, with a focus on improvements to biotic and abiotic stresses [2]. Since the first common bean whole-genome sequencing in 2014 [3], several whole-genome assemblies have been released (https://phytozome-next.jgi.doe.gov/). However, to our knowledge, all accessions sequenced so far originate from the Mesoamerican or Andean gene pools, which correspond to the major genetic centers of diversity of common bean. Here, we released the first whole-genome assembly of a common bean accession originating from Europe, which is considered as a secondary center of diversity for common bean [4].

2. Data Description

We report the whole genome sequencing of *P. vulgaris* cv. 'Flavert'. The raw data consisted of 1,730,172 HiFi circular consensus sequencing (ccs) DNA reads with N50=17,032 bp, and 4,281,715 Iso-Seq ccs RNA reads with N50=2,459 bp. Before use, Iso-Seq reads shorter than 150 bp, bearing polyX and/or corresponding to ribosomal RNA were filtered out, resulting in 3,643,805 cleaned ccs RNA reads submitted to SRA. The assembled genome size was 566,238,753 bp, with a GC content of 35.9%, 51X mean coverage and 99.2% completeness according to BUSCO [5]. Scaffolding led to 13 molecules including 11 chromosomes plus the mitochondrial and chloroplastic genomes (Table 1). Annotation predicted 29,549 protein coding genes and 6,958 non-coding RNA. In addition, we released 1,218 unplaced, unannotated contigs totaling 49,465,140 bp.

Table 1
Assembly and annotation statistics.

Identifier	Molecule	Size (bp)	GC%	Genes	Coding	Non-coding
PvulFLAVERTChr01	Chromosome	55,176,340	35.6	3,493	3,025	468
PvulFLAVERTChr02	Chromosome	51,416,083	34.0	3,938	3,581	357
PvulFLAVERTChr03	Chromosome	55,910,709	34.5	3,504	3,219	285
PvulFLAVERTChr04	Chromosome	54,932,568	37.2	2,221	1,847	374
PvulFLAVERTChr05	Chromosome	46,978,305	37.0	2,584	2,068	516
PvulFLAVERTChr06	Chromosome	36,418,383	35.1	4,377	2,455	1,922
PvulFLAVERTChr07	Chromosome	60,817,382	36.8	3,509	2,996	513
PvulFLAVERTChr08	Chromosome	63,359,058	36.3	3,683	3,146	537
PvulFLAVERTChr09	Chromosome	39,136,314	32.5	3,162	2,930	232
PvulFLAVERTChr10	Chromosome	45,100,080	37.7	2,943	1,855	1,088
PvulFLAVERTChr11	Chromosome	56,448,427	37.6	2,765	2,254	511
PvulFLAVERTMT	Mitochondrial	395,782	45.1	179	76	103
PvulFLAVERTCP	Chloroplastic	149,322	35.4	149	97	52
PRJNA931244	Whole genome	566,238,753	35.9	36,507	29,549	6,958

3. Experimental Design, Materials and Methods

3.1. Genomic DNA and Total RNA Extraction and Sequencing

High molecular weight genomic DNA was extracted from a pool of young leaves using the protocol by Russo *et al.* [6].

Total RNA was extracted from different tissues (*i.e.* first leaves, trifoliate leaves, roots, seedlings, flowers, young pods, developing seeds and mature seeds), using the Nucleospin RNA Plant and Fungi kit (Macherey-Nagel, Hoerdt, France) following manufacturer's recommendations.

For both DNA and RNA samples, library preparation and sequencing were performed at the platform Gentyane (UMR INRAE/UCA GDEC), Clermont-Ferrand, France. An HiFi SMRTBell library (for DNA) and an Iso-Seq library (for RNA) were prepared using the SMRTbell Express Template prep kit 2.0. Single-molecule real-time sequencing was performed on a PacBio Sequel II machine using Sequel II Sequencing plate v2.0, with Sequel II binding kit 2.2, sequencing primer v5 for HiFi sequencing and Sequel II binding kit 2.1, sequencing primer v4 for Iso-Seq.

3.2. Genome Assembly and Annotation

HiFi PacBio reads were filtered using the PacBio SMRT Tools v11.0.0.146107 toolkit, then assembled into contigs using Canu v2.2 [7]. Scaffolding was performed with ALLMAPS [8] by using the masked genome of *Phaseolus vulgaris* 5-593 v1.1 DOE-JGI (https://phytozome-next.jgi.doe.gov/info/Pvulgaris5_593_v1_1) as reference, after organelle exclusion, and a set of *in-silico*

genome markers. Chromosome assembly was checked using Minimap2 [9] on the genome of *Phaseolus vulgaris* 5-593 v1.1, and manual editing was performed to fix an inverted centromeric region in chromosome 10 and a missing contig in chromosome 06.

