



OPEN

Comparative analysis of RAD-seq methods for SNP discovery and genetic diversity assessment in oil seed crop safflower

Pooja Pathania¹, Gaddam Prasanna Kumar¹, Nishu Gupta¹, R. Parimalan¹, J. Radhamani¹, Rajesh Kumar¹, Sunil Shriram Gomashe², Palchamy Kadirvel³ & S. Rajkumar¹✉

Safflower (*Carthamus tinctorius* L.) is an important oilseed crop with diverse uses and the potential for genetic improvement. This study aimed to optimize genotyping-by-sequencing (GBS) for safflower via in silico and in vitro methods with two restriction site-associated DNA sequencing (RAD-seq) approaches, i.e., single restriction site-associated DNA sequencing (sdRAD-seq) and double-digest RAD sequencing (ddRAD-seq) and three restriction enzyme combinations (ApeKI, NlaIII_MseI, and EcoRI_MseI). Forty-two safflower accessions were selected for this study. In silico testing revealed that NlaIII_MseI generated the largest number of DNA fragments, followed by ApeKI and EcoRI_MseI. The in vitro results showed that ddRAD-seq outperformed sdRAD-seq in terms of raw read count, alignment rate, depth and breadth of coverage, and SNP detection. An alignment-free analysis using k-mer counting and sketching based on genetic distance further confirmed the superiority of ddRAD-seq. Gene-level k-mer validation identified more core genes in the ddRAD-seq data. Variant calling resulted in 6,721, 173,212, and 221,805 single nucleotide polymorphic sites (SNPs) for ApeKI, NlaIII_MseI, and EcoRI_MseI, respectively. SNP annotation and distribution analysis revealed that EcoRI_MseI captured more SNPs with fewer missing observations. Principal component analysis via ddRAD-seq data explained 30.29% and 33.98% of the total genetic variation in NlaIII_MseI and EcoRI_MseI, respectively. This study demonstrated that ddRAD-seq with the EcoRI_MseI enzyme combination is the most suitable GBS approach for genome sampling and SNP genotyping in safflower.

Keywords *Carthamus tinctorius*, Genetic diversity, Genome sampling, RAD-seq, Restriction enzymes, SNP discovery

Safflower (*Carthamus tinctorius* L.), an oilseed crop from the family Asteraceae, is considered a versatile crop owing to its benefits and uses, and has significant potential for future exploitation in terms of its genetic diversity. Traditionally, safflower has been cultivated for its floral pigments (red and yellow), which are utilized in industries such as textiles, food, cosmetics, painting, and pharmaceuticals^{1,2}. Safflower is also used as livestock feed², seed oil for biodiesel production³, straw for biogas and bioethanol refineries^{4,5}, and phytoremediation¹. The commercial importance of safflower is related primarily to its fatty acid profile, which includes two unsaturated fatty acids (oleic acid and linoleic acid), accounting for approximately 90% of the total fatty acids, with the remaining 10% comprising saturated fatty acids such as palmitic acid and stearic acid⁶. Safflower is cultivated in more than 60 countries under diverse climatic conditions, with Kazakhstan, the USA, Mexico, India, Turkey, and China being the highest producers⁷. Safflower crops are also recognized for their tolerance to drought, high temperatures⁸, and salinity⁹.

DNA polymorphism resulting from restriction length fragment polymorphism (RLFP), simple sequence repeat (SSR), and single nucleotide polymorphism (SNP) forms the basis of modern genetics. The discovery and analysis of these genomic variations, such as SNPs, insertions/deletions (InDels), and structural variations (SVs), among individuals and between populations, constitute the foundation for recognizing different genotypes and linking them to phenotypes^{10,11}. Reduced representation sequencing (RRS) is one of the most cost-effective and efficient high-throughput genotyping methods¹². Current RRS methods include single restriction site-

¹ICAR- National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi 110012, India. ²ICAR- National Bureau of Plant Genetic Resources, RS-Akola, Akola, India. ³ICAR-Indian Institute of Oilseeds Research, Rajendranagar 500030, Hyderabad, India. ✉email: s.rajkumar@icar.org.in

associated DNA sequencing (sdRAD-seq)¹³, double-digest RAD sequencing (ddRAD-seq)¹⁴ and Diversity Array Technology (DART)¹⁵. Both methods reduce genome complexity via restriction enzymes (REs). The utilization of different single and paired restriction enzymes in various combinations renders these approaches more suitable for addressing diverse types of differences in the genomes of different crops. The selection of restriction enzymes is also crucial because it can affect genome coverage¹² and SNP genotyping¹⁶. Compared with previous fragment size analysis methods or sequence comparisons of individual nuclear or chloroplast DNA regions, the genotyping-by-sequencing (GBS) approach has the potential to provide more precise relationships both within and among closely related species¹⁷.

The process of building GBS libraries involves restriction enzymes (REs) to reduce the complexity of the genome¹⁸, as well as a barcoding system to multiplex template DNA to produce representative libraries¹³. This genomic reduction, achieved via REs, represents a significant technical challenge because the genomic distribution of SNPs depends on the precise combination of the REs employed¹⁹. Selecting appropriate REs is crucial for species-specific GBS optimization. By utilizing suitable REs, repetitive regions of the genome can be avoided, and regions with fewer copies can be targeted with two to three times greater efficiency for accessing important regions of the genome²⁰. When selecting optimal REs for GBS optimization, factors such as genome size, the methylation of repeat elements, and genome organization are considered²¹. In plants, the REs ApeKI, PstI, and EcoRI are generally utilized and combined with frequent cutter enzymes such as MspI, MseI, and HpaII, depending on the genomic specificity of each species²².

While various marker systems have proven effective in assessing genetic diversity in safflower, there is insufficient information to directly compare SNP markers to other systems for this crop. However, given the success of SNP markers in other species and their generally high informativeness, it is likely that they would also be valuable for safflower genetic diversity studies. SNP markers have shown substantial genetic variation and have been useful in identifying population structure and genetic diversity²³. In the present study, two RAD-seq approaches (sdRAD-seq and ddRAD-seq) with three restriction enzyme combinations (ApeKI, NlaIII_MseI and EcoRI_MseI) were evaluated for GBS optimization via in silico and in vitro methods. The data were analysed utilizing both alignment-free and alignment-based approaches. The comparison between single-digest RAD-seq (sdRAD-seq) and double-digest RAD-seq (ddRAD-seq) was chosen to evaluate their relative performance and efficiency in genomic analysis. These methods were selected because they are widely used reduced representation sequencing techniques that allow for cost-effective genome-wide genotyping²⁴. The main difference between these approaches lies in the number of restriction enzymes used: sdRAD-seq uses a single enzyme, while ddRAD-seq employs two enzymes. This distinction can significantly impact the resulting genomic coverage, number of loci sampled, and overall sequencing efficiency²⁵. By comparing these methods, researchers can assess their relative strengths and limitations in various applications, such as SNP discovery, genetic mapping, and population genomics studies. Interestingly, while both methods have their merits, ddRAD-seq has gained popularity due to its flexibility and potential for more balanced genomic sampling^{24,25}. However, the choice between sdRAD-seq and ddRAD-seq often depends on specific research goals, target species, and available resources. By directly comparing these methods, researchers can make informed decisions about which approach is most suitable for their particular study, considering factors such as genome size, desired marker density, and budget constraints.

