Data in Brief 66 (2026) 112746

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Whole genome sequence data on diverse pearl millet accessions across the African continent

Madhav Subedi[a], Oumar Diack[b], Adama Faye[b], Yagouba Diao[b], Arlyn J. Ackerman[c], Meseret Wondifraw[a], Elisabeth A.M.C. Diop[b], Seny Diop Mbengue[b], Elhadji Moussa Seck[b], Rebecca Cubitt[c], Jared L. Crain[d], Craig Beil[c], Moira J. Sheehan[c,*]

[a] *Breeding Insight, Cornell University, Ithaca, 14853, NY, USA*
[b] *Centre National de Recherches Agronomiques de Bambey (CNRA), Institut Sénégalais de Recherches Agricoles (ISRA), BP 53, Bambey, Sénégal*
[c] *Breeding Insight, University of Florida, Gainesville, 32611, FL, USA*
[d] *Department of Plant Pathology, Kansas State University, Manhattan, 66506, KS, USA*

ARTICLE INFO

Article history:
Received 4 March 2026
Accepted 31 March 2026
Available online 3 April 2026

Dataset link: Pearlmillet_CRCIL (Original data)

Keywords:
West Africa
High-coverage sequencing
Genotyping
Marker panel development
Genetic mapping
Mildew resistance
Adaptation
Flowering time
Pearl millet

ABSTRACT

Pearl millet (*Pennisetum glaucum* (L.) R. Br.) is a major staple cereal in Asia and West Africa but remains underrepresented in genomic research compared with other major cereals. This dataset presents whole-genome sequencing (WGS) variant data for a diverse panel of 231 pearl millet accessions collected across Africa. A total of 7.1 Tb of sequencing data were generated resulting in high-coverage (20x per sample) paired-end sequencing reads. The data were aligned to two reference genomes, '843B' and the updated 'Tift $23D_2B_1$-P1-P5' assembly, and processed variant files were generated for each reference. These variant datasets provide a foundational genomic resource and will be utilized for a targeted DArTag marker panel, with marker coordinates defined relative to both reference genomes. The panel was also phenotyped for flowering time, downy mildew response, and panicle length across four field locations in Senegal, enabling downstream association analyses. All genomic and phenotypic datasets are

* Corresponding author.
E-mail address: moirasheehan@ufl.edu (M.J. Sheehan).

https://doi.org/10.1016/j.dib.2026.112746

publicly available to support genetic research, marker development, and breeding efforts aimed at improving pearl millet adaptation and productivity.

Specifications table

Subject	Biology
Specific subject area	Whole-genome sequencing and variant discovery in pearl millet
Type of data	Raw, Analyzed, Filtered, Processed
Data collection	Leaf tissue was collected from 288 accessions and genomic DNA was extracted. Paired-end libraries were prepared and sequenced on a DNBSEQ-T7 platform. Reads were quality-filtered and aligned to the 843B and updated Tift $23D_2B_1$-P1-P5 reference genomes, followed by variant calling and filtering to generate genome-wide variants on 231 accessions with a mean sequencing depth of ~20× (total of 7.1 Tb sequence data). Phenotypic data was collected on downy mildew, flowering time and panicle length from four experimental locations in Senegal
Data source location	Field phenotyping was conducted at four locations in Senegal, West Africa, situated along a rainfall gradient. DNA extraction, sequencing and bioinformatic analyses were performed at institutions within the United States. Data are stored in public repositories hosted by Zenodo and NCBI.
Data accessibility	Repository name: NCBI SRA and Zenodo Data identification number: SRA: PRJNA1426382, Zenodo DOI: 10.5281 Direct URL to data: SRA: https://www.ncbi.nlm.nih.gov/sra/PRJNA1426382 Zenodo: 10.5281/zenodo.18602603
Related research article	None