Mitochondrial and chloroplastic reads were retrieved by mapping and Blastn (99%-100% identity over 80% length) on the reference mitochondrial and chloroplastic genomes [10,11], respectively. Subsamples to obtain 60X (mitochondrion) or 40X (chloroplast) coverage were assembled using Canu v2.2 [7]. The mitochondrion genome was circularized using a homemade script while the chloroplast genome remained uncircularized.

Structural gene annotation was performed by the using the LIPME Bioinformatics Team (https://en.lipme.fr/bioinfo) Eugene-EP v2.0.2 pipeline [12–14], which combined four training sets for structural annotation, including the cleaned Iso-Seq data (after rDNA removal, trimming of polyX, minimum size of 150bp) and three different protein databases: SwissProt, TrEMBL Plants, and predicted proteins from the *Brachypodium distachyon* reference genome. Functional annotation was done using a combination of InterPro v91 [15] and BlastP against both the 'Viridiplantae' section of non-redundant NCBI database and the UniProt 'Fabaceae' database [16]. Gene ontologies were retrieved using Blast2GO [17]. Enzyme categories were annotated using the E2P2 v3.1 prediction pipeline (https://github.com/carnegie/E2P2) [18]. Kinases, transcription factors and transcription regulators were annotated using the iTAK and PlantTFcat databases [19,20]. Gene completeness was assessed using BUSCO v5.4.3_cv1 against the embryophyta_odb10 lineage [5].

Ethics Statements

This work did not involve data from animals or human.

CRediT Author Statement

Sébastien Carrère: Conceptualization, Methodology, Software, Data curation, Writing – review & editing; **Baptiste Mayjonade**: Methodology; **David Lalanne**: Methodology; **Sylvain Gaillard**: Resources; **Jérôme Verdier**: Supervision, Writing – review & editing; **Nicolas Chen**: Supervision, Project administration, Writing – original draft, Writing – review & editing.

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.

Data Availability

DNA raw reads (SRR23332461) (Original data) (NCBI Sequence Read Archive (SRA)).
RNA raw reads (SRR23332460) (Original data) (NCBI Sequence Read Archive (SRA)).
P. vulgaris cv. Flavert WGS (Original data) (NCBI GenBank).
P. vulgaris cv. Flavert Annotation (Original data) (JBrowse).

Acknowledgments

We thank the French National Research Agency (ANR) for funding this project under the "Programme Prioritaire de Recherche" grant ID 20-PCPA-0009 – SUCSEED "Stop the Use of pestiCides

in seeds". Sequencing was performed at the GENTYANE sequencing platform (INRAE, Clermont-Ferrand, France).

References

[1] C.H. Foyer, H.M. Lam, H.T. Nguyen, K.H.M. Siddique, R.K. Varshney, T.D. Colmer, W. Cowling, H. Bramley, T.A. Mori, J.M. Hodgson, J.W. Cooper, A.J. Miller, K. Kunert, J. Vorster, C. Cullis, J.A. Ozga, M.L. Wahlqvist, Y. Liang, H. Shou, K. Shi, J. Yu, N. Fodor, B.N. Kaiser, F.L. Wong, B. Valliyodan, M.J. Considine, Neglecting legumes has compromised human health and sustainable food production, Nat. Plants. 2 (2016) 16112, doi:10.1038/NPLANTS.2016.112.

[2] T. Assefa, A. Assibi Mahama, A.V. Brown, E.K.S. Cannon, J.C. Rubyogo, I.M. Rao, M.W. Blair, S.B. Cannon, A review of breeding objectives, genomic resources, and marker-assisted methods in common bean (Phaseolus vulgaris L.), Mol. Breed. (2019) 39, doi:10.1007/s11032-018-0920-0.

[3] J. Schmutz, P.E. McClean, S. Mamidi, G.A. Wu, S.B. Cannon, J. Grimwood, J. Jenkins, S. Shu, Q. Song, C. Chavarro, M. Torres-Torres, V. Geffroy, S.M. Moghaddam, D. Gao, B. Abernathy, K. Barry, M. Blair, M.A. Brick, M. Chovatia, P. Gepts, D.M. Goodstein, M. Gonzales, U. Hellsten, D.L. Hyten, G. Jia, J.D. Kelly, D. Kudrna, R. Lee, M.M.S.S. Richard, P.N. Miklas, J.M. Osorno, J. Rodrigues, V. Thareau, C.A. Urrea, M. Wang, Y. Yu, M. Zhang, R.A. Wing, P.B. Cregan, D.S. Rokhsar, S.A. Jackson, A reference genome for common bean and genome-wide analysis of dual domestications, Nat. Genet. 46 (2014) 707–713, doi:10.1038/ng.3008.

