RESOURCES

Special Section: Modern Improvement of Tropical Crops

Double-digest restriction-associated DNA sequencing-based genotyping and its applications in sesame germplasm management

Pradeep Ruperao[1] | Prasad Bajaj[1] | Rashmi Yadav[2] | Mahalingam Angamuthu[3] | Rajkumar Subramani[2] | Vandana Rai[4] | Kapil Tiwari[5] | Abhishek Rathore[6] | Kuldeep Singh[7] | Gyanendra Pratap Singh[2] | Ulavappa B. Angadi[8] | Sean Mayes[1] | Parimalan Rangan[2,9] ⓘD

[1]Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

[2]ICAR-National Bureau of Plant Genetic Resources, PUSA Campus, New Delhi, India

[3]TNAU-Regional Research Station, Vriddhachalam, India

[4]ICAR-National Institute of Plant Biotechnology, PUSA Campus, New Delhi, India

[5]Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar, India

[6]Excellence in Breeding Platform, CIMMYT, Hyderabad, India

[7]Genebank, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

[8]ICAR-Indian Agricultural Statistical Research Institute, New Delhi, India

[9]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, Queensland, Australia

Correspondence
Parimalan Rangan, ICAR-National Bureau of Plant Genetic Resources, PUSA Campus, New Delhi 110012, India.
Email: r.parimalan@icar.gov.in

Present address
Prasad Bajaj, Piatrika Biosystems, Hyderabad 500081, India.

Assigned to Associate Editor Manish Roorkiwal.

Abstract

Sesame (*Sesamum indicum* L.) is an ancient oilseed crop belonging to the family *Pedaliaceae* and a globally cultivated crop for its use as oil and food. In this study, 2496 sesame accessions, being conserved at the National Genebank of ICAR-National Bureau of Plant Genetic Resources (NBPGR), were genotyped using genomics-assisted double-digest restriction-associated DNA sequencing (ddRAD-seq) approach. A total of 64,910 filtered single-nucleotide polymorphisms (SNPs) were utilized to assess the genome-scale diversity. Applications of this genome-scale information (reduced representation using restriction enzymes) are demonstrated through the development of a molecular core collection (CC) representing maximal SNP diversity. This information is also applied in developing a mid-density panel (MDP) comprising 2515 hyper-variable SNPs, representing almost equally the genic and non-genic regions. The sesame CC comprising 384 accessions, a representative set of accessions with maximal diversity, was identified using multiple criteria such

---

Abbreviations: BWA, burrows-wheeler aligner; CC, core collection; CHGC, CoreHunter3 and GenoCore; ddRAD-seq, double-digest restriction-associated DNA sequencing; MAF, minor allele frequency; MDP, mid-density panel; NBPGR, National Bureau of Plant Genetic Resources; NJ, neighbor-joining; RAD, restriction associated DNA; SCC1, sesame core collection 1; SNP, single nucleotide polymorphism.

- -

as k-mer (subsequence of length "k" in a sequence read) diversity, observed heterozygosity, CoreHunter3, GenoCore, and genetic differentiation. The coreset constituted around 15% of the total accessions studied, and this small subset had captured >60% SNP diversity of the entire population. In the coreset, the admixture analysis shows reduced genetic complexity, increased nucleotide diversity (π), and is geographically distributed without any repetitiveness in the CC germplasm. Within the CC, India-originated accessions exhibit higher diversity (as expected based on the center of diversity concept), than those accessions that were procured from various other countries. The identified CC set and the MDP will be a valuable resource for genomics-assisted accelerated sesame improvement program.

Plain Language Summary

We have used a reduced representation strategy (double-digest restriction associated DNA sequencing) and identified the diverse set of sesame accessions (384). A set of minimal number of single nucleotide polymorphisms (SNPs) (2515) capturing maximal diversity were identified and represented equally the genic and non-genic regions having application in marker-assisted selection. These set of markers also contain SNPs linking genes associated with high oil content and can be used for screening genotypes with high oil content.

1 | INTRODUCTION

Sesame (*Sesamum indicum* L.) is one of the oldest oilseed crops being cultivated widely, grown in tropical regions across the globe, having Indian subcontinent as its center of origin and diversity along with its progenitor (Bedigian, 2003). However, Africa is the center of origin for most of the wild relatives of sesame, other than the progenitor. It is known as til in Hindi, nuvvulu (Telugu), ellu (Tamil), tal (Gujarathi), zhima (China), goma (Japan), chamkkae (Korean), and kunzhut (Russian), and it has many other names according to regional usage. Ancient Indian literature records the common use of sesame in religious ceremonies, indicating a much older (than 5000 years) era of sesame cultivation (Pathak et al., 2014). Based on the available germplasm, the core collection (CC) samples were developed using phenotype data in India (I. S. Bisht et al., 1998) and China (Xiurong et al., 2000).

The cultivation of sesame across the globe, besides Indian subcontinent, had led to the crop diversification in array pattern during its domestication process (Mondal et al., 2016; Pathak et al., 2015; Rangan et al., 2023). Variability for seed coat color, capsule length, plant type, carpel number, monostem nature, flower color, determinate nature, leaf shape, oil content, sesamin content, and other biochemical traits, abiotic and biotic stress tolerance traits, is reflected in the genebank holdings for sesame (Bhunia et al., 2015; Mondal et al., 2016; Ruperao et al., 2023; Uzun et al., 2008; Yadav et al., 2022;

Yol & Uzun, 2012). Handling these collections for efficient utilization and management requires a large-scale drive to identify the most diverse and promising accessions. Therefore, developing a CC representing the most diverse sub-set capturing most of the variability will favor efficient utilization and management of plant genetic resources (I. Bisht et al., 2004; Xiurong et al., 2000; H. Zhang et al., 2011).

The ICAR-National Bureau of Plant Genetic Resources (NBPGR) is a national institute that facilitates efficient management of the germplasm being conserved in its genebank. The NBPGR hosts nearly 10,000 sesame germplasm accessions, including a few wild relatives. The diversity of these germplasm was studied earlier in smaller numbers, not more than a few hundred (Bhat et al., 1999; Williams & Holden, 1984; P. Zhang et al., 2011). These studies involved either using a standard descriptor or using molecular markers based on phenotype scoring, especially the random amplified polymorphic DNAs (RAPDs). Although these were promising output at those time, those strategies cannot support handling the larger germplasm sets, say in thousands, that is, being conserved in the genebanks. In this scenario, the genomics-assisted approaches like double-digest restriction-associated DNA sequencing (ddRAD-seq) for genotyping in sesame would be most appropriate to handle 1000s of germplasm (Ruperao et al., 2023). The present study reports generation of sequence data at low coverage and reduced representation strategy, ddRAD-seq, for 2496 sesame germplasm and

demonstrate its application in efficient management of sesame genetic resources through development of a CC set and a mid-density panel (MDP) of markers for its future use in accelerated sesame improvement program. This will reduce the complexity of handling huge genebank collection for field-based screening to identify promising genotypes.

2 | MATERIALS AND METHODS

2.1 | Plant materials and DNA extraction

A total of 2496 sesame accessions were genotyped using ddRAD-seq approaches, as reported by us recently with a pilot-scale of 48 sesame accessions comparing ddRAD- and sdRAD-seq methods (Ruperao et al., 2023). Most accessions constituting the 2496 set were collected from India (1081), Singapore (1025), and Bangladesh (62) (Table S1). These selected sesame accessions were selfed for one generation at the Tamil Nadu Agricultural University (TNAU) regional research station, Virudachalam, to enhance the genetic purity of each of the accession studied. Hence, purified seeds were sown in a germination paper towel for its germination. The genomic DNA was extracted from the 5–7 days old seedlings using DNeasy plant mini kit (Qiagen). The quality check for the extracted DNA was analyzed using spectrophotometer and the agarose gel electrophoresis at 0.8% using 1xTAE buffer.

2.2 | Rad-seq data generation

The ddRAD data generation for the extracted genomic DNA of the 2496 sesame genotypes was outsourced to AgriGenomics Pvt. Ltd. The ddRAD data generation workflow includes the adapters preparation, similar to the earlier report (Elshire et al., 2011). The 1 µg of genomic DNA was digested with *Sph1* and *MluC1* restriction enzymes (Peterson et al., 2012), and the digested product was cleaned with Ampure beads. The ligation, pooling, size selection, PCR amplification, and QC check were done as reported by us earlier (Ruperao et al., 2023). The final pooling and sequencing were performed on a HiSeqX or NovaSeq6000 platform.