1. Value of the Data

- We report a 7.1 Tb resource of sequence data resulting in20x whole-genome sequencing data on a set of 231 diverse accessions of pearl millet (*Pennisetum glaucum* (L.) R. Br. also known as *Cenchrus americanus* (L.)). Despite the importance of pearl millet in African diets, it remains understudied and underfunded in terms of global genomic resources. This dataset significantly increases the available genomic information for the crop.
- The diversity of accessions present in the panel will enable crucial research into identifying favourable alleles associated with climate adaptation, stress tolerance, grain yield and quality, and the discovery of native traits (e.g., disease resistance).
- The genome-wide sequencing and high depth of coverage achieved in this study will allow accurate genotype calls, discovery of single nucleotide polymorphism (SNP) markers, and presence/absence variation, allowing the data to be used to create a targeted marker panel, perform genetic mapping, engage in marker assisted breeding activities, and provide the foundational genomic data needed to pursue genomic selection for pearl millet crop improvement.

2. Background

Pearl millet (*Pennisetum glaucum*, also known as *Cenchrus americanus*) is an important grain and staple food source for Asia and West Africa. Global production is led by India contributing to 40% of the global production from 29% of harvested area (FAOSTAT, 2024). Similarly, Africa

collectively contributes to 43% of global production, with West Africa alone contributing for approximately 33%. Despite its importance culturally and nutritionally, scientific research on pearl millet has lagged behind other cereal crops (corn, sorghum, wheat and rice) in both research and investment.

There is a continuing need for research on climate adaptability and grain yield potential in pearl millet. As the crop is primarily grown in arid and semi-arid regions, synchronizing flowering time with inter- and intra-annual rainfall patterns can help for optimum plant development and yield stability. Pearl millet grown in West Africa can be predominantly categorized into two groups; the early flowering Souna morphotypes, which flower approximately 50 to 60 days after sowing (DAS) and the later flowering Sanio morphotypes, which flower around 80 to 110 DAS [1]. Studying these morphotypes can help to understand the flowering time-based adaptation mechanism of pearl millet across diverse environments.

The advancements in next-generation DNA sequencing over the past two decades has made whole genome sequencing a feasible tool for species with complex genomes, high levels of heterozygosity, or large genomes. Pearl millet has a large diploid (2n=2x=14) genome, estimated at 1.85 Gb for the 'Tift $23D_2B_1$-P1-P5' accession [2]. It is a predominantly out-crossing species due to protogynous flowering that discourages self-pollination, such that the female stigmas on the panicle are receptive before dehiscence of the male anthers on the same panicle [3]. Unlike other self-pollinating millets (e.g., proso millet, finger millet, and others), this predominates outcrossing reproductive strategy leads to high levels of heterozygosity in resulting populations and progeny, which can further complicate genomic analyses.

To date, most pearl millet genotyping efforts have relied on low-coverage or reduced-representation approaches, which limit variant discovery frequency and accuracy [4–7]. Here we describe the creation of a high-coverage whole-genome sequencing data set from 231 accessions from Senegal, West Africa, and greater Africa. The accessions include landraces important to small-holder farmers, key breeding parents and founders, and various accessions from the Pearl Millet Inbred Germplasm Association Panel (PMiGAP) population [8]. The rationale for launching this project was to 1) survey the available adapted germplasm accessible to West African breeders as well as diversity from Africa, 2) investigate key alleles linked to important traits through genome-wide-association study analyses (GWAS), and 3) develop a fixed DArTag (Diversity Arrays Technology, Bruce, Australia) marker panel that would allow affordable and timely genotyping for African pearl millet breeders to use in their crop improvement efforts through marker assisted selection (MAS) or genomic selection (GS).

3. Data Description

3.1. Germplasm and data availability

The panel comprised 288 pearl millet accessions collected from 22 different countries across the African continent (Fig. 1). These include landraces, breeding materials and improved cultivars which have been categorized into four germplasm groups: Senegalese core collection (n = 57), Senegalese germplasm (n = 114), West African material (n = 24) and PMiGAP (n = 93). The complete list of samples and metadata associated with each accession (including locations and origin) can be found at Zenodo repository (10.5281/zenodo.18602603). The FASTQ files for the 231 accessions with adequate data generated in this project were deposited in the National Center for Biotechnology Information Sequence Read Archive (SRA) under Bioproject (https://www.ncbi.nlm.nih.gov/sra/PRJNA1426382). Processed variant files are provided for alignments to both the '843B' and updated 'Tift $23D_2B_1$-P1-P5' pearl millet reference genomes [2]. These genome-wide variants will be utilized for the development of a mid-density DArTag marker panel, with marker coordinates defined relative to both reference assemblies. The panel was also phenotyped for flowering time, downy mildew and panicle length at four different lo-