[4] E. Bitocchi, D. Rau, E. Bellucci, M. Rodriguez, M.L. Murgia, T. Gioia, D. Santo, L. Nanni, G. Attene, R. Papa, Beans (Phaseolus ssp.) as a model for understanding crop evolution, Front. Plant Sci. 8 (2017) 1–21, doi:10.3389/fpls.2017.00722.

[5] M. Manni, M.R. Berkeley, M. Seppey, F.A. Simão, E.M. Zdobnov, BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes, Mol. Biol. Evol. 38 (2021) 4647–4654, doi:10.1093/molbev/msab199.

[6] A. Russo, B. Mayjonade, D. Frei, G. Potente, R.T. Kellenberger, L. Frachon, D. Copetti, B. Studer, J.E. Frey, U. Grossniklaus, P.M. Schlüter, Low-input high-molecular-weight DNA extraction for long-read sequencing from plants of diverse families, Front. Plant Sci. 13 (2022) 1–12, doi:10.3389/fpls.2022.883897.

[7] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation, Genome Res. 27 (2017) 722–736, doi:10.1101/gr.215087.116.

[8] H. Tang, X. Zhang, C. Miao, J. Zhang, R. Ming, J.C. Schnable, P.S. Schnable, E. Lyons, J. Lu, ALLMAPS: Robust scaffold ordering based on multiple maps, Genome Biol. 16 (2015) 1–15, doi:10.1186/s13059-014-0573-1.

[9] H. Li, Minimap2: Pairwise alignment for nucleotide sequences, Bioinformatics 34 (2018) 3094–3100, doi:10.1093/bioinformatics/bty191.

[10] C. Bi, N. Lu, Y. Xu, C. He, Z. Lu, Characterization and analysis of the mitochondrial genome of common bean (Phaseolus vulgaris) by comparative genomic approaches, Int. J. Mol. Sci. 21 (2020) 1–20, doi:10.3390/ijms21113778.

[11] X. Guo, S. Castillo-Ramírez, V. González, P. Bustos, J.L. Fernández-Vázquez, R. Santamaría, J. Arellano, M.A. Cevallos, G. Dávila, Rapid evolutionary change of common bean (Phaseolus vulgaris L) plastome, and the genomic diversification of legume chloroplasts, BMC Genomics 8 (2007) 1–16, doi:10.1186/1471-2164-8-228.

[12] S. Carrere, J. Gouzy, LIPME Functional Annotation Pipeline (1.0.1), 2023, doi:10.5281/ZENODO.7603192.

[13] E. Sallet, J. Gouzy, T. Schiex, EuGene: An automated integrative gene finder for eukaryotes and prokaryotes, Methods Mol. Biol. 1962 (2019) 97–120, doi:10.1007/978-1-4939-9173-0_6.

[14] S. Carrere, J. Gouzy, Eukaryote EuGene pipeline Version 2 (2.0.2), 2023, doi:10.5281/ZENODO.7515746.

[15] T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B.L. Pinto, G.A. Salazar, M.L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D.H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D.A. Natale, C.A. Orengo, A.P. Pandurangan, C. Rivoire, C.J.A. Sigrist, I. Sillitoe, N. Thanki, P.D. Thomas, S.C.E. Tosatto, C.H. Wu, A. Bateman, InterPro in 2022, Nucleic Acids Res. 51 (2022) 418–427, doi:10.1093/nar/gkac993.

[16] The Uniprot Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Res. 51 (2023) D523–D531, doi:10.1093/nar/gkac1052.

[17] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676, doi:10.1093/bioinformatics/bti610.

[18] C. Hawkins, D. Ginzburg, K. Zhao, W. Dwyer, B. Xue, A. Xu, S. Rice, B. Cole, S. Paley, P. Karp, S.Y. Rhee, Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae, J. Integr. Plant Biol. 63 (2021) 1888–1905, doi:10.1111/jipb.13163.

[19] Y. Zheng, C. Jiao, H. Sun, H.G. Rosli, M.A. Pombo, P. Zhang, M. Banf, X. Dai, G.B. Martin, J.J. Giovannoni, P.X. Zhao, S.Y. Rhee, Z. Fei, iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases, Mol. Plant. 9 (2016) 1667–1670, doi:10.1016/j.molp.2016.09.014.

[20] X. Dai, S. Sinharoy, M. Udvardi, P.X. Zhao, PlantTFcat: An online plant transcription factor and transcriptional regulator categorization and analysis tool, BMC Bioinformatics 14 (2013), doi:10.1186/1471-2105-14-321.