Raw sequence data were quality trimmed using trimmomatic (Bolger et al., 2014) with low-quality bases (Q20) and adapters if any were removed, a sliding 4 bp window was applied to trim the bases when the average quality score dropped below 15, and the remaining clean reads were mapped to the sesame reference genome assembly (L. Wang et al., 2016) with BWA mem (Li & Durbin, 2010). The basic stats for raw reads were generated with the in-house developed script (Raspberry) (https://github.com/CEG-ICRISAT/Raspberry) and quality check was performed with fastqc (Andrews, 2015), and results were compiled and assessed

Core Ideas

- A coreset of 384 genetically diverse sesame accessions was identified for crop improvement.
- A set of 2515 single nucleotide polymorphisms (SNPs) capturing maximal diversity were identified and selected for a mid-density panel.
- These selected sets of SNP markers include genes associated with high oil content for oil enhancement.
- Double-digest restriction associated DNA sequencing (ddRAD-seq) SNPs covers ca. 16% of the total protein-coding genes, though they cover only 2.81% of the sesame genome.
- Accessions from India and Singapore were represented majorly in the core collection.

with multiQC (Ewels et al., 2016). The mapping rate was assessed with qualimap (Okonechnikov et al., 2016), Samtools (Li et al., 2009), and the variants were called using Stacks pipeline (Catchen et al., 2013) with sequence quality filtering (minimum Phred scores 20, allowing a single mismatch in the adapters). The SNPs were filtered to have the biallelic variants with a call rate of at least 70% in the population with minor allele frequency (MAF) >0.01 frequency using the vcftoolsv-0.1.17 (Danecek et al., 2011) and annotated with the in-house developed script and SNPeff (Cingolani et al., 2012). The density plots were generated with circus (Krzywinski et al., 2009), and the R package "euclidean" method with 1000 bootstrap was used to estimate the genetic distance between the accessions and NJ tree created with the R "amap," "labdsv," and "ape" packages and visualized using an iTOL tree viewer (Letunic & Bork, 2019). The population structure analysis was performed with Admixture using default parameter settings (Liu et al., 2020). The raw data for the 2496 samples along with their IDs are listed in Table S10.

2.3 | K-mer analysis

For the k-mer analysis, the cleaned reads were subjected to the k-mer count in each accession, and distinct k-mers were identified using Jellyfish (Marçais & Kingsford, 2011) (bloom filters: -m 27 -s 100G -C -t 25 -F 2 and count filters: -m 27 -s 100G -C -t 15). The single copy of the k-mer in an accession was filtered out as it was likely to be an artifact sequence or error. The distinct k-mers were compared between the accessions to find the conserved k-mer set and the most variable k-mers (a k-mer is absent in at least one of the accessions). The variable k-mers frequency was measured, and the accessions were ranked accordingly to find the distinct accession. The

k-mers appearing in more sesame samples were considered the conserved sequence, whereas the sequences appearing in fewer samples are the variable sequences. The higher frequency of unique k-mer sequence in samples was reported as more diverse than the lower frequency k-mers.

2.4 | Core collection

The CC was defined with diverse sesame accessions identified with different factors, including the CoreHunter3, GenoCore, heterozygosity, variable k-mer sequences, genetic distance (Euclidean, correlation, Manhattan, maximum, and Pearson), and Mash distance. Initially, the construction and composition of the CC were identified with the commonly used R package "CoreHunter3" (de Beukelaer et al., 2018) and GenoCore (Jeong et al., 2017). The CoreHunter3 and Geno-Core reported the list of core components (list of accessions), and the common accessions were used as the initial coreset. Later, the uniquely identified accessions from both the software, were pooled and examined for heterozygosity, followed by genetic distance. The two set of accessions were combined to make the final CC set (Figure 3A). The resulting coreset of accessions was examined for any biological and technical replicates in the CC since the strategy used was reduced representation. The genome-wide heterozygosity was called using vcftools (Danecek et al., 2011), and the accessions having >60k variant sites with $H_O > H_E$ was used for core accessions' assessment. Shannon's diversity index is the mathematical method to measure the diversity index in a population (Shannon, 1948).

SNP coverage is the percentage of retained SNP markers in core accessions relative to the SNPs observed in the full collection. The SNP coverage was measured using the vcftools package (Danecek et al., 2011; Thachuk et al., 2009).

2.5 | Mid-density panel

From the overall SNP calls, the monomorphic and co-dominant SNPs were removed and retained only biallelic SNPs. The 70% overall call rate as the cut-off with a mean depth of four per site was retained to further filter the minor allele count or frequency of 5%, MAF 0.05, transversions, >20% heterozygosity, and ambiguous calls on 100 bp on both sides of flanking regions were used to filter the set of SNPs for an MDP. The flanking sequences were extracted for the finally selected MDP's SNPs set, and the GC count was measured with in-house developed scripts. The above rules were applied to reduce the genome-wide SNP density with vcftools (Danecek et al., 2011), bcftools (Danecek et al., 2021), and in-house developed bash on the fly scripts. A matrix containing the SNP information for all the 64,910 SNPs (in each row)

against all the 2496 accessions (in each column) is provided along with a header note for the information on the set of SNPs included in MDP and the set of accessions included in the derived coreset (Table S11).

3 | RESULTS

3.1 | Genotyping

A total of 2496 sesame accessions being conserved at the national genebank of the ICAR-NBPGR is genotyped through this study to develop the hyper-variable single nucleotide polymorphisms (SNPs) for its application to identify the set of genetically diverse accessions for use in crop improvement programs, commonly called as CC. Also, a set of SNPs were selected for a panel of markers at mid-density level. These 2496 accessions represented various countries and constituted mainly from India (1078) and Singapore (1025), while Bangladesh (62), China (19), Sri Lanka (18), and other countries have minor representations (Figure 1A) (Table S1). These sets of accessions also include some replicates for a few accessions (20). From the 2496 accessions, a total of 8419 million raw sequence reads were generated using ddRAD-seq strategy, and the sequence data were acquired through 150 bp paired-end sequencing on the Illumina sequencing platform. The raw data are available in the public database through the following project IDs: INRP000062 or PRJEB61739.

3.2 | SNP calling

The sequence reads for 2496 sesame accessions were filtered to exclude low-quality reads, and adaptor sequences were removed, resulting in 5655 million reads (filtered reads). A total of around 5109 million (mapped reference regions) RAD tags were obtained from the 2496 sesame samples, with an average of 2 million per sample. The RAD tags were mapped to the sesame reference genome of the cv. Zhongzhi13 (L. Wang et al., 2016) using the software burrows-wheeler aligner "(BWA)" (Li & Durbin, 2010), with an average mapping rate of 86% reads mapping to the reference sequence (Tables S2 and S3). The sequence reads from these samples got mapped to the reference with an average depth of 14.32x vertically and covered each chromosome horizontally, an average of 564.83 kb (Figure 1B). The SNPs were called and from the raw 290,527 SNPs, a total of 64,910 high-quality SNPs (with biallelic and 70% genotype call rates) were identified among sesame 2496 accessions. Of these RAD tags containing polymorphic SNPs (62,881), 96.87% were located on the 13 assembled chromosomes of the sesame genome, while the remaining polymorphic SNPs (2029) were from the unassigned scaffolds (Figure 1C). The number of SNPs per
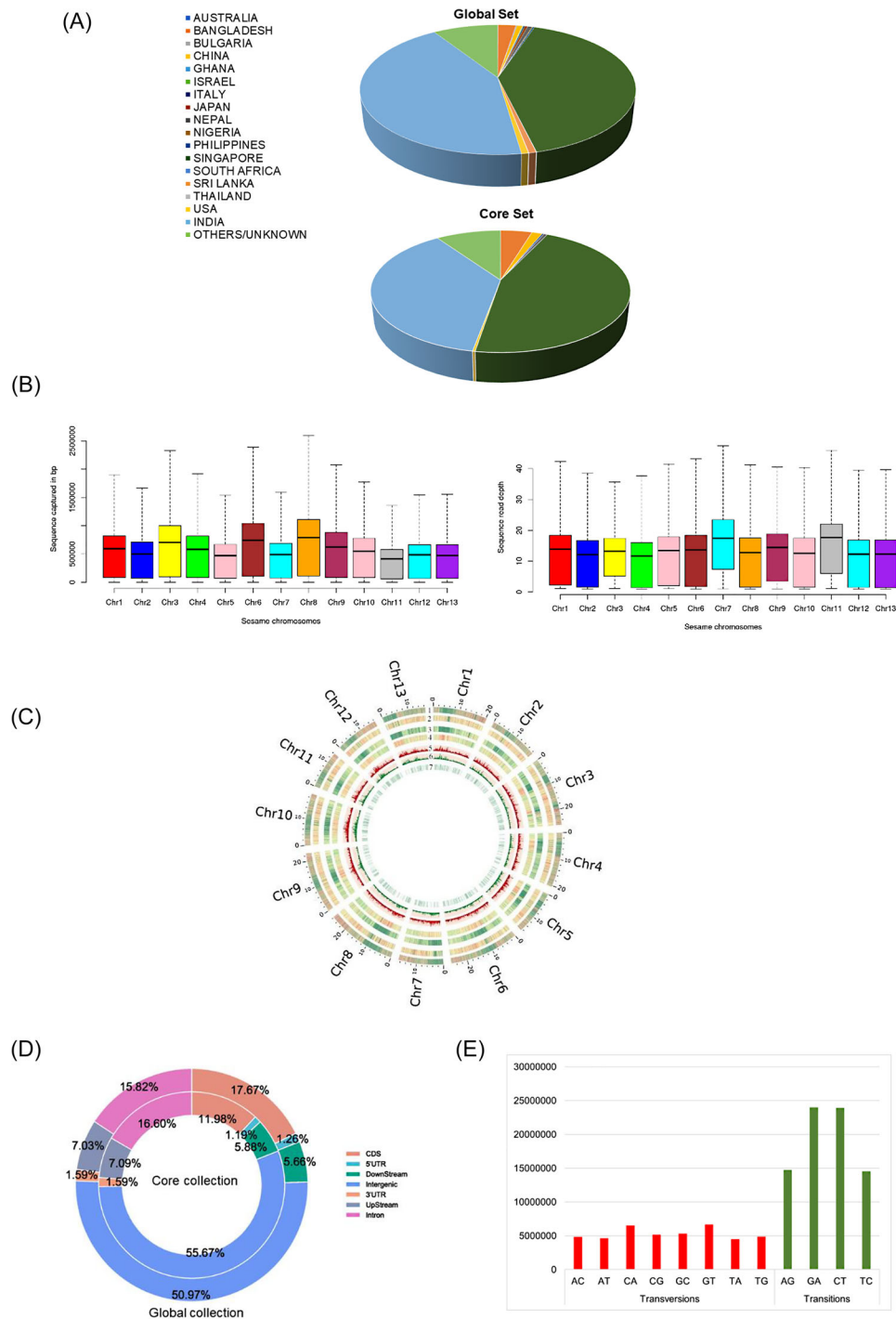
FIGURE 1 (A) Geographical distribution of the 2496 sesame accessions and its representation in the generated coreset. (B) Reference genome sequence captured with double-digest restriction associated DNA sequencing (ddRAD) sequence data (horizontal and vertical coverage). (C) A circus plot with the distribution of genes, restriction associated DNA (RAD) tags, and single nucleotide polymorphisms (SNPs) at different levels. Heatmaps produced with the color scale of red for maximum, yellow for medium, and green indicates the low value (1. Gene density, 2. Overall RAD tags on the reference genome, 3. RAD tags in the genic region, 4. Intergenic RAD tags, 5. Overall SNP density, 6. SNP density from the core accessions, 7. Mid-density panel SNPs distribution on the sesame reference genome). (D) SNP annotations between global (outer ring, 2496 accessions) and core (inner ring, 384 accessions) collection with distribution pattern among the intergenic, and (E) the overall transitions and transversions count in the global collection.