Materials and methods

Materials

A total of 42 diverse safflower accessions were selected for the study (Supplementary Table S1) and were obtained from the National Gene Bank of the Indian Council of Agriculture Research, National Bureau of Plant Genetic Resources, New Delhi (ICAR-NBPGR). Of these, 19 accessions originated from different countries, whereas the remaining 23 accessions originated from various states of India. All the accessions were purified by selfing (single-seed descent method) for two generations, at the Akola, Regional Station of the NBPGR. Seeds were soaked overnight in water and germinated on towel paper for 5–7 days. The seedlings were subsequently subjected to DNA extraction using the DNeasy Plant Kit (Qiagen, USA). The quality and quantity of the extracted DNA were assessed by electrophoresis.

Selection of restriction enzymes (REs)

Three pairs of enzymes were selected to evaluate the efficacy of the sdRADseq and ddRADseq methods. For sdRADseq, the ApeKI restriction enzyme was utilized, whereas, for ddRADseq, a combination of a rare and frequent cutter, specifically the EcoRI_MseI and NlaIII_MseI pair of restriction enzymes, was employed (Table 1). In vivo studies on number of restriction site mapping in the reference genome provided EcoRI and NlaIII were having higher number of sites along with MseI and also commonly used enzymes in RAD-seq due to their distinct recognition sites, whereas MseI was paired with later two enzymes for its compatibility²².

Enzyme	Recognition site	Length	Methylation sensitivity
ApeKI	G CWGC	5	CpG
EcoRI	G AATTC	6	CpG
NlaIII	CATG	4	No
MseI	T TAA	4	No

Table 1. Summary of the different restriction enzymes used in this study.

In Silico simulation

In silico analyses of sdRADseq and ddRADseq methodologies using ApeKI, EcoRI_Msel, and NlaIII_Msel restriction enzymes were conducted via the R package simRAD²⁶. These analyses were performed using safflower reference genome version 1 obtained from The Genome database of *Carthamus tinctorius*²⁷.

In vitro testing (sdRADseq and DdRADseq protocol)

In vitro testing was conducted on 42 diverse accessions common to both methods (sdRADseq and ddRADseq) via restriction enzymes (ApeKI, NlaIII_Msel, and EcoRI_Msel). The sdRAD-seq and ddRADseq protocols were similar, differing only in the digestion step with respect to the restriction enzymes used in this study. DNA samples (200 ng of DNA/sample) were digested with ApeKI for sdRAD-seq and double digested using rare-cutting NlaIII/EcoRI and frequent-cutting Msel enzymes (New England BioLabs, US) for ddRADseq. The digested DNA fragments were subsequently ligated with P1 and P2 adapters via T4 ligase (New England BioLabs, US). The ligation reaction involved overnight incubation (> 12 h) at room temperature (approximately 21 °C) followed by heat deactivation of the enzyme at 65 °C for 10 min. To eliminate unincorporated adapters and small DNA fragments (< 300 bp), the ligation products were purified using 0.8X volume of Agencourt AMPure XP SPRI magnetic beads (Beckman Coulter, US). A unique combination of dual-indexed barcodes was attached to the purified fragments through 14 PCR cycles. Indexed PCR products were pooled in equal volumes, and fragments between 300 and 700 bp in size were selected using Agencourt AMPure XP SPRI magnetic beads. The library concentration was determined using a Qubit.3 fluorometer (Life Technologies, US) with a Qubit dsDNA HS (High Sensitivity) Assay Kit (Thermo Fisher Scientific, US). Library quality assessment was performed using an Agilent D5000 ScreenTape System (Agilent, US) in a 4150 TapeStation System (Agilent, US). The library qualification criteria included the presence of a broad peak in the range of 300 bp to 1000 bp, with an average size of 400 bp in the Agilent 4150 TapeStation system, and Qubit concentrations above 2 ng/μl or 10 nmol. Final pooling and sequencing were conducted on a NovaSeq 6000 platform.

Sequencing and data processing

Sequence data were generated via NovaSeq 6000. Data quality was assessed using the FastQC software²⁸. The raw reads were trimmed using trimmomatic²⁹. For alignment-based analysis, the cleaned reads were mapped to safflower reference genome version 1²⁴ using the bwa mem algorithm³⁰. For each sample, the breadth and depth of coverage were determined using the SAM tool package³¹. For alignment-free analysis, a k-mer analysis was conducted. Cleaned reads were subjected to k-mer counting with a k-mer size of 27 nt, and distinct and unique k-mers were identified using jellyfish³². The k-mers that appeared only once in the samples were filtered out, as they were likely attributable to sequencing errors. The k-mer-based genetic distance between the 42 samples was measured using Mash³³. Gene-level k-mer sequence validation was performed by gene presence/absence variation (PAV) in each sample using SGSGeneLoss³⁴. Variants from both the sdRAD-seq and ddRAD-seq datasets were called using the bayesian genotype caller and stacks software using “ref_map.pl” pipeline with the remove duplicates option. The “population” pipeline of stacks software³⁵ was further utilized to export the detected SNPs in VCF format. Vcftools³⁶ was employed to filter the variants, which were subsequently plotted via CMplot³⁷. Variant effects were predicted using SnpEff software³⁸ and gene annotation of the safflower genome version 1²⁷. Principal component analysis (PCA) was performed and plotted using the R package SNPRelate³⁹.

Results

In Silico simulation

Enzymatic digestion of the reference safflower genome, which was simulated with the four different restriction enzymes, revealed that the Msel enzyme has the highest number of restriction sites in the safflower genome, followed by NlaIII, ApeKI, and EcoRI (Supplementary Fig. S1). In silico digestion showed that the highest total number of DNA fragments was generated by the NlaIII_Msel combination, followed by ApeKI and EcoRI_Msel. After the size selection range of 300 to 700 bp, similar patterns were observed, where NlaIII_Msel resulted in the highest number of fragments, and the lowest number for EcoRI_Msel. A comparison of the fragment size distributions revealed that after size selection, only 5% of the fragments were retained for NlaIII_Msel, whereas for ApeKI and EcoRI_Msel, 12% of total fragments were retained (Supplementary Fig. S2).

In vitro results

Sequencing data

sdRAD-seq by ApeKI and ddRAD-seq by NlaIII_Msel and EcoRI_Msel resulted in total raw read counts of 175,46, 424.77 and 332.46 million respectively for the set of 42 diverse safflower accessions. sdRAD-seq by ApeKI resulted in a minimum of 48,978 (Y15) reads and a maximum of 1, 29, 74, and 484 (Y95) reads, with an average of 4.17 million reads per sample. NlaIII_Msel produced 5,664 (L17) lowest and 3,77,84,410 (G54) the highest number of reads, with an average of 10.11 million reads per sample. The EcoRI_Msel combination reported a minimum of 78,854 (X8) and a maximum of 1,91,48,138 (Y16) read counts, with an average of 7.92 million reads per sample (Supplementary Fig. S3).