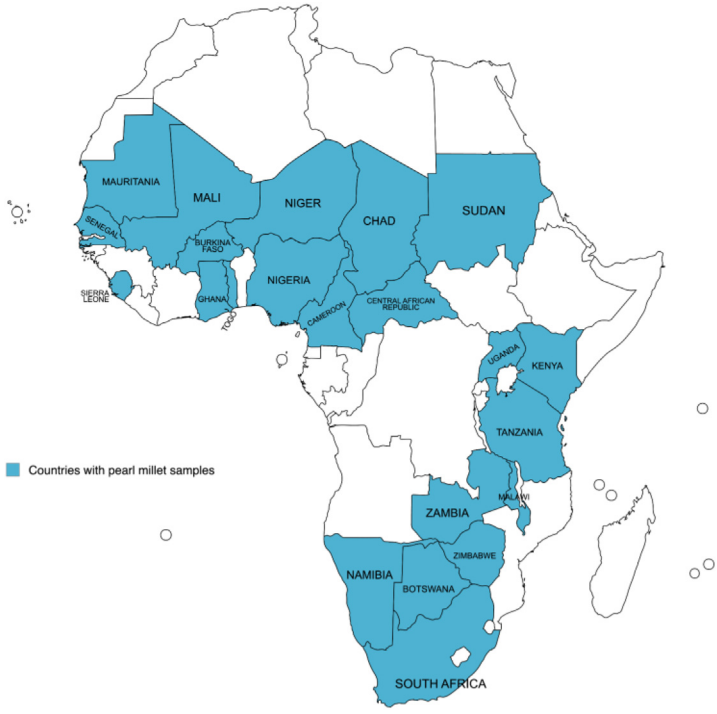

Fig. 1. Geographic distribution of pearl millet samples used in this study across Africa. Highlighted countries indicate the countries of origin of the sampled accessions.

cations in Senegal, which will be utilized for downstream association analyses. Both variant files and phenotypic datasets are also available at Zenodo repository (10.5281/zenodo.18602603).

3.2. Genetic diversity

PCA based on genome-wide SNPs discovered in this study indicate a genetically diverse pearl millet panel (Fig. 2). The PMiGAP and West African material form a cluster and are differentiated from the majority of Senegalese core collection along the PC1. Furthermore, Sanio and Souna morphotypes within the Senegalese core collection show separation along the PC2 indicating genetic differences which is consistent with known differences among these groups for agro-ecological adaptation [1,9].

3.3. Dataset overview

This dataset greatly enhances the WGS data available to the public for pearl millet. The alignment to both '834B' and updated 'Tift $23D_2B_1$-P1-P5' will provide the maximum usability to research groups using either reference genome. The distribution of the 231 accessions across the African continent reflects key pearl millet growing areas and sources of germplasm from small holders in those nations (Fig. 1). Analysis of the genomic data have already yielded some unexpected results in that the two main morphotypes, Souna and Sanio, seem to be almost completely isolated from each other (Fig. 2B), suggesting that they may be representing two heterotic