TABLE 1 The distribution of the restriction associated DNA (RAD) tags is called single nucleotide polymorphisms (SNPs) across the sesame genome reference (Zhongzhi13). In last row, 4436 represents number of unassigned scaffolds in the reference genome.

Chromosome	Length (bp)	Number of SNPs	SNP density/kb	Transition SNPs (Ts)	Transversion SNPs (Tv)	Ts/Tv ratio
Chr1	20,257,639	4446	0.22	2868	1578	1.82
Chr2	18,415,740	4240	0.23	2699	1541	1.75
Chr3	25,850,335	4993	0.19	3177	1816	1.75
Chr4	20,582,917	5031	0.24	3302	1729	1.91
Chr5	16,584,689	4641	0.28	3041	1600	1.90
Chr6	25,967,286	5728	0.22	3647	2081	1.75
Chr7	16,756,707	4475	0.27	2909	1566	1.86
Chr8	26,180,356	5976	0.23	3745	2231	1.68
Chr9	22,847,643	4996	0.22	3181	1815	1.75
Chr10	19,487,738	5848	0.30	3841	2007	1.91
Chr11	14,047,238	3595	0.26	2370	1225	1.93
Chr12	16,275,532	4718	0.29	3092	1626	1.90
Chr13	16,472,772	4194	0.25	2708	1486	1.82
Chr00	13,008,389	2029	0.16	1290	739	1.75

chromosomes ranged from 3595 (chr 11) to 5976 (chr 8). The average SNP density was 0.24 per kb, with the lowest and highest SNP densities observed on chromosomes 3 and 10, respectively, of 0.19 and 0.30 (Table 1). Roughly, half of the SNPs (50.97%) were identified to be from intergenic regions, while nearly one-fifth SNPs in the coding region (coding DNA sequence [CDS], 17.67%) (Figure 1E). The maximum and minimum SNP frequency occurred as G/A and T/A, respectively (Table S2) (Figure 1E). The proportion of transition (average 30,905 sites per sample) was greater than transversions (average 16,994 per sample) (Table S3).

3.3 | Genomic distribution and annotation of SNPs

Based on the reported sesame genome annotation result, 27,150 protein-coding genes were distributed on the 13 chromosomes, with the gene number per chromosome ranging from 1403 (Chr 11) to 2696 (Chr 6) (L. Wang et al., 2016). Of the 64,910 SNPs located in the genome-wide polymorphic RAD tags, 49.03% were located within genic regions of 4270 genes (5.6 SNP markers per gene on an average). Although average horizontal genome coverage for all 2496 samples is 2.81% (Table S4), the ddRAD-seq strategy helped cover roughly 16% of the total protein-coding genes annotated. This underscores the efficiency or cost-effectiveness of the reduced representation strategy used in our study. Among all the chromosomes, chromosome 6 carried the maximum number of SNPs (2490) within the genic regions (457 genes), with an average of 5.44 SNPs per gene, and is followed by chromosome 3 (2467 SNPs in 451 genes) and chromosome 8 (2337

SNPs in 439 genes). The minimum number of SNP markers per chromosome is for chromosome 13, with 1196 SNPs. So, across the sesame genome, there is an average number of 1832 SNPs per chromosome within the genic regions.

3.4 | Defining the core collection of the 2496 sesame accessions

Many approaches for selecting a CC using SNP datasets were proposed, and the most commonly used were CoreHunter3 (de Beukelaer et al., 2018), PowerCore (Kim et al., 2007), and GenoCore (Jeong et al., 2017). Different data types are also in practice to define the core samples, such as genealogical data in wheat (*Triticum aestivum* L.) (Stehno et al., 2006), agronomic data in groundnut (*Arachis hypogaea* L.) (Kuo et al., 2021), and molecular data in rice (*Oryza sativa* L.) (J.-C. Wang et al., 2007). In this study, besides using these CC tools, a coreset was designated with 384 accessions by combining multiple approaches such as heterozygosity, genetic distance (k-mer [subsequence of length 'k' in a sequence read] mash and distance calculation methods Euclidean, Manhattan, Pearson, etc.), accessions with distinct k-mer content, prior sample information such as geographical origin, and sample replicate information.

3.5 | Heterozygosity analysis

In the core and complete sesame sample set, the 42 and 947 SNPs exhibit heterozygosity in at least 20% of the samples, respectively. The heterozygosity assessment showed that

1143 accessions exhibit a higher observed heterozygosity (H_O) value than the expected heterozygosity (H_E). While the remaining accessions exhibited a lower H_O than the H_E values. The negative inbreeding coefficient (F-value) for 564 accessions indicates a low level of inbreeding, hence a higher gene flow for these accessions compared with other accessions (Table S5).

3.6 | K-mer analysis

The ddRAD-Seq data for 2496 sesame accessions were subjected to k-mer analysis (Ruperao et al., 2023), resulting in 19.42 billion k-mer sequences with an average of 7.7 million distinct k-mers per accession. On the k-mer sequence comparison between the 2496 accessions, 189 million k-mers were commonly shared between 2494 (maximum) accessions indicating the conserved k-mers (possibly belonging to the core-part of the genome), 6 million k-mers are shared among 70% (1747) of accessions as a softcore k-mer sequence, and 562 million k-mers were present only in 30% of sesame accessions. This is an indirect measure of the sequence variability between the sesame accessions reported in this study (Figure 2A). For example, the 175 million k-mers were commonly reported from any two sesame accessions, but the rate of k-mer sequence commonness descends with the increased number of accessions, as expected. The variable or diverse k-mer sequences were informative and are an indirect representation of the genetic diversity among the accessions studied. This strategy identifies the set of accessions that carry the diverse k-mer sequences to exhibit the most diverse set among the total studied accessions. The 748 accessions with the most diverse k-mers were identified as the representative core samples, on k-mer basis, from the 2496 accessions studied.

On the other hand, the sketch of k-mers was constructed from the collection of k-mers in the reads and compared to the reference genome of the cv. Zhongzhi 13 (L. Wang et al., 2016) sketch database. The pairwise genetic distance and p-value significance test enabled clustering and estimated the genetic distance, which helped identify the 748 most diverse accessions. Of which, 124 accessions (5%) of the sesame population were reported to be the most genetically distant accessions.