Alignment based analysis

For alignment-based analysis, raw reads were trimmed for adapter content, and low-quality reads were aligned to the safflower reference genome. The alignment rate for sdRAD-seq was not uniform among the safflower accessions, varying from 14.21% (N90) to 92.99% (I72). However, in the case of ddRAD-seq, an alignment rate of > 82% was reported. The alignment rates for the NlaIII_Msel and EcoRI_Msel restriction enzyme combination were highly related, and EcoRI_Msel showed a nearly uniform alignment rate for 42 samples (Supplementary Fig.

S4). Among the aligned bam files, 46.78% of the ApeKI reads were primary aligned; however, the percentages of the primary aligned NlaIII_MseI and EcoRI_MseI reads were 96.68% and 98.91%, respectively (Supplementary Table S2). Alignment scores for ApeKI, NlaIII_MseI and EcoRI_MseI are 45.56, 106.20 and 119.43 respectively.

Depth and breadth of coverage

The depth of coverage for each sample was determined using the aligned bam files. The per-sample depth of coverage was lowest for ApeKI, that is, 0.24X with a range from 0.0004X (A82) to 0.6596X (A15). The NlaIII_MseI depth of coverage varied from 0.0005X to 3.9485X, with an average of 1.10X per sample. In the case of EcoRI_MseI, the depth coverage ranged from 0.0091X (X8) to 1.781X (Y16), with a per-sample average of 0.785X (Supplementary Fig. S5 a). The mean breadth coverage for ApeKI was 0.68%, and NlaIII_MseI and EcoRI_MseI had 24.57% and 7.05% coverage breadth, respectively (Supplementary Fig. S5 b).

Alignment free analysis

K-mer (Distinct & Unique) analysis

K-mer frequencies in 42 samples with sdRAD-seq using ApeKI and ddRAD-seq using NlaIII_MseI and EcoRI_MseI were recorded on the basis of distinct and unique K-mer counts. sdRAD-seq by ApeKI resulted distinct k-mers counts ranging from 0.06 million (A15) to 59.91 million (Z13) with an average of 25.92 million per sample (Supplementary Fig. 6a). ddRAD-seq by NlaIII_MseI reported average distinct counts of 172.20 million per sample with a range of 0.34 million (L17) to 397.22 (J77) million k-mers (Supplementary Fig. 6b) whereas for EcoRI_MseI, distinct k-mer frequency varies from 2.39 million (X8) to 141.22 million (Y16), with an average of 76.20 million per sample (Supplementary Fig. 6c). Unique K-mers comparison shown that sdRAD-seq by ApeKI generated more common K-mers (88.22%) than the unique K-mers (11.78%) for 42 safflower samples. However, ddRAD-seq by NlaIII_MseI captured 46.40% unique K-mers and 58.12% EcoRI_MseI.

K-mer sketching based on genetic distance

The k-mer-based genetic distance between each pair of samples was calculated and sketched to create a reduced representation while unique or low-copy k-mers were removed (minimum copies of each k-mer to filter unique k-mers were specified as 2), since these k-mers are likely be the result of sequencing errors. Genetic distance measures variation between accessions. For the ApeKI k-mer-based genetic distance between each pair of samples, A15, A82, C54, B91, and K95 presented high genetic distances (Fig. 1a). L17 and F18 showed high genetic distance for NlaIII_MseI (Fig. 1b) and EcoRI_MseI (X8 and A13), whereas samples with a low number of k-mers reflected high genetic distance (Fig. 1c). In addition to the aforementioned samples, the maximum variability in ApeKI was 0.19%. For NlaIII_MseI, apart from the L17 and F18 samples, the maximum variability reported was 0.19%. Similarly, for EcoRI_MseI, except for samples X8 and A13, the maximum variability reported was 0.13%.

Gene level validation (distribution and annotation) of k-mers

The sdRAD-seq (ApeKI) and ddRAD-seq (NlaIII_MseI and EcoRI_MseI) k-mer sequence data were assessed for gene variability in 42 safflower samples, where genes covered in all samples by k-mer data were considered core genes and genes absent in any of the 42 safflower samples were considered variable genes. sdRAD-seq by ApeKI identified 16 core and 17,393 variable genes. Sixteen core genes were able to cover chromosomes 1,2,3,6,10 and chromosome 12, with 3, 4,1,1,5 and 2 genes, respectively. ddRAD-seq via NlaIII_MseI restriction enzyme

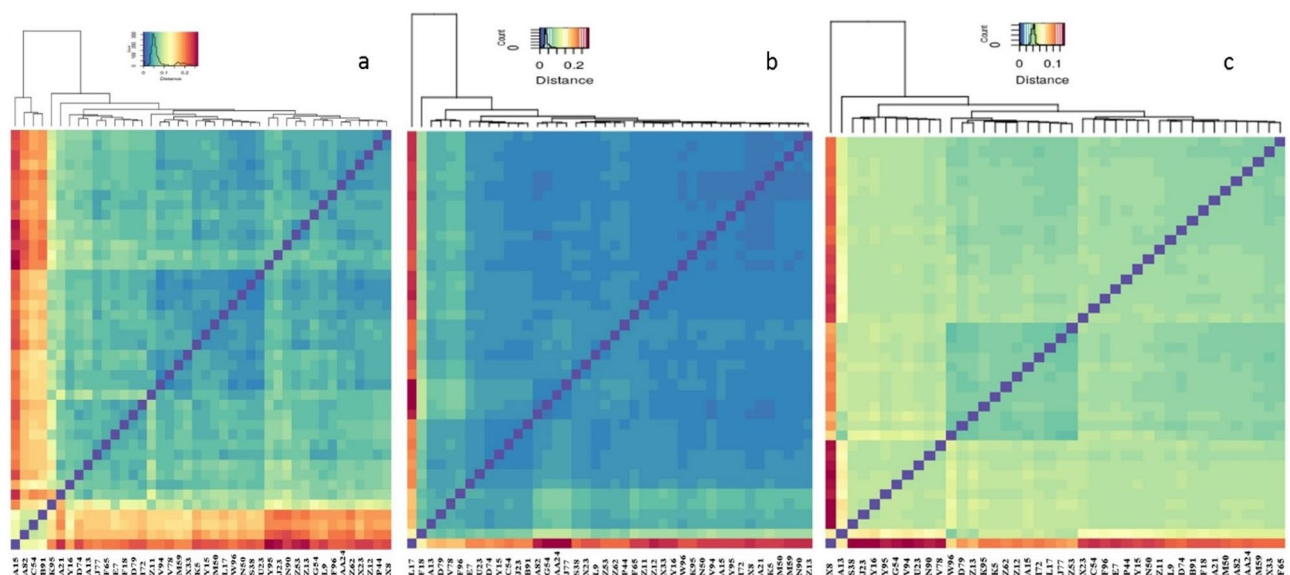


Fig. 1. Mash based genetic distance, (a) ApeKI, (b) NlaIII_MseI and (c) EcoRI_MseI.