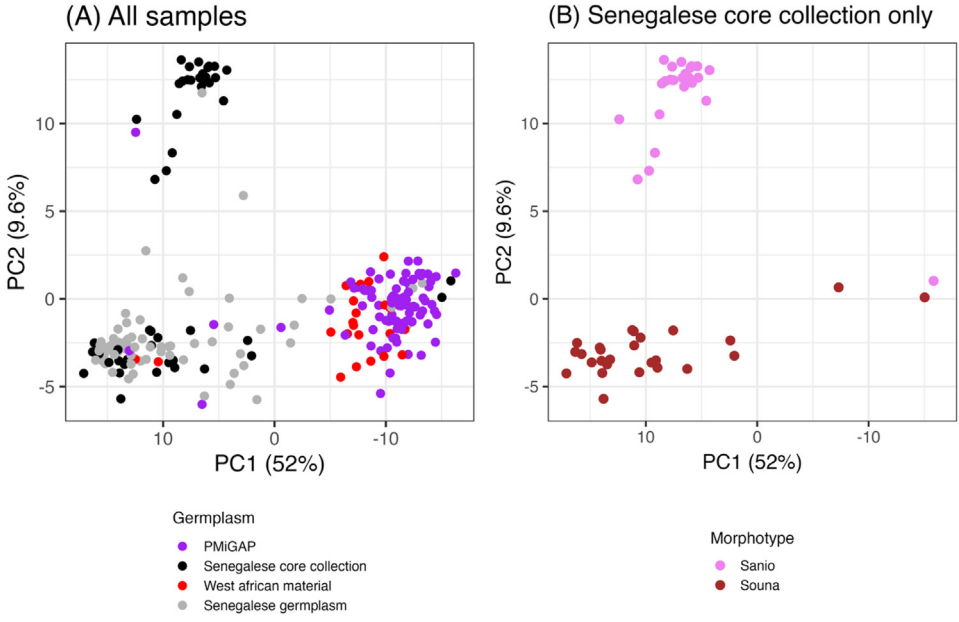

Fig. 2. Principal component analysis (PCA) of 231 pearl millet accessions based on genome wide SNP data. Colors are used to distinguish the germplasm categories (PMiGAP, Senegalese core collection, West African material, and Senegalese germplasm) and the two morphotypes, Sanio and Souna, within Senegalese core collection.

groups. Also, the passport information for the three Senegalese Core samples that cluster with the PMiGAP may be mislabelled.

The accompanying multi-location phenotypic dataset on flowering time, downy mildew and panicle length compliments the genomic dataset by enabling integrative analysis of genotype, phenotype and environment. Together, these datasets provide a comprehensive resource for future genetic and genomic studies aimed at improving crop adaptability and yield potential of pearl millet. For instance, genome-wide association studies (GWAS) will be conducted to identify significant marker–trait associations for the phenotyped traits, which can be utilized in marker-assisted selection within pearl millet breeding programs.

4. Experimental Design, Materials and Methods

4.1. Tissue sampling

Pearl millet is a highly heterozygous crop [10], and the degree of inbreeding among these accessions is uncertain, particularly for the landrace accessions. To the assay the potential genetic diversity that may exist in the available seed lots, we planted six seeds from each accession in seedling trays grown under shade cloth outdoors at the National Agricultural Research Center (CNRA) of the Senegalese Agricultural Research Institute (ISRA) in Bambey (14°42′25“N 16°28′46″W). At the 7-leaf stage, 4 weeks after sowing, we collected one hole punch per plant and pooled the six-tissue punches from each plant into a single tube of a 96-well tissue collection plate, to represent each accession. The collected tissues were desiccated by placing them in an oven at 40 °C for 48 h, followed by drying with silica gel, and then shipped to the Genomics Facility within the Biotechnology Resource Center, Ithaca, New York.

4.2. DNA extraction and whole genome sequencing

DNA was extracted from each accession using E-Z Plant DNA DS kit (https://omegabiotek.com/) and quantified using Promega Quantiflour and a plate reader, with DNA yields ranging from 0–296 ng/uL. A total of 37 samples did not yield sufficient DNA and were excluded from the analysis. Genomic DNA was shipped to Solis AgriGenomics (St. Louis, MO, USA) for whole-genome sequencing. PCR based sequencing libraries were prepared using the NEBNext UltraExpress FS DNA Library Prep Kit (New England Biolabs) according to the manufacturer's instructions. Further, 20 samples failed the library preparation step due to low fragment size. The sequencing was performed using Complete Genomics DNBSEQ-T7 (https://www.completegenomics.com/) to generate 150 bp paired-end reads targeting 20x coverage. The remaining 231 samples yielded an average of 161.3 M reads per sample (7.1 Tb total project sequence).