3.6.1 | Pairwise genetic distance computation

To find the most diverse accessions, the genetic distance matrix was computed through many iterations ($n = 1000$) with five methods (Euclidean, Manhattan, Pearson, correlation, and maximum distance method) (https://CRAN.R-project.org/package = labdsv) (Figure 2B). The distance between the samples were compared between the methods and compiled to score each accession (Table S6). A total of 450 accessions exhibited divergence supported by a minimum of two of the five distance methods studied. The 98 accessions were found to be the most common set of diverse accessions identified through all the five methods (indicating the stability of these methods on input data). A maximum of 393 accessions were identified to be in common between Euclidean and Manhattan distances. The Euclidean distance method ranged between 377 (maximum, B32-Z28) and 4.42 (minimum, N96-Z78). A similar pattern of accessions was also observed in the Manhattan distance method (but with a wider range 19–74482), indicating that these results also support the Euclidean distance method.

3.7 | CoreHunter3 and GenoCore (CHGC) classification

CoreHunter3 (de Beukelaer et al., 2018) analysis had identified 749 accessions as the coreset from the set of 2496 sesame accessions studied, while the GenoCore (Jeong et al., 2017) analysis had identified 1471 accessions and is >50% of the accessions studied (Figure 2C). CoreHunter3 uses a non-deterministic algorithm to report core accessions, while GenoCore uses modified statistical measures related to genetic coverage and diversity. In comparison, 651 common accessions among both methods were identified and chosen for further analysis to classify the coreset that could be of potential use for the sesame improvement program.

3.7.1 | Sample replication of the coreset

Twenty-one accessions were chosen (seven biological and 14 technical replicates) within this study, at random, to replicate the experiment. The 286 common diverse accessions predicted with the k-mer diversity analysis and combined CoreHunter3 and GenoCore analysis, I58 biological replicated, were excluded for the further downstream filtering analysis. From the overall 384 coreset, with two (M32-H69 and N95-I58) and six (N56-A25, N62-A76, N72-F70, O04-J10, O16-N71, and U22-T46) biological and technical replicates, respectively, retaining a single copy of accession in the final coreset indicates the representative sample from the global sesame collection. Although these were replicates, reason behind their representation in the coreset is analyzed at sequence level, and we found that there are mutually exclusive genomic regions to which the replicated samples' RAD tags were mapped. Hence, the tools subjected to identify the coreset identified them as diverse accessions. This is detailed in Section 4.1.
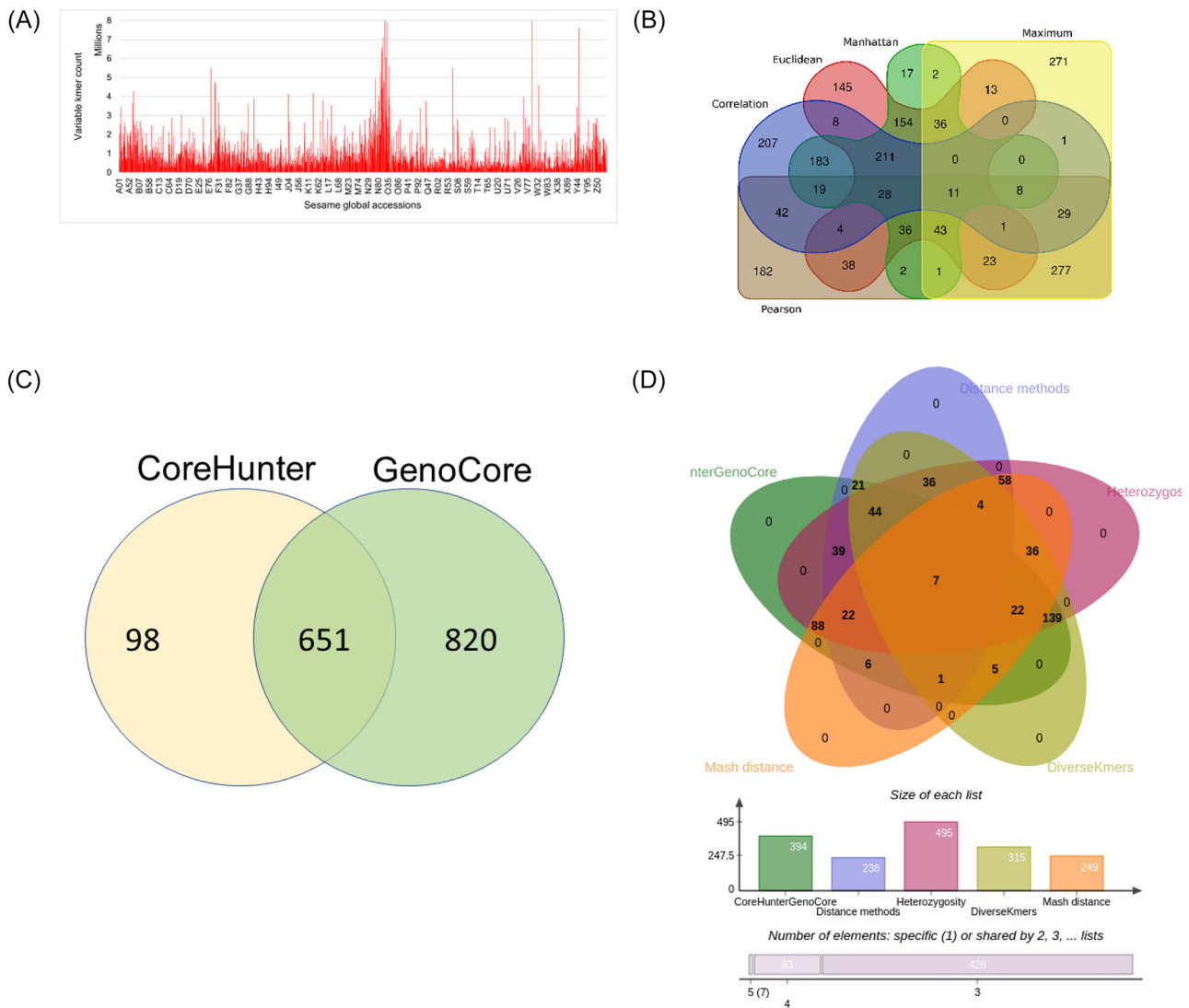
FIGURE 2 (A) Variable subsequence of length 'k' in a sequence read (k-mer) counts per accession. (B) Genetic distance computing methods comparison for the set of genetically distant accessions identified from each method. (C) CoreHunter3 and GenoCore assessment. (D) Venn and histogram showing the comparison of representative samples obtained using all the five factors.

3.7.2 | Evaluation of core

Using five different methods (CHGC, k-mer, heterozygosity, genetic distance using five measures, and Mash distance; Figure 2D), the candidate coreset has been designated, and using the common and unique factors, a final designated coreset of 384 accessions using the SNP dataset is identified. Here, four different methods yielded a different set of coresets, namely, CHGC (651 accessions), k-mer diversity (748 accessions), heterozygosity (1930 accessions), k-mer-based mash distance, and distance-based analysis-pool of five measures (450 accessions) (Figure 2D). Since the mash genetic distance calculated between the reference genome and sesame accessions, which do not explain the distance between the sesame accessions, the mash distance method was excluded for subsampling the sesame population. Though there was a common

set of accessions among all four methods used, the equal number of different accessions was there. So, we devised novel strategies to combine these methods in a stepwise fashion and applied the appropriate filters in each step to streamline the coreset designation process.

In the first step, a comparison of the set of accessions between the CHGC and k-mer diverse samples was made since CHGC is the specialized core methods and k-mer-based method is a novel application in diversity assessment (Ruperao et al., 2023). This revealed a total of 286 accessions in common between these two methods, CHGC (651 accessions) and K-mer (748 accessions). A cross-check for biological or technical replicate accessions was made and found that an accession, I58, is a biological replicate of N95, and hence former was excluded to avoid the accession duplication and yielded 285 accessions as a prospective core. In the
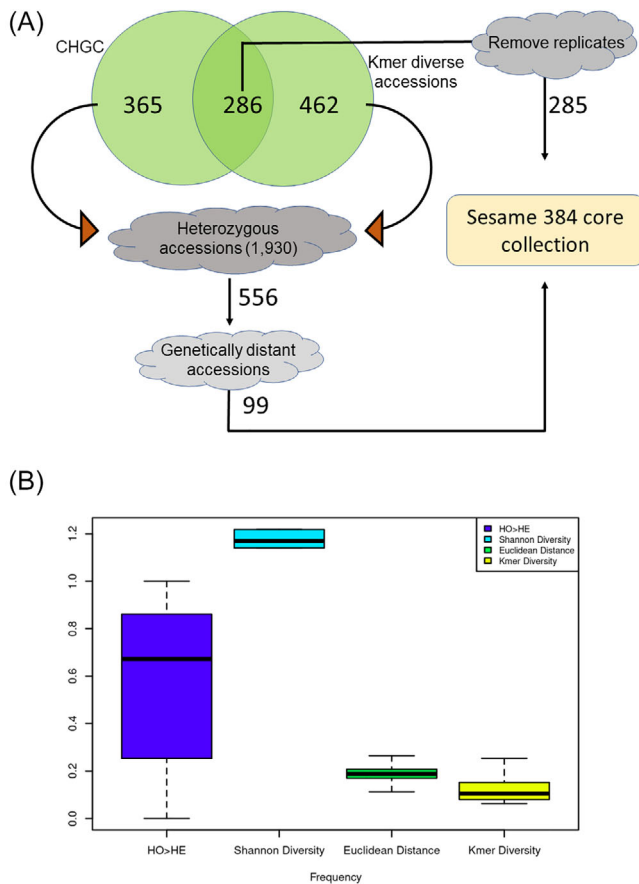
FIGURE 3 (A) Sesame core collection 1 (SCC1) development flow chart and (B) SCC1 frequency evaluation with observed heterozygosity (H_O), Shannon–Weaver index, Euclidean distance, single nucleotide polymorphism (SNP) coverage and subsequence of length 'k' in a sequence read (k-mer) diversity.