Statistic	ApeKI	NlaIII_Msel	EcoRI_Msel
Total SNPs	6,721	1,73,212	2,21,805
Total aligned reads in bam files	10,35,40,252	47,18,49,895	33,83,30,252
Mean aligned reads/SNP	15,405	2,724	1,525
Mean aligned reads/SNP/sample	366	65	36
SNP with MAF 5% and genotype coverage 80%	289	8,292	37,226
SNP with MAF 5% and genotype coverage 100%	-	1	306

Table 2. SNP statistic for ApeKI and two restriction enzyme combinations (NlaIII_Msel and EcoRI_Msel).

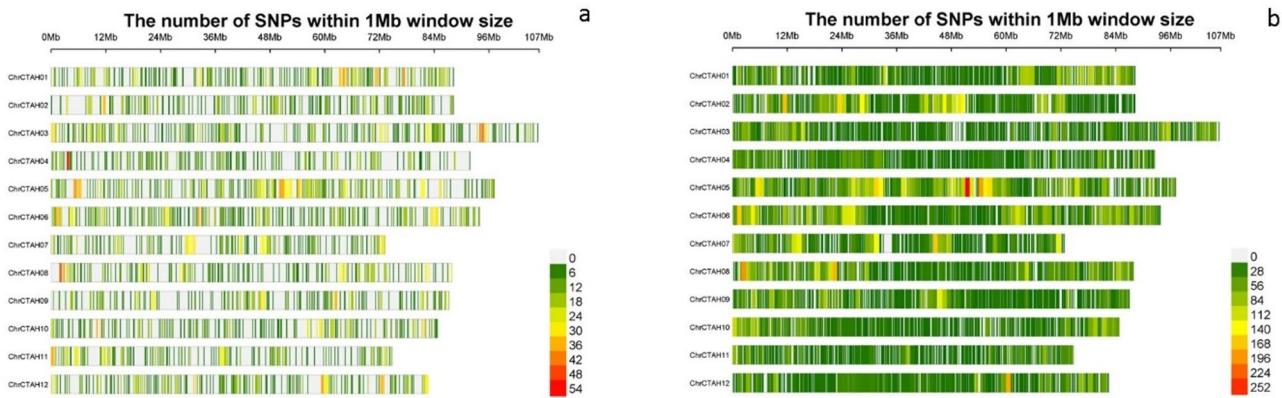


Fig. 2. Density of SNPs along chromosomes, (a) NlaIII_Msel and (b) EcoRI_Msel.

combination was able to capture 258 core genes and 32,998 variable genes. With 19, 45, 23, 25, 13, 26, 15, 15, 6, 31, 18, and 22 genes corresponding to chromosomes 1 through 12, 258 core genes were able to cover all the chromosomes. ddRAD-seq via EcoRI_Msel described 946 core and 30,217 variable genes in the set of 42 samples. All 12 chromosomes were covered by 946 core genes, which included 72, 93, 89, 83, 66, 105, 60, 66, 67, 105, 68 and 72 genes for 12 chromosomes (Supplementary Fig. S7 a). The gene ontology of sdRAD-seq (ApeKI) core genes involved in four molecular functions and two biological processes associated with three cellular components and were mainly related to photosynthesis (photosystem II reaction center protein A and B and photosynthetic electron transfer A) and ATP metabolic processes (proton transmembrane transporter activity) (Supplementary Fig. S7 b). The core genes of ddRAD-seq by NlaIII_Msel engaged in 41 molecular functions and 30 biological processes and were found in 16 cellular components (Supplementary Fig. S7 c), mainly from biological processes such as photosynthesis, transcription, translation, and DNA integration, whereas core genes from ddRAD-seq through EcoRI_Msel had 66 molecular functions, 49 biological processes, and 23 cellular components and were similarly involved in biological processes such as photosynthesis, transcription, translation, DNA integration, and oxidation reduction processes (Supplementary Fig. S7 d).

Variant calling

Variant calling from the aligned reads for sdRAD-seq (ApeKI) and ddRAD-seq (NlaIII_Msel and EcoRI_Msel) resulted in 6,721, 173,212, and 221,805 SNPs, respectively. Subsequently, SNPs were filtered using minor allele frequency (MAF) 5% and genotype coverage of 80%, which retained 289, 8,292, and 37,226 SNPs from the sdRAD-seq (ApeKI) and ddRAD-seq (NlaIII_Msel and EcoRI_Msel) datasets, respectively (Table 2). SNP statistics based on total aligned reads from the sequenced data indicated that ddRAD-seq generated more SNPs with fewer reads per SNP per sample, and the enzyme combination EcoRI_Msel exhibited superior performance compared with NlaIII_Msel. SNP analysis to assess the extent of missing data in SNP genotyping revealed that ddRAD-seq performed better than sdRAD-seq. EcoRI_Msel restriction enzyme combination produced more SNPs i.e. approximately 28% higher than NlaIII_Msel with fewer missing observations (Table 2), reflecting its genotype coverage potential. A higher number of SNP and larger sample coverage will give better representation of populations genetic diversity.

Polymorphic information content (PIC), SNP distribution & annotation for ddRAD-seq data

SNP statistics revealed that ddRAD-seq generated a higher number of SNPs, PIC analysis and annotation was performed exclusively for the ddRAD-seq data. In ddRAD-seq, NlaIII_Msel and EcoRI_Msel produced 8,292 and 37,226 SNPs, respectively. In both the enzymes combinations, the average PIC content was 0.30, with a range from 0.09 to 0.38. For NlaIII_Msel, the variant rate was 1 variant per 127,486 bases, with the highest number of SNPs on chromosome 5 (986 SNPs) and the lowest number of SNPs observed on chromosome 11 (536 SNPs) (Fig. 2a, Supplementary Table S3). The majority of the SNPs (73.94%) were located in intergenic regions, of which

13.015% were upstream and 14.441% were found in the downstream region of an open reading frame (ORF). A limited percentage of SNPs were observed in the intronic (16.121%) and exon regions (7.47%) compared with the 3'UTR (1.074%) and 5'UTR (0.936%) (Supplementary Fig. 8S a). Based on nucleotide substitutions (transitions (Ts) and transversions (Tv)), 5,589 transitions and 2,703 transversions were detected with a transition/transversion (Ts/Tv) ratio of 2.06. The transition frequency was highest for G/A, and the transversion frequency was highest for A/T (Supplementary Fig. 8S b). With respect to EcoRI_MseI, the variant rate was one variant per 28,397 bases, with chromosome 5 having the largest number of SNPs (5,515 SNPs) and chromosome 11 having the lowest number of SNPs (1,717 SNPs) (Fig. 2b, Supplementary Table S2). The highest concentration of SNPs was positioned in intergenic regions, comprising 67.15%; of these, 13.107% were upstream and 15.282% were downstream of an open reading frame (ORF). A moderate proportion of SNPs was observed when the intronic (20.485%) and exon regions (9.237%) were compared with the 3'UTR (1.262%) and 5'UTR (1.315%) (Supplementary Fig. S8 a). A total of 23,496 transitions and 13,730 transversions were identified on the basis of nucleotide substitutions (transitions (Ts) and transversions (Tv)). The ratio of transitions to transversions (Ts/Tv) is 1.71. G/A presented the highest transition frequency, whereas C/A presented the highest transversion frequency (Supplementary Fig. S8 b). Nevertheless, similarity was observed in the chromosome distribution pattern of SNPs in both enzyme combinations of ddRAD-seq and in the investigation of the structural and functional relevance of SNPs, which also followed a comparable trend.