4.3. Bioinformatic processing

The raw data files (FASTQ) for the 231 samples were processed by Breeding Insight (RRID: SCR_026645) at Cornell University. The raw reads were taken through a series of quality control steps, including adaptor removal and exclusion of low-quality bases (Q score of <20) for read trimming using Cutadapt (v5.1). Any reads smaller than 30 bp after trimming were excluded from further analysis. Trimmed reads were aligned to the pearl millet '843B' and updated 'Tift $23D_2B_1$-P1-P5' reference genome separately [2] using BWA (v0.7.17). We obtained a minimum of 9x and a mean of 19x genome coverage across both reference genomes. Any additional reads that aligned to the reference genome with the same start and end positions as reads previously mapped were marked as duplicates using Picard-tools (v3.4.0) and removed during variant discovery. The filtered reads were used to call SNP variants across the 231 accessions using GATK (v4.6.2).

For both reference genomes, identical variant filtering thresholds were applied throughout all filtering steps. Variants were first filtered using standard GATK hard-filtering criteria, excluding sites with QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5 and ReadPosRankSum < -8.0. Variants passing these filters were further processed by retaining only biallelic SNPs and removing SNPs located within 5 bp of indels. Additionally, SNPs were filtered based on depth (5–95%), a maximum missing rate of 20%, and a minor allele frequency (MAF) $\geq 5\%$. After filtering, 22.2 M variants were retained for the 843B reference genome and 20.8 M variants for the updated 'Tift $23D_2B_1$-P1-P5' reference genome.

4.4. Phenotypic data

All 288 pearl millet accessions were evaluated during the 2024 growing season across four experimental locations in Senegal: Bambey, Nioro, Sefa, and Sinthiou maleme (Fig. 3). In each site, the experimental design was a 9×32 alpha lattice with three replications. In each replication each genotype was grown in a single row of fifteen hills. The distance between the rows and between the plants in the row was 80 cm. In each trial, flowering time (FLO), Downy mildew incidence (DM) and panicle length (LEP) were recorded following the recommendations of IPGRI and ICRISAT (1993). FLO was measured in days after sowing when 50% of the plants in a plot had reached flowering stage. DM was measured as the ratio of infected plants to the total number of plants per plot, scaled from 0 (no infection) to 1 (complete infection). LEP was the length in centimeter (cm) of the panicle at physiological maturity.

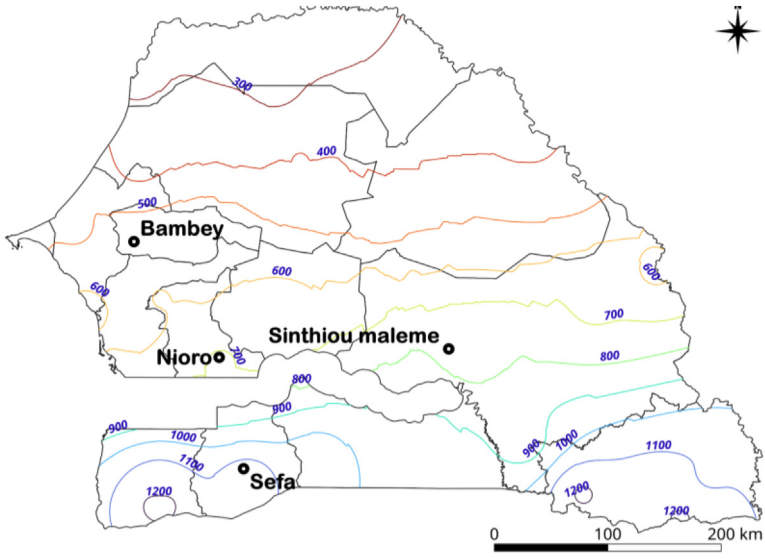

Fig. 3. Location of four pearl millet experimental sites in Senegal with mean annual rainfall (mm) isohyets.