second step, the set of excluded (unique) ones from CHGC (365 accessions) and K-mer (462 accessions) were pooled together and tested for the common set of accessions among these 827 accessions with the core based on heterozygosity (1930 accessions), and 556 accessions were found to be in common between heterozygosity and (CHGC & k-mer combined). These 556 accessions were identified to exhibit higher H_O. In the third step, these selected 556 accessions were further compared for commonalities with the core derived from the distance-based analysis pool of five measures (450 accessions). This helped identify the set of 99 accessions exhibiting higher H_O and were genetically distant as well. These 99 accessions were augmented to the initially identified prospective coreset (285 accessions) to obtain the final designated coreset containing 384 accessions, which is 15.38% of the 2496 accessions studied (Figure 3A). These final 384 accessions are designated as the sesame core collection 1 (SCC1), the true representative set of the 2496 accessions capturing maximum allelic variation with a minimum number of acces-

sions (Figure 3A) (Table 2). Among the 2496 accessions, four samples (U11, Y31, Y49, and Y52) were considered as the most diverse accessions with all five factors of the distance-based measures and are part of the 384 SCC1, which acts as an indirect validation (Figure 3A) (Table S6). Among the SCC1, CHGC and k-mer diversity analyses report 89.84% and 84.37% of diverse accessions, respectively, with 99.70% accuracy (Table 2).

The diversity was assessed among the set of SCC1 accessions using the Shannon–Weaver diversity index, and the results indicate a high diversity ($H = 1.218199$) (Figure 3B), and Shannon–Equability index (EH) was 0.58583 for the accessions included in the SCC1.

Also, SNP coverage was measured for the initial collection (2496 accessions) and compared with the SNP coverage from the SCC1. The SNP coverage measures the proportion of the SNPs retained in the coreset relative to the initial collections. From the genome-wide variant calling, on average 3683 SNPs per site (total SNPs count across all samples) were recorded (with an average maf of 0.0164947), of which 981 (on average) SNPs were retained in the SCC1 per site (with an average maf of 0.0252642), indicating 26.64% of overall SNP coverage (Figure 3B).

3.8 | Population structure analysis in sesame

Using this dataset, we had studied these 2496 accessions for their population structure behavior for both initial and SCC1 sets using principal coordinate analysis (PCoA). After applying the SNP filter, we obtained 64,910 and 41,347 SNPs to conduct PCoA on global and SCC1 (Figure 4A). This indicates that the SCC1, comprising 15.38% of the total collection (2496 accessions), retained 63.7% of the filtered SNPs. The genetic variation was explained by 10 principal components (PC1–10), and the three components explained 58.96% and 77.70% of the variation for global and SCC1. After enriching the diverse accessions in the SCC1, the PC1 alone accounts for 64.44%, PC2 (7.10%), and PC3 (6.15%). Further, we investigated the population structure using ADMIXTURE1.3.0 (Liu et al., 2020), assuming the subpopulations were in the range of 2–20. The lowest cross-validation error was observed when 14 genetic groups were assumed ($K = 14$) for the global set, whereas for the SCC1, $K = 7$ was defined (Figure 4B,C). From this analysis, no obvious geographical pattern was found within total and SCC1 accessions.

3.9 | Genetic diversity and phylogeny

The estimates of nucleotide diversity (pi) for the total collection were 0.0274378, whereas the core accessions were

TABLE 2 A comparison on the sesame coreset on its geographical distribution.

	CHGC	k-mer diversity	$H_O > H_E$	Genetic distance (5 methods)	Final set of accessions
Initial core collection[a]	651	748	1930	450	
Finalized core samples (SCC1)	345	316	314	173	384
Origin					
Bangladesh	14	14	16	11	18
China	4	6	5	4	6
India	145	132	154	105	172
Italy	1	1	1	1	1
Nepal	2	2	2	0	2
Singapore	170	153	129	48	175
USA	1	1	1	0	1
Unknown country of origin	8	7	6	4	9

Abbreviations: k-mer, subsequence of length "k" in a sequence read; H_E, expected heterozygosity; H_O, observed heterozygosity; SCC1, sesame core collection 1.

[a]A initial representative sample identified with different methods individually.

significantly higher (0.0488387) than the overall accession collection. In SCC1, the pi values were higher for accessions collected from India (0.0493762) when compared to Singapore (0.0466954) and Bangladesh (0.0422634). This indicates that 172 and 175 accessions were represented in the SCC1 from India and Singapore (Table 2), respectively,; since more heterozygous accessions were represented from India (154) when compared to the ones from Singapore (129), nucleotide diversity is higher. The higher pi value and higher heterozygosity value of the Indian accessions when compared with the accessions from other countries indicate that collection from India was genetically more diverse among the set of accessions studied. Our exotic collection database suggests that the germplasm imported from Singapore is a set of material augmented at the regional level (South–East Asia, Oceania, and Pacific Islands) by the then Asia, Pacific, and Oceania Regional office of the International Board for Plant Genetic Resources, Singapore. These regions could only represent the secondary center of diversity, as the progenitors of this crop ([*S. indicum* subsp. *malabaricum* (Burm.) Bedigian] and *S. mulayanum* N.C. Nair) are evidently not distributed (naturally) here. These diversity values indicate that the collections from India represented in the SCC1 are highly diverse compared to the accessions from other countries (Table 2). The increased pi values in SCC1 indicate the accumulation of the diverse set of accessions in the coreset (Figure 4D). The overall inbreeding coefficient of total collections was significantly lower (below zero) for most accessions, indicating that their H_O is greater than the expected heterozygosity. This also underscores the cross-movement of alleles in a population leading to outbred nature and up to 50% cross-pollination in sesame was reported (Van Rheenen, 1980),

which corroborates with the inbreeding coefficient values obtained in our study (Figure 4E). Differentiation coefficient (F_{ST}) among total and CC is 0.001218 (mean F_{ST} estimate), indicating there is not much deviation among these two sets.

To visualize the genetic relatedness among accessions, a neighbor-joining (NJ) tree was generated with Euclidean function from labdsv package in R (described in methods session). The NJ tree consists of 2496 sesame accessions grouped into five clusters. All groups had accessions represented from different geographic origins, indicating these accessions cannot be clustered geographically. This underscores the lesser diversification because of lesser artificial selection pressure on various traits (Figure 4F). Alternatively, the sesame crop might also resist selection pressure, as is a known fact among the oilseed and pulse crops being reflected in its lesser variability (Rangan et al., 2023).

3.9.1 | Mid-density panel (SNP panel for chip)