Genetic diversity and relationships

The genetic parameters (observed and expected heterozygosity, homozygosity, and gene diversity) were investigated for 8,292 and 37,226 SNP markers recorded by NlaIII_MseI and EcoRI_MseI, respectively, for the 42 safflower accessions. The SNP data for NlaIII_MseI revealed an average gene diversity of 0.38, ranging from 0.38 to 0.39, with the results indicating that the highest gene diversity (0.39) was observed for accessions (B91, F96, X23, and A13) from India. Additionally, the observed heterozygosity (H_o) ranged from 0 to 0.06, with an average of 0.04. With EcoRI_MseI, an average gene diversity of 0.29 was recorded, with a range from 0.29 to 0.30. The highest gene diversity (0.30) was observed for 12 accessions, of which five (C54, D79, F18, L9, and N90) originated from India, four (D74, K95, P44, and S38) from the USA, and one each from Italy (K5) and Afghanistan (Z13). The observed heterozygosity (H_o) in this case ranged from 0.01 to 0.04, with an average of 0.02. For both restriction enzyme combinations for ddRAD, the expected heterozygosity (H_e) was higher than the observed heterozygosity (H_o) (Table S4). The principal component analysis (PCA) for ddRAD-seq indicated that in NlaIII_MseI, the first principal component (PC) accounted for 19.70% of the total variation (Fig. 3a), whereas in EcoRI_MseI, the PC1 accounted for 20.54% of the total variation (Fig. 3b). The cumulative variation in the first three principal components for NlaIII_MseI and EcoRI_MseI was 30.29% and 33.98%, respectively. Phylogenetic analysis utilizing the Neighbor-Joining method revealed the clustering of 42 accessions into three primary groups in both cases. For NlaIII_MseI, the first two clusters each comprised a single accession (X23 and G54), whereas the remaining 40 accessions were grouped into a single cluster. Conversely, for EcoRI_MseI, the first cluster contained two accessions (N90, F96), the second cluster included five accessions (J77, G54, W96, M59, A21), and the remaining 35 accessions formed a separate, larger cluster. (Fig. 4).

Discussion

Genotyping-by-sequencing (GBS) approaches are designed to sequence genomes to explore genetic variations and perform genome-wide association mapping¹⁶. The reduced representation methods of GBS offer the advantages of being expeditious, efficient, and economical for genome-wide genetic variation analysis and

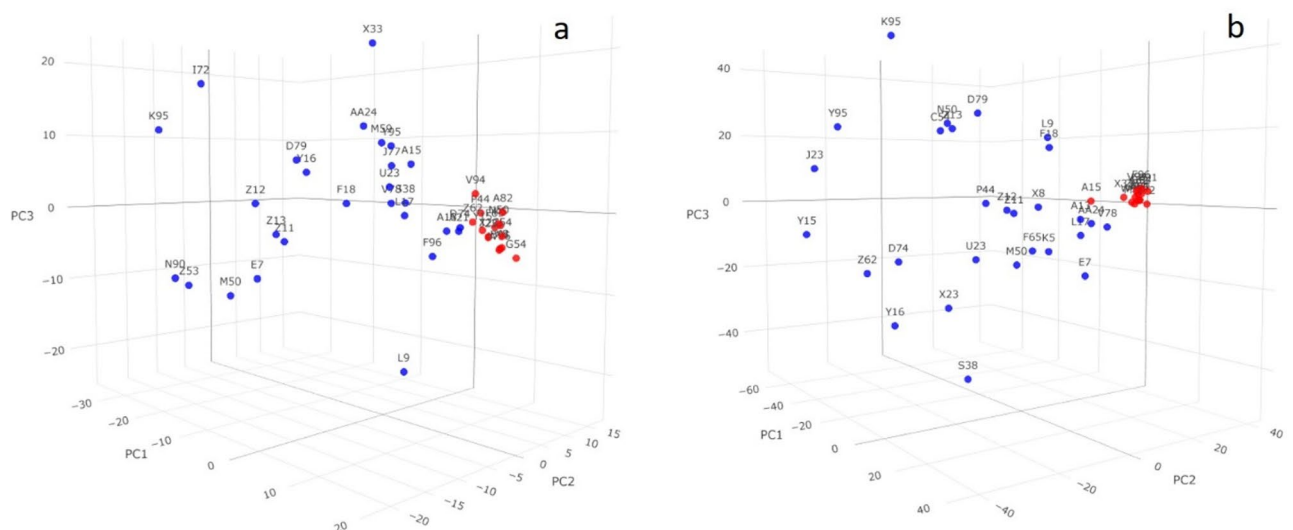


Fig. 3. Scatter plot of three principal components (PCs) for ddRAD, (a) NlaIII_MseI, (b) EcoRI_MseI. In both enzyme combinations, accessions that are clustered together are highlighted as red dots.

