BMC Genomic Data

DATA NOTE Open Access

Development of a composite core collection from 5,856 Sesame accessions being conserved in the Indian National Genebank

Pradeep Ruperao[1], Kapil Tiwari[2,12], Vandana Rai[3], Rashmi Yadav[4], Mahalingam Angamuthu[5], Anuj Kumar Singh[2], Bhemji P. Galvadiya[2], Anshuman Shah[3], Nitin Gadol[3], Ajay Kumar[4], Rajkumar Subramani[4], Harinder Vishwakarma[4], Pradheep Kanakasabapathi[6], Senthilraja Govindasamy[5], Rasna Maurya[4], Tamanna Batra[4], Aravind Jayaraman[4], Senthil Ramachandran[7], Abhishek Rathore[8], Kuldeep Singh[7], Rakesh Singh[4], Sanjay Kalia[9], Ulavappa B. Angadi[10], Sean Mayes[1], Gyanendra Pratap Singh[4] and Parimalan Rangan[4,11*]

Abstract

Objectives A composite core collection (CCC) in sesame (*Sesamum indicum* L.) will help utilize genetic resources efficiently. This study reports, using genomics tools, a representative minimal set (CCC) that capture maximal genetic diversity from a set of 5,856 sesame accessions being conserved at the National Genebank (NGB) of the ICAR-NBPGR. The CCC will serve as a valuable resource for researchers and breeders to facilitate sesame improvement for traits such as yield, disease resistance, stress resilience, and nutritional content. Ultimately, this work contributes to the broader goal of improving sesame for an ever-increasing demand for vegetable oil, to meet our food security challenges.

Data description This study presents ddRAD-seq data for a total of 5,856 sesame accessions that includes 2,496 accessions (a subset of 5,856 accessions) that was reported by us recently. Using next-generation sequencing (NGS) short-reads over 2.16 Terabases of sequence data were generated, with each sample averaging 1.2 million reads. The study identifies a set of 1,768 sesame accessions as the CCC that captures maximal diversity, genotypic and phenotypic. This will aid researchers in trait discovery, association studies, pre-breeding, and parental selection for complex traits viz., yield, disease resistance, stress resilience, and other economically important traits.

Keywords Core collection, ddRAD, Genetic diversity, Sesame breeding, Sesame composite core collection, Trait-association

*Correspondence:
Parimalan Rangan
r.parimalan@icar.org.in
[1]Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India
[2]Bio Science Research Centre, Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar 385506, India
[3]ICAR-National Institute of Plant Biotechnology, PUSA Campus, New Delhi 110012, India
[4]ICAR-National Bureau of Plant Genetic Resources (NBPGR), PUSA Campus, New Delhi 110012, India
[5]TNAU-Regional Research Station, Vriddhachalam 606001, India
[6]ICAR-NBPGR, Regional station, KAU Campus, Thrissur 680656, India
[7]Genebank, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502324, India
[8]Excellence in Breeding Platform, CIMMYT, Hyderabad 502324, India
[9]Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi 110012, India
[10]ICAR-Indian Agricultural Statistical Research Institute, PUSA Campus, New Delhi 110012, India
[11]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, QLD 4072, Australia
[12]Institute of Biotechnology, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu (SKUAST-Jammu), Chatha, Jammu, India

Objective

Sesame, a member of the *Pedaliaceae* family, is one of the oldest oilseed crop being cultivated for its vegetable oil in the pan-tropics [1]. Sesame seed composition is unique and comprises oil (42–55%), protein (18–25%), and a high content of lignans (sesamolin, sesamin, or sesamol); hence valued for its nutritional content [1–3]. The genetic gain through breeding approaches for oilseed and pulse crops is lower when compared to cereal crops, arguing for the need of supplemental tools to accelerate crop improvement. However, its cultivation faces numerous challenges, including yield instability and susceptibility to biotic and abiotic stresses, lack of high-yielding and locally adapted varieties [4, 5].

Identification of a diverse set of accessions for desirable traits from a reservoir of genetic resources for their utilization in breeding programs will be useful in improving the genetic gain. Such collections have been the source for developing varieties with improved yield and productivity, oil quality, early maturity, and resistance to diseases/pests [4].