A total of 2515 SNPs were found to be the most diverse markers of biallelic nature, non-transversion, and non-ambiguous SNPs, which are represented from most of the diverse accessions. Of these 2515 SNPs, 2489 were chromosomal SNPs, with Chr6 (255) and Chr7 having the maximum and minimum SNPs (Figure 1C). The 100 bp of both flanking sequences for each marker has an average GC percent of 41.28 (Table S7). Among the set of 2515 SNPs, 1360 SNPs were inter-genic and 1155 SNPs were intra-genic nature (Table S8). This indicates that the selected MDP set of SNPs exhibit near-equal representation of the genic and non-genic regions across the whole
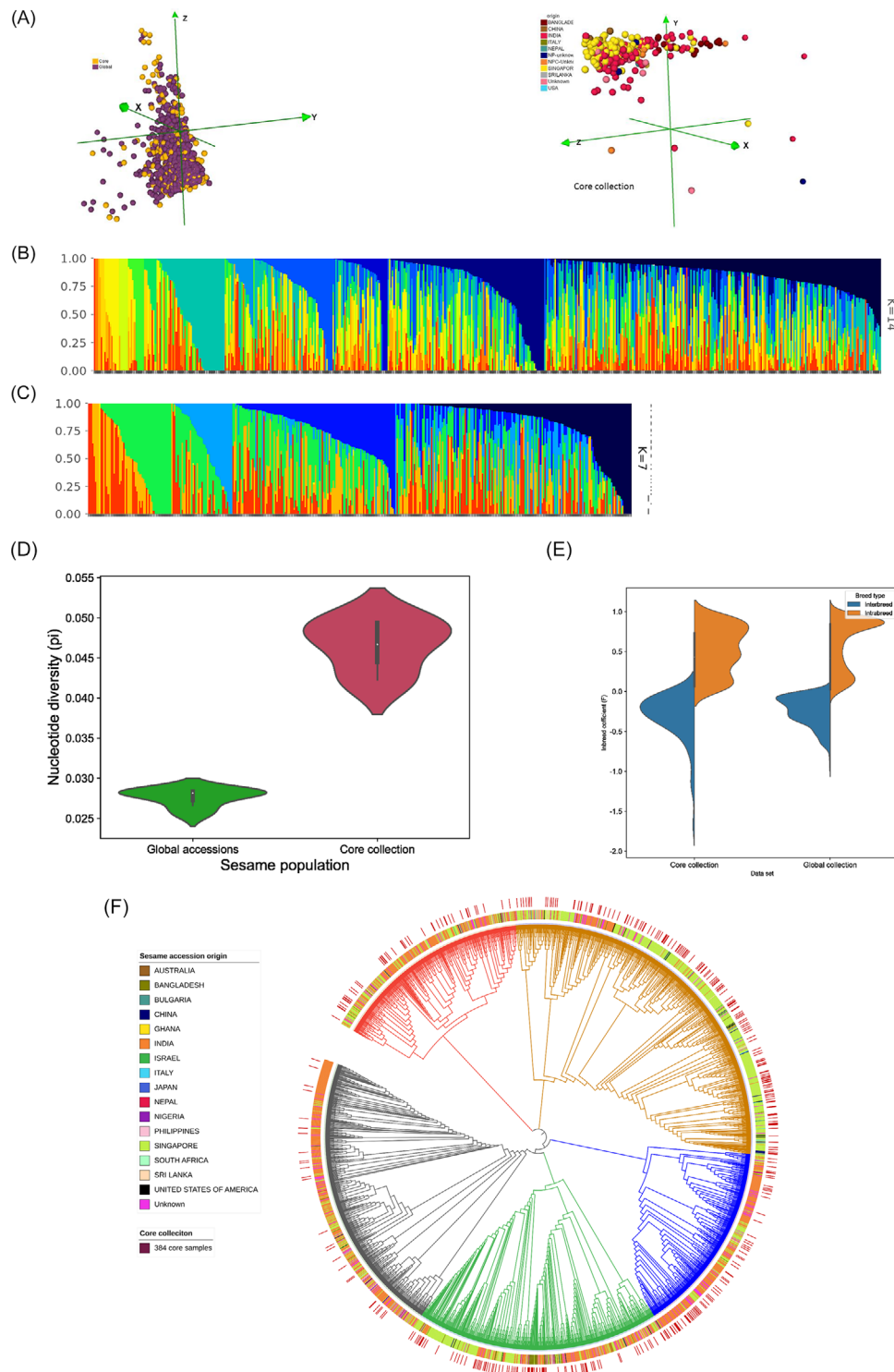
FIGURE 4 (A) A PCO analysis for global and core collection accessions. Admixture analysis of (B) global accessions and (C) core collection accessions. (D) Nucleotide diversity. (E) Heterozygosity in both global and core collection. (F) A genetic relationship of sesame accessions for the 2496 accessions and core accessions.

genome without any bias on the genic or inter-genic regions of the genome, linking with the choice of the restriction enzymes (Peterson et al., 2012; Pootakham, 2023; Pootakham et al., 2016). The SNP details for the designated SCC1 set of samples (384 accessions) for the MDP set of SNPs are tabulated in Table S9.

4 | DISCUSSION

4.1 | Designating core collection

Irrespective of the method or data type used to identify a coreset, most of the CC sizes range from 5% to 20% of the population from which they were derived but represent the maximal variability of the population as developed for rice (H. Zhang et al., 2011), *Glycine max* (L.) Merr. (soybean) (Oliveira et al., 2010), groundnut (Holbrook et al., 2000), and sesame (I. S. Bisht et al., 1998) crops. Various sampling strategies and methods have been utilized to develop CC (Franco-Duran et al., 2019). In this study, we combined the results of the most used core samples defining software, such as CoreHunter3 (de Beukelaer et al., 2018) and GenoCore (Jeong et al., 2017). Additionally, we have made use of k-mer analysis, H_O, and genetic distance to generate a stepwise CC with various filters to finally attain a set of 384 accessions as the SCC1. In the initial screening step, the CHGC reports the 26% accessions as the core component. Of these, 286 were in common with the k-mer-based diverse accessions. The other criterion, such as H_O and genetic distance between the accessions, was also measured to find the possible diverse accessions. The H_O is the observed heterozygosity with SNPs from the parental accessions and is used to measure the inbreeding coefficient.

On the other hand, the genetic distance is a measure of genetic divergence between the accessions representing the degree of differentiation and can be calculated with Euclidean distance, correlation, Manhattan, maximum, and Pearson measures. Of the five measures of genetic distance used in this study, Euclidean and Manhattan were found to be more consistent than the other three methods. With the above criterion, we report nearly 15% of the sesame accessions as the core accessions, covering 63.70% of the SNP variation. The SCC1 comprising 384 accessions is more extensive than the Mediterranean SCC, as reported by Basak et al. (2019). A core collection of 10%–30% of the total collection is likely to contain at least 70% of the variation of the total or global collection (Brown, 1989a, 1989b). An efficient non-random selection of these germplasm as a CC allows using a limited set of genotypes to represent the whole collection with minimum repetitiveness.

The inclusion of biological and technical replicates in the entire population and the visibility of only a single copy of an accession in the SCC1 indicates the unique genetic resource in the CC, validating the methodologies used to derive the SCC1. The inclusion of the biological replicate in one single instance is due to the reduced representation strategy wherein, most of the genomic regions were non-overlapping during reduced representation and hence represented as diversity, as a rare instance. For example, H69 and M32 are biological replicates and are expected to have a similar ddRAD sequence-read alignment pattern on the reference genome. However, on chromosome 1 at 4.6 Mb position, the H69 read alignments are missing, but the M32 sample reads show the alignment with the variant SNPs at two places (Figure 5). This indicates that the ddRAD sequence representations within the replicates (biological and technical replicates) have different coverage (mutually exclusive genomic regions) between the replicate samples and the so-called variant SNPs. However, probability of such occurrence is very low. As evident from our study, even if thousands of germplasm were involved, the ddRAD-seq strategy is cost-effective and efficient in genotyping through sequence-based approaches. The CC developed was proven to be useful for broadening the genetic base, target the high-yielding genotypes, commercially important traits, early flowering, adaptability (Yol & Uzun, 2012), abiotic stress in chickpea (*Cicer arietinum* L.) (Upadhyaya et al., 2013), sorghum [*Sorghum bicolor* (L.) Moench] (Upadhyaya et al., 2009), groundnut (Upadhyaya et al., 2014), and high oil content (Yol et al., 2015).

4.2 | Genetic diversity in sesame

In this study, the sesame genetic diversity captured in non-repetitive accessions exhibits greater diversity, as reflected by the nucleotide diversity values. The admixture analysis has defined the $K = 14$ in the total collections but has reduced only to half ($K = 7$) in the CC, while the collections were reduced to only 15%. This also reflects the variation percentage (variant SNPs) captured in the non-core accessions, which is (36.30%) much less compared to the core accessions (63.69%). The CC (PCoA) P1 component alone shares 64.44% of the CC, which is higher than the overall variation percent in the global population. This high genetic difference in a single component reflects the diversity in the sesame genetic variation. However, the PCoA and phylogenetic tree did not show clear separate clusters based on the sesame accession origin.