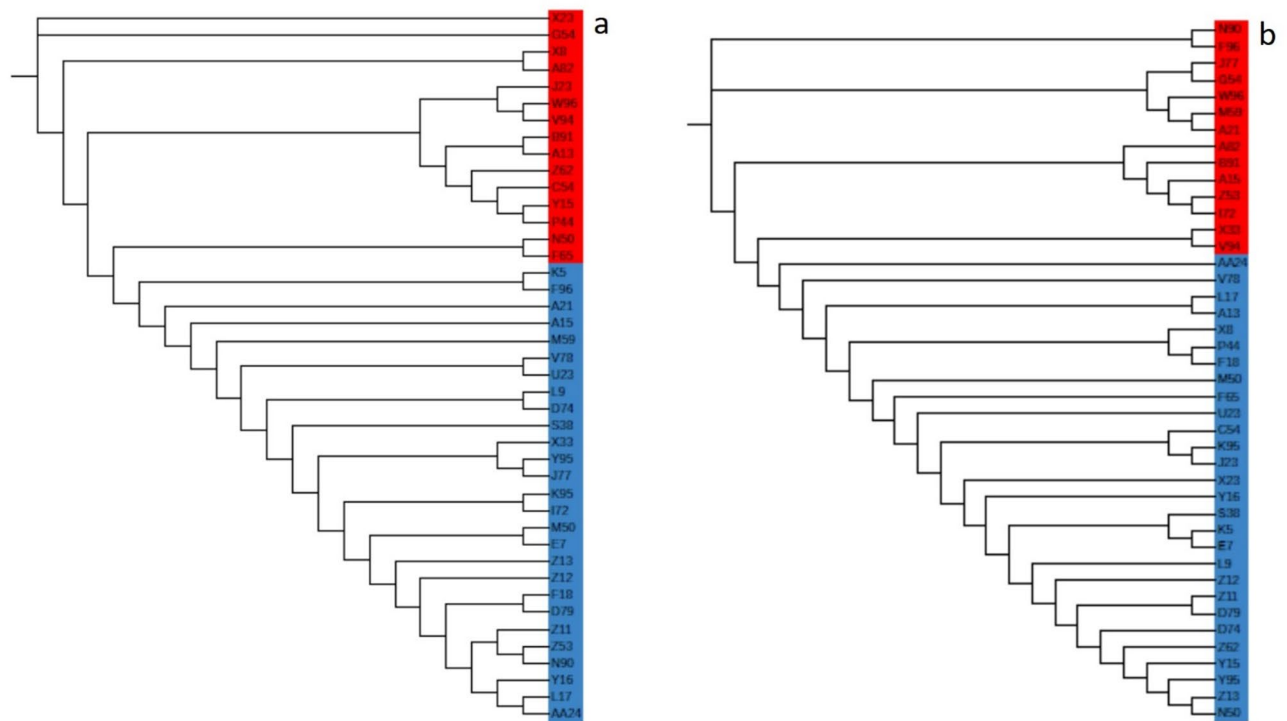


Fig. 4. Neighbor-Joining clustering for 42 safflower accessions, **a)** NlaIII_MseI, **(b)** EcoRI_MseI. Accessions that are clustered together in the PCA plots (highlighted in red) are also marked in red here, demonstrating consistent groupings across these two analyses.

association studies^{13,40} via restriction enzyme (RE)-assisted genome reduction^{41,42} and SNP calls with or without a sequenced genome^{18,43}. Selecting the appropriate restriction enzyme is a crucial initial step in the development of a GBS protocol⁴⁴. The high-quality genome assembly of safflower reported²⁷ could serve as a valuable reference for such ddRAD studies, enabling more precise genetic analyses and potentially uncovering novel aspects of safflower diversity and evolution. This suggests that high-throughput sequencing approaches like ddRAD could potentially provide even more comprehensive insights into safflower genetics. Its application could potentially overcome limitations of previous marker systems and provide more detailed insights into safflower genetics.

The simulated prediction of safflower genome complexity reduction utilized a combination of rare cutter enzymes (NlaIII and EcoRI) and frequent cutter enzymes (MseI), and a comparison of the results with a single restriction reaction (ApeKI), revealed a higher number of fragments with NlaIII_MseI ranging between 300 and 700 bp. However, different restriction enzyme combinations have been employed in previous studies, depending on the experimental requirements and analyses. For instance, GBS-based phylogenetic analyses of *C. tinctorius* together with wild species of *Carthamus*, that is, *C. boissieri*, *C. glaucus*, *C. lanatus*, *C. oxyacantha*, *C. palaestinus*, and *C. tenuis*, utilized the PstI_MspI restriction enzyme combination and obtained 1,582 loci for a dataset of 60 individuals¹⁷. To study the diversity, population structure, and marker-trait associations of 94 safflower accessions from the United States Department of Agriculture (USDA) via DArTseq, the PstI-MseI restriction enzyme combination was utilized⁴⁵. In our, in vitro experiments, the overall sequence read representation was lower in the sdRAD-seq (ApeKI) data than in the ddRAD-seq data. The NlaIII_MseI enzyme combination obtained the highest average number of raw reads per sample (> 10 million), followed by EcoRI_MseI, with more than 7 million raw reads per sample. ddRAD-seq demonstrated superior performance compared with sdRAD-seq under actual experimental conditions. ApeKI (G/CWGC), used for sdRAD-seq, is a GC-rich sequence recognition enzyme that cannot cleave AT-rich DNA regions. Additionally, ApeKI exhibits partial methylation sensitivity, which affects its digestion under wet laboratory conditions, resulting in outcomes that differ from those of in silico analysis. Similar results were observed in soybean, where ApeKI produced fewer fragments in actual experiments than in silico simulations did⁴⁴. The frequent cutter MseI (T/TAA) used for ddRAD-seq recognizes AT-rich sequences and is not sensitive to DNA cytosine methylation (Dcm), deoxyadenosine methylation (Dam), or CpG methylation. Therefore, it is not influenced by DNA methylation⁴⁶.

Based on the alignment percentage, the performance of the tested restriction enzymes demonstrated that the overall performance of EcoRI_MseI exhibited a similar trend across 42 samples, whereas for NlaIII_MseI and ApeKI, the alignment percentage varied across a broad scale. For alignment-based analysis, measuring the depth of sequencing coverage is critical for genomic analyses. NlaIII_MseI reported a higher depth of coverage because the coverage parameter was directly influenced by the read number, indicating that a higher read number corresponds to a higher depth of coverage⁴⁷. To expedite genotyping and minimize reference bias, an alignment-free algorithm was employed⁴⁸ for genome assembly and comparison. The most widely used alignment-free

method is based on the k-mer count, where sequence information is transformed into numerical values such as the k-mer count, and their frequencies are further utilized to calculate distances between samples⁴⁹. Our results indicate that ddRAD-seq performed better than sdRAD-seq in producing both distinct and unique k-mers. Gene-level validation of the k-mer data revealed that ddRAD-seq by EcoRI_MseI was able to cover more genes, that is, 946 genes in 42 samples, whereas ddRAD-seq by NlaIII_MseI covered only 258 genes in all the samples.

The improvement in SNP genotyping through restriction enzyme combinations can be assessed by SNP analysis with respect to the extent of missing data. In our study, NlaIII_MseI covered a smaller proportion of the genome than EcoRI_MseI (Fig. 2). SNP statistics revealed that EcoRI_MseI displayed more SNPs, with 80% genotype coverage and no missing observations. The H_e represent gene diversity and H_o represents observed heterozygosity. This helps in understanding the population structure and genetic diversity of population. In safflower using the enzyme combination we could be able to derive higher gene diversity (H_e) and lower heterozygosity (H_o).