The core collection studies reveal significant genetic diversity and population structures. The characterization of phenotypic and molecular genetic diversities in sesame leads to the development of core collections, as reported earlier [6]. Furthermore, researchers also aim to determine diversity for specific traits of interest like seed oil content, using a broader germplasm collection [7]. Germplasm collections in large numbers warrants for a prior study on genetic diversity and population structure for its efficient utilization [8]. The NGB, located in India, conserves thousands of sesame accessions. Identifying a diverse subset using phenotypic and genotypic approaches will facilitate its use in breeding programs for crop improvement. Here, we describe a composite core collection (CCC), using genomics-assisted approaches, from a set of 5,856 sesame accessions.

Data description

Recently, we genotyped 2,496 sesame accessions using a double-digest Restriction-site Associated DNA-sequencing (ddRAD-seq) approach [9], with the methodology finalized through a pilot-scale study involving 48 accessions [10]. We report here the genotyping of an additional set of 3,360 accessions, providing 5,856 genotyped accessions in total, using the same strategy as reported earlier [9]. Here, we have pooled all these 5,856 accessions (Dataset 1 [11]), and performed genotyping analysis using short-read sequencing technology. With the ddRAD-seq approach, 2.16 terabases (Tb) of data were generated, with an average data of 369.9 Mbp per sample (DataSet 2 [11–13]). This data was used for genotyping through variant calling using Zhongzhi 13 [14] as a reference. For detailed methodology on variant calling and genomics-assisted coreset development, readers may refer Ruperao et al., (2024) [9]. The raw variants were separated as single nucleotide polymorphism (SNPs) and InDels. The SNPs were further filtered to be biallelic, with minor allele frequency (MAF) more than 0.01 and Qual more than 30 using Bcftools v 1.17 [15]. This narrowed the number of SNPs to 205,295 filtered SNPs spanning the sesame genome (Dataset 3 [11]). The frequency of SNP coverage was estimated at one SNP per 1.6 Kbp with more transitions than transversions (Dataset 3, 4; [11]). Among the genome-wide SNPs, it was observed that Chr. 2 has the largest number of SNPs (39,129), for chromosome-wise details, please refer Dataset 3. All these SNPs were structurally annotated with reference to the genomic regions they belong to, for easier further utilization (Dataset 4 [11]). Within the genome's uneven gene density (Dataset 5 [11]), 13,820 were reported as genic, and the remaining were intergenic SNPs (Datasets 4, 5, 6; [11]). Furthermore, 4,117,836 raw indels were observed with an average size of 3 bp length of indels (Dataset 3 [11]).

The SNP variants were subjected to the core-development pipeline as described in Ruperao et al. (2024) [9], wherein we have compared the diversity between total collection and core collection (generated using SNP dataset) that supports the strength of genomics-assisted core development. A genomics-assisted coreset was developed using the complete set of 5,856 accessions (1,163 accessions) (Dataset 1 [11]). In parallel, a coreset comprising of 773 sesame accessions was developed independently using phenotypic data (Dataset 1 [11]). In addition, a trait-specific set of accessions (206) was identified to possess desirable trait features (Dataset 1 [11]).

Using both these coresets (genotypic and phenotypic) and the trait-specific set, a composite coreset collection comprising 1768 sesame accessions was established after excluding overlaps (Dataset 1 [11]). This set will be of great utility for sesame researchers to utilize in crop improvement programs and trait association studies to mine novel alleles or genes and their linked markers.

Limitations

The limitation of this approach is the lack of coverage of the genome throughout. So, the results presented in this study pertains to the regions of the sesame genome covered through the ddRAD-seq approach. Although using the whole genome resequencing data would give a robust dataset for generating a CCC, when we consider the cost-benefit balance, ddRAD-seq approach is the most popular one when we handle thousands of germplasms. This is because, the ddRAD-seq approach presumes that the rate of the nucleotide variation (SNPs) across the genome is near uniform. Hence, it considers that the fractional-part of the genome is a true-representative of the whole genome with special reference to the rate of SNPs. However, the choice of the restriction enzymes for genome enrichment in a ddRAD-seq technique may vary depending on the species.