4.3 | Sesame MDP development

We have identified 2515 large fragments of ddRAD sequence-based markers (100 bp ± flanking sequences) for future molecular-based breeding applications, derived from the 2496 accessions. This was the first report of developing the MDP with the ddRAD sequence data in sesame using 2496 accessions. Numerous studies have been reported analyzing the genetic diversity in higher plants like rice (Arbelaez et al., 2019), chickpea (Deokar et al., 2014), sorghum (Magalhaes et al., 2007), maize (*Zea mays* L.) (Bukowski et al., 2018), and pearlmillet [*Pennisetum glaucum* (L.) R.Br.] (Semalaiyappan et al., 2023). In such studies, majorly the SNP from the whole genome sequence data have been widely used for
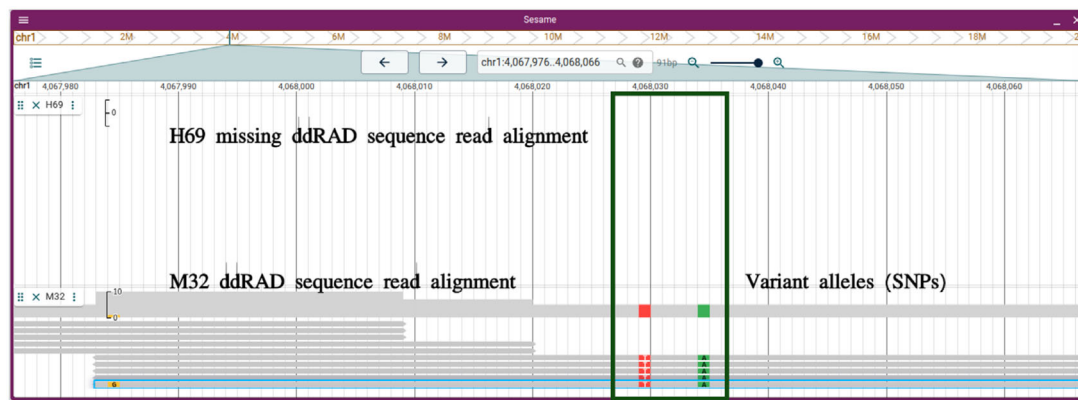
FIGURE 5 A panel showing the mapping of the restriction associated DNA (RAD) tags for the biologically replicated samples H69 and M32 wherein the RAD tags were only present in the sample M32, leading to the assumption of the variant identified in this region is unique (due to the reduced representation technique used here). ddRAD, double-digest restriction associated DNA; SNPs, single nucleotide polymorphisms.

constructing the genetic profiles and selecting the genetic background. In the reduced representation sequence used in this study, we identified 62,881 SNP, which was reduced to 2515 SNP by applying various filters as described in Section 2. Nearly half of these SNPs were tagged with the genic regions and the other half in non-genic regions, which is a perfect fit for its utilization (both foreground and background selection) as a marker-assisted selection strategy. The SNPs associated with the two genes, SIN_1005540 (*Oleosin 1-like*) and SIN_1023410 (*Dir1* [defective in induced resistance1]), known to be differentially regulated in high oil yielding genotypes when compared to the low oil yielding ones (Nawade et al., 2022), form a part of the MDP's 2515 SNPs. This underscores the potential application in screening and identification of the genotypes with SNPs carrying the alleles for high oil yield and its potential implication in sesame improvement. Appropriate choice of restriction enzymes will differentially enrich genic or non-genic regions depending on the need (Pootakham, 2023; Pootakham et al., 2016). Sooner, this will help accelerate the breeding programs toward enhanced crop improvement. When compared to the MDP from other crop plants along with its genome size, the present one with 2515 markers is sufficiently a good number with reference to its genome size. The number of markers in the MDP varies, as sorghum has 3.4K SNP markers for the MAS selection, maize (3.3K), pigeonpea [*Cajanus cajan* (L.) Millsp.] (2K), wheat (3.9K, 2.4K), rice (800), common bean (*Phaseolus vulgaris* L.) (1.8K), groundnut (2.5K), cowpea [*Vigna unguiculata* (L.) Walp.] (2.6K), and potato (*Solanum tuberosum* L.) (2.1K) (https://excellenceinbreeding.org).

5 | CONCLUSIONS

The novel and multiple strategies combining various parameters were utilized and demonstrated here in making an effective coreset using sequence-based genotyping method involving reduced representation strategy. We have demonstrated through this study on the utility of the ddRAD-seq-based genotyping strategy for various applications like coreset development and the generation of an MDP of markers. Hence, identified coreset with 384 accessions effectively reduces the number of sesame accessions when compared to the total collections but efficiently retains maximal SNP diversity and can be potentially useful in accelerated sesame crop improvement. Also, development of an MDP panel comprising of 2515 markers will be an effective aid for the MAS-based molecular breeding program. Since this set also contains the SNPs associated with genes known to be differentially regulated in high oil yielding genotypes, it has potential applications in screening the germplasm for identification of germplasm with high oil content. This panel is also useful in the genetic diversity assessment of sesame germplasm across the globe in any genebank to identify the set of most diverse accessions that can be utilized in the breeding program at lesser cost than genotyping through sequencing approaches. Thus, the modified strategy demonstrated here for the coreset identification with multiple approaches, compared to the routinely used single software-based core, is very well applicable to other crop systems as well for an efficient management and utilization of larger collection of plant genetic resources.

AUTHOR CONTRIBUTIONS
Pradeep Ruperao: Formal analysis; software; writing—original draft; writing—review and editing. **Prasad Bajaj**: Formal analysis; investigation. **Rashmi Yadav**: Funding acquisition; resources. **Mahalingam Angamuthu**: Methodology. **Rajkumar Subramani**: Resources; supervision. **Vandana Rai**: Methodology. **Kapil Tiwari**: Methodology. **Abhishek Rathore**: Resources. **Kuldeep Singh**: Funding acquisition. **Gyanendra Pratap Singh**: Funding acquisition; resources. **Ulavappa B Angadi**: Resources. **Sean Mayes**:

Resources; supervision. **Parimalan Rangan**: Conceptualization; data curation; funding acquisition; investigation; methodology; project administration; supervision; writing—review and editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interests.

DATA AVAILABILITY STATEMENT

The short-read raw data (2×150 bp) generated for 2496 samples with ddRAD strategy were submitted to the public repository bearing the project id INRP000062 or PRJEB61739 (Table S10). The complete set of filtered SNP details (64,910 SNPs) against the 2496 samples reported here is provided in a matrix form (Table S11).

ORCID

Parimalan Rangan 🄳 https://orcid.org/0000-0003-1660-8072

REFERENCES

Andrews, S. (2015). *A quality control tool for highthroughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arbelaez, J. D., Dwiyanti, M. S., Tandayu, E., Llantada, K., Jarana, A., Ignacio, J. C., Platten, J. D., Cobb, J., Rutkoski, J. E., Thomson, M. J., & Kretzschmar, T. (2019). 1k-RiCA (*1K-Rice Custom Amplicon*) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. *Rice*, *12*(1), 1–15.

Basak, M., Uzun, B., & Yol, E. (2019). Genetic diversity and population structure of the Mediterranean sesame core collection with use of genome-wide SNPs developed by double digest RAD-Seq. *PloS One*, *14*(10), e0223757.

Bedigian, D. (2003). Evolution of sesame revisited: Domestication, diversity and prospects. *Genetic Resources and Crop Evolution*, *50*, 779–787.

Bhat, V. K., Babrekar, P. P., & Lakhanpaul, S. (1999). Study of genetic diversity in Indian and exotic sesame (*Sesamum indicum* L.) germplasm using random amplified polymorphic DNA (RAPD) markers. *Euphytica*, *110*, 21–34.

Bhunia, R. K., Chakraborty, A., Kaur, R., Gayatri, T., Bhat, K. V., Basu, A., Maiti, K., & Sen, S. K. (2015). Analysis of fatty acid and lignan composition of Indian germplasm of sesame to evaluate their nutritional merits. *Journal of the American Oil Chemists' Society*, *92*(1), 65–76.

Bisht, I., Bhat, K., Lakhanpaul, S., Biswas, B., Pandiyan, M., & Hanchinal, R. (2004). Broadening the genetic base of sesame (*Sesamum indicum* L.) through germplasm enhancement. *Plant Genetic Resources Characterization and Utilization*, *2*(3), 143–151.

Bisht, I. S., Mahajan, R. K., Loknathan, T., & Agrawal, R. (1998). Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. *Genetic Resources and Crop Evolution*, *45*, 325–335.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Brown, A. (1989a). The case for core collections. In A. Brown, O. Frankel, D. Marshall, & J. T. Williams (Eds.), *The use of plant genetic resources* (pp. 136–156). Cambridge University Press.

Brown, A. (1989b). Core collections: A practical approach to genetic resources management. *Genome*, *31*(2), 818–824.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., & Xie, C. (2018). Construction of the third-generation *Zea mays* haplotype map. *Gigascience*, *7*(4), gix134.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008. 10.1093/gigascience/giab008

de Beukelaer, H., Davenport, G. F., & Fack, V. (2018). Core hunter 3: Flexible core subset selection. *BMC Bioinformatics*, *19*, 1–12.

Deokar, A. A., Ramsay, L., Sharpe, A. G., Diapari, M., Sindhu, A., Bett, K., Warkentin, T. D., & Tar'an, B. (2014). Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. *BMC Genomics*, *15*, 1–19.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048.

Franco-Duran, J., Crossa, J., Chen, J., & Hearne, S. J. (2019). The impact of sample selection strategies on genetic diversity and representativeness in germplasm bank collections. *BMC Plant Biology*, *19*(1), 1–17.

Holbrook, C. C., Timper, P., & Xue, H. (2000). Evaluation of the core collection approach for identifying resistance to *Meloidogyne arenaria* in peanut. *Crop Science*, *40*(4), 1172–1175.