Limitations

None.

Ethics Statement

The authors have read, and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRediT Author Statement

Madhav Subedi: Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing- Original Draft Preparation, Writing- Review & Editing. **Oumar Diack:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing- Original Draft Preparation, Writing- Review & Editing. **Adama Faye**: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision. **Arlyn J. Ackerman:** Data curation, Project administration, Supervision, Writing- Review & Editing. **Yagouba Diao:** Data curation, Investigation, Resources. **Meseret Wondifraw**: Methodology, Software, Writing- Review & Editing. **Elisabeth A. M. C. Diop**: Investigation, Resources. **Seny Diop Mbengue**: Investigation, Resources. **Elhadji Moussa Seck**: Investigation, Resources. **Rebecca Cubitt:** Resources, Writing- Review & Editing. **Jared L. Crain**: Funding acquisition**,** Writing- Review & Editing. **Craig Beil**: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing- Review & Editing. **Moira J. Sheehan**: Conceptualization, Funding acquisition, Project administration, Supervision, Writing- Original Draft Preparation, Writing- Review & Editing.

Data Availability

Pearlmillet_CRCIL (Original data) (Zenodo).

Acknowledgements

Drs. Subedi and Wondifraw were supported by the U.S. Government, Department of State under the terms of Cooperative Agreement No. 7200AA23LE00003. Drs. Ackerman, Beil, and Sheehan, and Ms Cubitt were supported through Breeding Insight (RRID: SCR_026645), a USDA-ARS initiative previously hosted by Cornell University under Cooperative Agreements (8062-21000-043-004-A, 8062-21000-052-002-A, and 8062-21000-052-003-A) and currently hosted at the University of Florida, Gainesville, under a Cooperative Agreement (8062-21000-052-020-A). The opinions expressed herein are those of the author(s) and do not necessarily reflect the views of the U.S. Government, Department of State, or the U.S. Department of Agriculture.

Declaration of Competing Interest

The author declares 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] O. Diack, et al., GWAS unveils features between early- and late-flowering pearl millets, BMC Genom. 21 (1) (2020) 777, doi:10.1186/s12864-020-07198-2.
[2] P. Ramu, et al., Improved pearl millet genomes representing the global heterotic pool offer a framework for molecular breeding applications, Commun. Biol. 6 (1) (2023) 902, doi:10.1038/s42003-023-05258-3.
[3] O.P. Yadav, et al., Genetic gains in pearl millet in India: insights into historic breeding strategies and future perspective, Front. Plant Sci. 12 (2021) 645038, doi:10.3389/fpls.2021.645038.
[4] Z. Hu, et al., Population genomics of pearl millet (Pennisetum glaucum (L.) R. Br.): comparative analysis of global accessions and Senegalese landraces, BMC Genom. 16 (1) (2015) 1048, doi:10.1186/s12864-015-2255-0.
[5] G. Kanfany, et al., Genomic diversity in pearl millet inbred lines derived from landraces and improved varieties, BMC Genom. 21 (1) (2020) 469, doi:10.1186/s12864-020-06796-4.
[6] Z. Liang, et al., Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids, G3 (Bethesda) 8 (7) (2018) 2513–2522, doi:10.1534/g3.118.200242.
[7] R.K. Varshney, et al., Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments, Nat. Biotechnol. 35 (10) (2017) 969–976, doi:10.1038/nbt.3943.
[8] D. Sehgal, et al., Exploring potential of pearl millet germplasm association panel for association mapping of drought tolerance traits, PLoS One 10 (5) (2015) e0122165, doi:10.1371/journal.pone.0122165.
[9] Y. Dussert, A. Snirc, T. Robert, Inference of domestication history and differentiation between early- and late-flowering varieties in pearl millet, Mol. Ecol. 24 (7) (2015) 1387–1402, doi:10.1111/mec.13119.
[10] F.T. Sattler, B.I.G. Haussmann, A unified strategy for West African pearl millet hybrid and heterotic group development, Crop Sci. 60 (1) (2020) 1–13, doi:10.1002/csc2.20033.