The first three principal components in the PCA explained 30.29% and 33.98% of the total genetic variation for samples sequenced by ddRAD-Seq using the NlaIII_MseI and EcoRI_MseI restriction enzymes, respectively. PCA results from NlaIII_MseI showed a slightly less distinct separation of accessions, as they were tightly grouped, whereas the PCA findings from EcoRI_MseI could be more readily utilized to identify duplicates in samples. Principal Coordinates Analysis (PCoA) and phylogenetic analysis for both enzyme combinations exhibit a strong correlation. For NlaIII_MseI, accessions highlighted in red on the phylogenetic tree were consistently clustered together in the PCoA plot. These clumped accessions represent samples from diverse geographical regions, including India, Ethiopia, the USA, Germany, and Australia. In the case of EcoRI_MseI, the accessions marked in red on the phylogenetic tree also appeared to cluster together in the PCoA analysis, with these accessions being exclusively from India and the USA. Notably, EcoRI_MseI demonstrated a greater ability to distinguish accessions based on their geographical origins, further emphasizing its utility in differentiating safflower accessions from distinct regions. These findings demonstrate that among the two different approaches of reduced representation, the ddRADseq method generated significantly more variants among the accessions compared to sdRADseq. The enzyme combination EcoRI_MseI generated variants with 5% MAF more effectively than NlaIII_MseI, demonstrating the potential of ddRAD with the restriction enzyme combination EcoRI_MseI for advanced genome sampling and SNP genotyping in safflower for GBS applications.

Overall, Double-digest RAD sequencing (ddRAD-seq) offers several advantages over single-digest RAD sequencing (sdRAD-seq) for plant genotyping. ddRAD-seq utilizes two restriction enzymes, providing more flexibility and balance in genomic loci analysis. This approach allows for the examination of hundreds of individuals, making it particularly suitable for large-scale genetic and genomic studies in plants. ddRAD-seq technique that involves the production of multiplexed libraries by enzymatically digesting whole genomic DNA and then binding specific adapters, resulting in reduced representation libraries. By restricting the portion of the genome sequenced, ddRAD-seq generates a comprehensive set of SNP markers that can be used to precisely assess genetic diversity and population structure in large germplasm collections. The use of double restriction digestion also facilitates paired-end sequencing of identical loci across multiple samples, enhancing read mapping accuracy compared to sdRAD-seq.

While ddRAD-seq has been successfully applied to various diverse crops, such as tomato⁵⁰ and sesame²³, its use in safflower germplasm investigations has been relatively scarce. The present study will provide insight in to genotyping of germplasm collection conserved at Indian National Gene Bank for developing core collection harbouring maximum genetic diversity to be used as parental lines for varietal development program. The robust cost effective ddRAD markers with suitable enzyme combinations will provides maximum SNPs with wide genomic coverage for precise development of core collection. This method is well-suited for large-scale genotyping and can be implemented in extensive breeding programs. However, its effectiveness depends on the number of reads generated and the suitable combination of restriction enzymes, which maximizes the reads and thus leads to effective coverage for discovering genomic regions associated with various traits.

Conclusion

This study compared the sdRAD-seq and ddRAD-seq methods for SNP discovery and genetic diversity assessment in safflower (*Carthamus tinctorius* L.) using in silico and in vitro approaches. Three restriction enzyme combinations (ApeKI, NlaIII_MseI, and EcoRI_MseI) were tested in 42 diverse safflower accessions. In silico testing revealed that NlaIII_MseI generated the highest number of DNA fragments. The in vitro results demonstrated that ddRAD-seq outperformed sdRAD-seq in read count, alignment rate, coverage, and SNP detection. Alignment-free k-mer analysis confirmed the superiority of ddRAD-seq, with EcoRI_MseI identifying more core genes than other combinations. EcoRI_MseI captured more SNPs with fewer missing observations. Principal component analysis using ddRAD-seq data explained substantial genetic variation. The study concluded that ddRAD-seq with EcoRI_MseI is the most suitable genotyping-by-sequencing approach for genome sampling and SNP genotyping in safflower.

Data availability

Sequence data that support the findings of this study have been deposited in the Indian Biological Data Centre (IBDC) (<https://ibdc.dbtindia.gov.in/inda/completeStudyDetailsById?studyid=INRP000244>) Accession No. INRP000244 and INSDC Project Accession No. ERP167403 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB83828>) PRJEB83828 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJEB83828>).