Jeong, S., Kim, J.-Y., Jeong, S.-C., Kang, S.-T., Moon, J.-K., & Kim, N. (2017). GenoCore: A simple and fast algorithm for core subset selection from large genotype datasets. *PloS One*, *12*(7), e0181420.

Kim, K.-W., Chung, H.-K., Cho, G.-T., Ma, K.-H., Chandrabalan, D., Gwag, J.-G., & Park, Y.-J. (2007). PowerCore: A program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics*, *23*(16), 2155–2162.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., & Marra, M. A. (2009). Circos: An information

aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645.

Kuo, H.-I., Dai, H.-Y., Wu, Y.-P., & Tseng, Y.-C. (2021). Peanut germplasm evaluation for agronomic traits and disease resistance under a two-season cropping system in Taiwan. *Agriculture*, *11*(12), 1277.

Letunic, I., & Bork, P. (2019). Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Liu, C.-C., Shringarpure, S., Lange, K., & Novembre, J. (2020). Exploring population structure with admixture models and principal component analysis. *Methods Molecular Biology*, *2090*, 67–86.

Magalhaes, J. V., Liu, J., Guimaraes, C. T., Lana, U. G., Alves, V. M., Wang, Y.-H., Schaffert, R. E., Hoekenga, O. A., Piñeros, M. A., Shaff, J. E., Klein, P. E., Carneiro, N. P., Coelho, C. M., Trick, H. N., & Kochian, L. V. (2007). A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nature Genetics*, *39*(9), 1156–1161.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Mondal, N., Bhat, K., Srivastava, P., & Sen, S. (2016). Effects of domestication bottleneck and selection on fatty acid desaturases in Indian sesame germplasm. *Plant Genetic Resources Characterization and Utilization*, *14*(2), 81–90.

Nawade, B., Kumar, A., Maurya, R., Subramani, R., Yadav, R., Singh, K., & Rangan, P. (2022). Longer duration of active oil biosynthesis during seed development is crucial for high oil yield—Lessons from genome-wide in silico mining and RNA-Seq validation in sesame. *Plants*, *11*(21), 2980.

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294.

Oliveira, M. F., Nelson, R. L., Geraldi, I. O., Cruz, C. D., & de Toledo, J. F. F. (2010). Establishing a soybean germplasm core collection. *Field Crops Research*, *119*(2-3), 277–289.

Pathak, N., Bhaduri, A., Bhat, K., & Rai, A. (2015). Tracking sesamin synthase gene expression through seed maturity in wild and cultivated sesame species—A domestication footprint. *Plant Biology*, *17*(5), 1039–1046.

Pathak, N., Rai, A. K., Kumari, R., Thapa, A., & Bhat, K. V. (2014). Sesame crop: An underexploited oilseed holds tremendous potential for enhanced food value. *Agricultural Sciences*, *5*(6), 46023.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One*, *7*(5), e37135.

Pootakham, W. (2023). Genotyping by sequencing (GBS) for genome-wide SNP identification in plants. In Y. Shavrukov (Ed.), *Plant genotyping: Methods and protocols* (pp. 1–8). Springer Nature.

Pootakham, W., Sonthirod, C., Naktang, C., Jomchai, N., Sangsrakru, D., & Tangphatsornruang, S. (2016). Effects of methylation-sensitive enzymes on the enrichment of genic SNPs and the degree of genome complexity reduction in a two-enzyme genotyping-by-sequencing (GBS) approach: a case study in oil palm (*Elaeis guineensis*). *Molecular Breeding*, *36*, 1–7.

Rangan, P., Henry, R., Wambugu, P., & Periyannan, S. (2023). Plant genetic and genomic resources for sustained crop improvement. *Frontiers in Plant Science*, *14*, 1266698.

Rangan, P., Pradheep, K., Archak, S., Smýkal, P., & Henry, R. (2023). Genomics and phenomics of crop wild relatives (CWRs) for crop improvement. *Frontiers in Plant Science*, *14*, 1221601.

Ruperao, P., Bajaj, P., Subramani, R., Yadav, R., Reddy Lachagari, V. B., Lekkala, S. P., Rathore, A., Archak, S., Angadi, U. B., Singh, R., Singh, K., Mayes, S., & Rangan, P. (2023). A pilot-scale comparison between single and double-digest RAD markers generated using GBS strategy in sesame (*Sesamum indicum* L.). *PloS One*, *18*(6), e0286599.

Semalaiyappan, J., Selvanayagam, S., Rathore, A., Gupta, S. K., Chakraborty, A., Gujjula, K. R., Haktan, S., Viswanath, A., Malipatil, R., Shah, P., Govindaraj, M., Ignacio, J. C., Reddy, S., & Singh, A. K. (2023). Development of a new AgriSeq 4K mid-density SNP genotyping panel and its utility in pearl millet breeding. *Frontiers in Plant Science*, *13*, 1068883.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Stehno, Z., Faberová, I., Dotlačil, L., Martynov, S., & Dobrotvorskaya, T. (2006). Genealogical analysis in the Czech spring wheat collection and its use for the creation of core collection. *Czech J Genet Plant Breed*, *42*, 117–125.

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., & Davenport, G. F. (2009). Core hunter: An algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinformatics*, *10*, 1–13.

Upadhyaya, H. D., Dronavalli, N., Dwivedi, S. L., Kashiwagi, J., Krishnamurthy, L., Pande, S., Sharma, H. C., Vadez, V., Singh, S., Varshney, R. L., & Gowda, C. L. L. (2013). Mini core collection as a resource to identify new sources of variation. *Crop Science*, *53*(6), 2506–2517.

Upadhyaya, H., Dwivedi, S., Vadez, V., Hamidou, F., Singh, S., Varshney, R., & Liao, B. (2014). Multiple resistant and nutritionally dense germplasm identified from mini core collection in peanut. *Crop Science*, *54*(2), 679–693.

Upadhyaya, H., Pundir, R., Dwivedi, S., Gowda, C., Reddy, V. G., & Singh, S. (2009). Developing a mini core collection of sorghum for diversified utilization of germplasm. *Crop Science*, *49*(5), 1769–1780.

Uzun, B., Arslan, Ç., & Furat, Ş. (2008). Variation in fatty acid compositions, oil content and oil yield in a germplasm collection of sesame (*Sesamum indicum* L.). *Journal of the American Oil Chemists' Society*, *85*(12), 1135–1142.

Van Rheenen, H. (1980). Aspects of natural cross-fertilization in sesame (*Sesamum indicum* L.). *Tropical agriculture. Trinidad and Tobago*, *57*(1), 53–59.

Wang, J.-C., Jin, H., Zhang, C.-f., & Zhang, S. (2007). Assessment on evaluating parameters of rice core collections constructed by genotypic values and molecular marker information. *Rice Science*, *14*(2), 101–110.

Wang, L., Xia, Q., Zhang, Y., Zhu, X., Zhu, X., Li, D., Ni, X., Gao, Y., Xiang, H., Wei, X., Yu, J., Quan, Z., & Zhang, X. (2016). Updated sesame genome assembly and fine mapping of plant height and

seed coat color QTLs using a new high-density genetic map. *BMC Genomics*, *17*(1), e31.

Williams, J. T., & Holden, J. H. (Eds.). (1984). *Crop genetic resources: Conservation & evaluation*. George Allen & Unwin.

Xiurong, Z., Yingzhong, Z., Yong, C., Xiangyun, F., Qingyuan, G., Mingde, Z., & Hodgkin, T. (2000). Establishment of sesame germplasm core collection in China. *Genetic Resources and Crop Evolution*, *47*, 273–279.

Yadav, R., Kalia, S., Rangan, P., Pradheep, K., Rao, G. P., Kaur, V., Pandey, R., Rai, V., Vasimalla, C. C., Langyan, S., Sharma, S., Thangavel, B., Rana, V. S., Vishwakarma, H., Shah, A., Saxena, A., Kumar, A., & Singh, K. (2022). Current research trends and prospects for yield and quality improvement in sesame, an important oilseed crop. *Frontiers in Plant Science*, *13*, 863521.

Yol, E., Toker, R., Golukcu, M., & Uzun, B. (2015). Oil content and fatty acid characteristics in Mediterranean sesame core collection. *Crop Science*, *55*(5), 2177–2185.

Yol, E., & Uzun, B. (2012). Geographical patterns of sesame accessions grown under Mediterranean environmental conditions, and establishment of a core collection. *Crop Science*, *52*(5), 2206–2214.

Zhang, H., Zhang, D., Wang, M., Sun, J., Qi, Y., Li, J., Wei, X., Han, L., Qiu, Z., Tang, S., & Li, Z. (2011). A core collection and mini core collection of *Oryza sativa* L. in China. *Theoretical and Applied Genetics*, *122*, 49–61.

Zhang, P., Li, J., Li, X., Liu, X., Zhao, X., & Lu, Y. (2011). Population structure and genetic diversity in a rice core collection (*Oryza sativa* L.) investigated with SSR markers. *PLoS One*, *6*(12), e27565.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.