Table 1 Overview of data file/data sets

Label	Name of data file/ data set	File type (file extension)	Data repository and identifier
Data file 1	DataSet1_Final.xlsx	Excel file (.xlsx)	https://doi.org/10.21421/D2/AS65TV [11]
Data file 2a	ddRAD-seq raw data set of 2496 accessions	Fastq files (.gz)	Sequence Read Archive (http://identifiers.org/bioproject: PRJEB61739) [12]
Data file 2b	ddRAD-seq raw data set of 3360 accessions	Fastq files (.gz)	Sequence Read Archive (http://identifiers.org/bioproject: PRJEB82853) [13]
Data file 2c	DataSet2_Final.xlsx	Excel file (.xlsx)	https://doi.org/10.21421/D2/AS65TV [11]
Data file 3	DataSet3_Final.xlsx	Excel file (.xlsx)	https://doi.org/10.21421/D2/AS65TV [11]
Data file 4	DataSet4_Final.xlsx	Excel file (.xlsx)	https://doi.org/10.21421/D2/AS65TV [11]
Data file 5	DataSet5_Final.pdf	Image file (.pdf)	https://doi.org/10.21421/D2/AS65TV [11]
Data file 6	DataSet6_Final.pdf	Image file (.pdf)	https://doi.org/10.21421/D2/AS65TV [11]

Abbreviations

CCC	Composite core collection
ddRAD-seq	Double digest restriction-associated DNA sequencing
GWAS	Genome-wide association studies
Indel	Insertion-deletion
MAF	Minor allele frequency
NGB	National Genebank
NGS	Next-generation sequencing
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
Ts/Tv ratio	It is a proportion of transitions (A<->G, C<->T) to transversions (A<->C, A<->T, G<->C, G<->T) in a given set of nucleotide substitutions

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-025-01347-w.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Acknowledgements

The authors duly acknowledge the Department of Biotechnology, Government of India for financial support through the project grant number-code 16113200037-1012166.

Authors' contributions

Pradeep Ruperao: Formal analysis, Data curation, Investigation, Software, Writing-original draft, Writing-review and editing; Kapil Tiwari: Methodology, resources, project administration, supervision; Vandana Rai: Methodology, resources, project administration, supervision; Rashmi Yadav: Funding acquisition, Resources; Mahalingam Angamuthu: Methodology; Anuj Kumar Singh: Methodology; Bhemji P. Galvadiya: Methodology; Anshuman Shah: Methodology; Nitin Gadol: Methodology; Ajay Kumar: Methodology; Rajkumar Subramani: Resources, Supervision; Harinder Vishwakarma: Resources; Pradheep Kanakasabapathi: Resources; Senthilraja Govindasamy: Methodology; Rasna Maurya: Methodology; Tamanna Batra: Methodology; Aravind Jayaraman: Resources; Senthil Ramachandran: Resources, Methodology; Abhishek Rathore: Resources; Kuldeep Singh: Conceptualization, Funding acquisition, Resources, Methodology, Supervision; Rakesh Singh: Resources; Sanjay Kalia: Resources, Supervision; Ulavappa B. Angadi: Resources; Sean Mayes: Funding acquisition, Resources, Supervision; Gyanendra Pratap Singh: Funding acquisition, Resources, Supervision; Parimalan Rangan: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing-review and editing.

Funding

Department of Biotechnology, Government of India for financial support through the project grant number-code 16113200037-1012166.

Data availability

The ddRAD sequence data for 5,856 accessions were deposited to INDA with the project ids INRP000062 or PRJEB61739–2496 samples from our earlier report (Ruperao et al. 2024; http://identifiers.org/bioproject: PRJEB61739) and INRP000184 or PRJEB82853 (http://identifiers.org/bioproject: PRJEB82853)–3,360 samples from this present study. Next-generation sequencing-based ddRAD-seq is a genotyping method that utilizes a reduced representation strategy to obtain maximal information with a minimal cost.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 20 January 2025 / Accepted: 17 July 2025