Received: 19 December 2024; Accepted: 10 June 2025

Published online: 02 July 2025

References

- Ciaramella, B. R., Corinzia, S. A., Cosentino, S. L. & Testa, G. Phytoremediation of heavy metal contaminated soils using safflower. *Agronomy* **12**, 2302. <https://doi.org/10.3390/agronomy12102302> (2022).
- Poodineh, M., Mahdi, N., Nezhad, Mohammadi-Nejad, G., Ali, B., Ebrahimi, F. & Fakheri, F. Identification of safflower (*Carthamus tinctorius* L.) QTL under drought stress and normal conditions. *Ind. Crop Prod.* **171**, 113889. <https://doi.org/10.1016/j.indcrop.2021.113889> (2021).
- Gongora, B. et al. Comparison of emissions and engine performance of safflower and commercial biodiesels. *Ind. Crop Prod.* **179**, 114680 (2022).
- Sajad, S. S., Karimi, K. & Mirmohamadsadeghi, S. Hydrothermal pretreatment of safflower straw to enhance biogas production. *Energy* **172**, 545–554 (2019).
- Sajad, S. S., Mirmohamadsadeghi, S. & Karimi, K. Biorefinery development based on whole safflower. *Plant. Renew. Energy*. **152**, 399–408. <https://doi.org/10.1016/j.renene.2020.01.049> (2020).
- Amirkhiz, K. F. et al. Evaluation of changes in fatty acid profile, grain, and oil yield of *Carthamus tinctorius* L. in response to foliar application of polyamine compounds under deficit irrigation conditions. *Ind. Crop Prod.* **161**, 113231. <https://doi.org/10.1016/j.indcrop.2020.113231> (2021).
- Mani, V., Lee, S. K., Yeo, Y. & Hahn, B. S. A metabolic perspective and opportunities in Pharmacologically important safflower. *Metabolites* **10**, 253. <https://doi.org/10.3390/metabo10060253> (2020).
- Zemour, K. et al. Effects of genotype and Climatic conditions on the oil content and its fatty acids composition of *Carthamus tinctorius* L. seeds. *Agronomy* **11** (2048). <https://doi.org/10.3390/agronomy11102048> (2021).
- Gengmao, Z. et al. Salinity stress increases secondary metabolites and enzyme activity in safflower. *Ind. Crop Prod.* **64**, 175–181 (2015).
- Scheben, A., Batley, J. & Edwards, D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant. Biotech. J.* **15**, 149–161. <https://doi.org/10.1111/pbi.12645> (2017).
- Zhang, H. et al. Genome-wide DNA polymorphism analysis and molecular marker development for the *Setaria italica* variety SSR41 and positional cloning of the *Setaria* white leaf sheath gene SiWLS1. *Front. Plant. Sci.* <https://doi.org/10.3389/fpls.2021.743782> (2021).
- Pavan, S. et al. Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. *Front. Genet.* **11** (447). <https://doi.org/10.3389/fgene.2020.00447> (2020).
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q. & Thureen, D. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).
- Truong, H. T. et al. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One*. **7**, e37565. <https://doi.org/10.1371/journal.pone.0037565> (2012).
- Raman, H. et al. Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*. *PLoS One*. **9**, e101673. <https://doi.org/10.1371/journal.pone.0101673> (2014).
- Fu, Y. B., Gregory, W., Peterson & Dong, Y. Increasing genome sampling and improving SNP genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. *G3* **6**, 845–856. <https://doi.org/10.1534/g3.115.025775> (2016).
- Sardouei-Nasab, S., Nemati, Z. & Mohammadi-Nejad, G. Phylogenomic investigation of safflower (*Carthamus tinctorius*) and related species using genotyping-by-sequencing (GBS). *Sci. Rep.* **13**, 6212. <https://doi.org/10.1038/s41598-023-33347-0> (2023).
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A. & Kawamoto, K. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. **6**, e19379 (2011).
- D'Agostino, N. & Tripodi, P. NGS-based genotyping, high-throughput phenotyping and genome-wide association studies laid the foundations for next-generation breeding in horticultural crops. *Diversity* **9** (38). <https://doi.org/10.3390/d9030038> (2017).
- Gore, M. et al. Evaluation of target Preparation methods for single-feature polymorphism detection in large complex plant genomes. *Crop Sci.* **47**, S135–S148. <https://doi.org/10.2135/cropsci2007.02.0085tpg> (2007).
- Rabbi, I. et al. Genetic mapping using genotyping-by-sequencing in the clonally propagated Cassava. *Crop. Sci.* **54**, 1384–1396. <https://doi.org/10.2135/cropsci07.0482> (2014).
- Nguyen, T. K., Yu, J., Choi, H. W., In, B. C. & Lim, J. H. Optimization of genotyping-by-sequencing (GBS) in chrysanthemums: selecting proper restriction enzymes for GBS library construction. *Hort Sci. Technol.* **28**, 108–114. <https://doi.org/10.12972/kjst.20180012> (2018).
- Sehgal, D., Rajpal, V. R., Raina, S. N., Sasanuma, T. & Sasakuma, T. Assaying polymorphism at DNA level for genetic diversity diagnostics of the safflower (*Carthamus tinctorius* L.) world germplasm resources. *Genetica* **135**, 457–470 (2009).
- Ruperao, P. et al. A pilot-scale comparison between single and double-digest RAD markers generated using GBS strategy in sesame (*Sesamum indicum* L.). *PLoS One*. e0286599, (2023). <https://doi.org/10.1371/journal.pone.0286599>
- Lajmi, A., Glinka, F. & Privman, E. Optimizing DdRAD sequencing for population genomic studies with DdRADer. *Mol. Eco Resor.* **1–13** <https://doi.org/10.1111/1755-0998.13870> (2023).
- Lepais, O., Weir, J. & SimRAD An R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Mol. Ecol. Resour.* <https://doi.org/10.1111/1755-0998.12273> (2014).
- Wu, Z. et al. The chromosome-scale reference genome of safflower (*Carthamus tinctorius*) provides insights into Linoleic acid and flavonoid biosynthesis. *Plant. Biotech. J.* <https://doi.org/10.1111/pbi.13586> (2021).
- Andrews, S. & FastQC A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170> (2014).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Li, H. The sequence alignment/map format and SAM tools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
- Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770. <https://doi.org/10.1093/bioinformatics/btr011> (2011).
- Ondov, B. D. et al. Mash: fast genome and metagenome distance Estimation using MinHash. *Genome Biol.* <https://doi.org/10.1186/s13059-016-0997-x> (2016).
- Golicz, A. A. et al. Gene loss in the fungal Canola pathogen *Leptosphaeria maculans*. *Funct. Integr. Genomics*. **15**, 189–196 (2015).
- Catchen, J., Hohenlohe, P., Bassham, S., Amores, A. & Cresko, W. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22** (7), 3124–3140 (2013).
- Danecek, P. et al. The variant call format and VCF tools. *Bioinformatics* **27**, 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330> (2011).
- Yin, L. et al. A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Geno Prot. Bioinfo.* <https://doi.org/10.1016/j.gpb.2020.10.007> (2021).
- Cingolani, P. et al. 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. (Austin)*. **6**, 80–92. <https://doi.org/10.4161/fly.19695> (2012).
- Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328. <https://doi.org/10.1093/bioinformatics/bts606> (2012).

40. Fu, Y. B. Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. *G3* **4**, 891–900 (2014).
41. Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J. & Baldwin, J. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–515 (2000).
42. van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F. & Snoeijers, S. Complexity reduction of polymorphic sequences (CROPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* **2**, e1172 (2007).
43. Fu, Y. B. & Peterson, G. W. Genetic diversity analysis with 454 pyrosequencing and genomic reduction confirmed the Eastern and Western division in the cultivated barley gene pool. *Plant. Gen.* **4**, 226–237 (2011).
44. Sonah, H., Bastien, M., Iqura, E., Tardivel, A. & Le 'gare', G. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* **8**, e54603. <https://doi.org/10.1371/journal.pone.0054603> (2013).
45. Ali, F. et al. Genetic diversity, population structure and marker-trait association for 100-Seed weight in international safflower panel using Silico dart marker information. *Plants* **9**, 652. <https://doi.org/10.3390/plants9050652> (2020).
46. Marinus, M. G. & Løbner-Olesen, A. DNA methylation. *Eco Sal Plus* **10** (1128). <https://doi.org/10.1128/ecosalplus.ESP-0003-2013> (2014).
47. Aballay, M. M., Aguirre, N. C. & Filippi, C. V. Fine-tuning the performance of ddRAD-seq in the Peach genome. *Sci. Rep.* **11**, 6298. <https://doi.org/10.1038/s41598-021-85815-0> (2021).
48. Hantze, H. & Horton, P. Effects of spaced k-mers on alignment-free genotyping. *Bioinformatics* **39**, i213–i221 (2023).
49. Zielezinski, A., Girgis, H. Z. & Bernard, G. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20** (44). <https://doi.org/10.1186/s13059-019-1755-7> (2019).
50. Esposito, S. et al. DdRAD sequencing-based genotyping for population structure analysis in cultivated tomato provides new insights into the genomic diversity of mediterranean 'da serbo' type long shelf-life germplasm. *Hortic. Res.* **7**, 134 <https://doi.org/10.1038/s41438-020-00353-6> (2020).

Acknowledgements

We are thankful to the Director, ICAR- National Bureau of Plant Genetic Resources for providing facilities to carry out this research work.

Author contributions

PP methodology, formal analysis and manuscript drafting; GPK methodology, manuscript drafting; NG methodology; RP Resources; JR resources; Rk resources; SSG resources; PK resources, funding acquisition; SR funding acquisition, conceptualization, validation, editing the manuscript.

Funding

This research was funded by Department of Biotechnology, Government of India, grant number BT/Ag/Network/Safflower/2019-20".

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-06706-2>.

Correspondence and requests for materials should be addressed to S.R.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025