Published online: 18 August 2025

References

1. Yadav R, Kalia S, Rangan P, Pradheep K, Rao GP, Kaur V, et al. Current research trends and prospects for yield and quality improvement in sesame, an important oilseed crop. Front Plant Sci. 2022;13:863521. https://doi.org/10.3389/fpls.2022.863521
2. Yermanos DM, Hemstreet S, Saleeb W, Huszar CK. Oil content and composition of the seed in the world collection of Sesame introductions. J Am Oil Chem Soc. 1972;49:20–23. https://doi.org/10.1007/BF02545131
3. Wei P, Zhao F, Wang Z, Wang Q, Chai X, Hou G et al. Sesame (Sesamum indicum L.): A comprehensive review of nutritional value, phytochemical composition, health benefits, development of food, and industrial applications. Nutrients. 2022;14:4079. https://doi.org/10.3390/nu14194079
4. Teklu DH, Shimelis H, Abady S. Genetic improvement in sesame (*Sesamum indicum* L.): progress and outlook: a review. Agronomy. 2022;12:2144. https://doi.org/10.3390/agronomy12092144
5. Rauf S, Basharat T, Gebeyehu A, Elsafy M, Rahmatov M, Ortiz R, et al. Sesame, an underutilized oil seed crop: breeding achievements and future challenges. Plants. 2024;13:2662. https://doi.org/10.3390/plants13182662
6. Zhang Y, Zhang X, Che Z, Wang L, Wei W, Li D. Genetic diversity assessment of Sesame core collection in China by phenotype and molecular markers and

extraction of a mini-core collection. BMC Genet. 2012;13:102. https://doi.org/10.1186/1471-2156-13-102

7. Teklu DH, Shimelis H, Tesfaye A, Shayanowako AIT. Analyses of genetic diversity and population structure of Sesame (Sesamum indicum L.) germplasm collections through seed oil and fatty acid compositions and SSR markers. J Food Compos Anal. 2022;110:104545. https://doi.org/10.1016/j.jfca.2022.104545
8. Seay D, Szczepanek A, De La Fuente GN, Votava E, Abdel-Haleem H. Genetic diversity and population structure of a large USDA Sesame collection. Plants. 2024;13:1765. https://doi.org/10.3390/plants13131765
9. Ruperao P, Bajaj P, Yadav R, Angamuthu M, Subramani R, Rai V et al. Double-digest restriction associated DNA sequencing-based genotyping and its applications in Sesame germplasm management. Plant Genome. 2024;17:e20447. https://doi.org/10.1002/tpg2.20447
10. Ruperao P, Bajaj P, Subramani R, Yadav R, Reddy Lachagari VB, Lekkala SP, et al. A pilot-scale comparison between single and double-digest RAD markers generated using GBS strategy in sesame (*Sesamum indicum* L). PLoS One. 2023;18:e0286599. https://doi.org/10.1371/journal.pone.0286599
11. Ruperao P, Tiwari K, Rai V, Yadav R, Angamuthu M, Singh AK et al. Development of a composite core collection from 5,856 Sesame accessions being conserved in the Indian National genebank. 2025. ICRISAT Dataverse. https://doi.org/10.21421/D2/AS65TV. Accessed 31 July 2025.
12. Ruperao P, Bajaj P, Yadav R, Angamuthu M, Subramani R, Rai V et al. The raw data fastq files for 2496 sesame accessions using ddRAD-seq approach. 2024. NCBI Sequence Read Archive http://identifiers.org/bioproject:PRJEB61739 (alternate id: INRP000062). Accessed 31 July 2025.
13. Ruperao P, Tiwari K, Rai V, Yadav R, Angamuthu M, Singh AK et al. The raw data fastq files for 3360 sesame accessions using ddRAD-seq approach. 2025. NCBI Sequence Read Archive http://identifiers.org/bioproject:PRJEB82853 (alternate id: INRP000184). Accessed 31 July 2025.
14. Wang L, Xia Q, Zhang Y, Zhu X, Zhu X, Li D, et al. Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. BMC Genomics. 2016. https://doi.org/10.1186/s12864-015-2316-4.
15. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017. https://doi.org/10.1093/bioinformatics/btx100.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.