High-throughput single nucleotide polymorphism (SNP) identification and mapping in the sesame (*Sesamum indicum* L.) genome with genotyping by sequencing (GBS) analysis

Ayse Ozgur Uncu · Anne Frary ·
Petr Karlovsky · Sami Doganlar

Received: 14 June 2016 / Accepted: 28 November 2016 / Published online: 20 December 2016

Abstract Sesame (*Sesamum indicum* L. syn. *Sesamum orientale* L.) is considered to be the first oil seed crop known to man. Despite its versatile use as an oil seed and a leafy vegetable, sesame is a neglected crop and has not been a subject of molecular genetic research until the last decade. There is thus limited knowledge regarding genome-specific molecular markers that are indispensible for germplasm enhancement, gene identification, and marker-assisted breeding in sesame. In this study, we employed a genotyping by sequencing (GBS) approach to a sesame recombinant inbred line (RIL) population for high-throughput single nucleotide polymorphism (SNP) identification and genotyping. A total of 15,521 SNPs were identified with 14,786 SNPs (95.26 %) located along sesame genome assembly pseudomolecules. By incorporating sesame-specific simple sequence repeat (SSR) markers developed in our previous work, 230.73 megabases (99 %) of sequence from the genome assembly were saturated with markers. This large number of markers will be available for sesame geneticists as a resource for candidate polymorphisms located along the physical chromosomes of sesame. Defining SNP loci in genome assembly sequences provides the flexibility to utilize any genotyping strategy to survey any sesame population. SNPs selected through a high stringency filtering protocol (770 SNPs) for improved map accuracy were used in conjunction with SSR markers (50 SSRs) in linkage analysis, resulting in 13 linkage groups that encompass a total genetic distance of 914 cM with 432 markers (420 SNPs, 12 SSRs). The genetic linkage map constitutes the basis for future work that will involve quantitative trait locus (QTL) analyses of metabolic and agronomic traits in the segregating RIL population.

Keywords SNP · SSR · Molecular markers · Oil crop · Oil seed · Recombinant inbred line

Electronic supplementary material The online version of this article (doi:10.1007/s11032-016-0604-6) contains supplementary material, which is available to authorized users.

A. O. Uncu · A. Frary · S. Doganlar (✉)
Department of Molecular Biology & Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey
e-mail: samidoganlar@iyte.edu.tr

P. Karlovsky
Molecular Phytopathology and Mycotoxin Research, Georg-August-University Göttingen, Göttingen, Germany

Introduction

Sesame (*Sesamum indicum* L. syn. *Sesamum orientale* L.) (2n = 26), is a member of the Pedaliaceae family and is an oil seed crop having the most ancient cultivation history (Bedigian 2003). Sesame is cultivated primarily for its oil. This edible oil is of exceptional quality with a very high degree of resistance against oxidative deterioration (Namiki 2007) and a high unsaturated fatty acid content (Anilakumar et al. 2010). Thus, sesame has gained a reputation as the "Queen of the oil seeds" (Bedigian and Harlan 1986). Sesame oil contains characteristic lipid-soluble antioxidant lignans, sesamin, sesamolin, and sesaminol, which not only protect oil quality but also contribute to its nutraceutical properties (Namiki 2007). The health-beneficial effects of sesame lignans, including antiaging, antihypertensive, and anticancer properties, were reviewed by Namiki (2007) and Anilakumar et al. (2010).

Sesame is indeed a versatile crop which provides a leafy edible vegetable in addition to its common use as an oil seed. The leaves have a rich carotenoid, ascorbic acid, iron, and calcium content and contain good amounts of protein (Bedigian 2003). Therefore, sesame is a potential food-security crop in rural regions of Africa and Asia, where its cultivation is most common. Nevertheless, sesame is not a stereotypical domestic plant and has retained wild plant characteristics such as seed shattering, indeterminate growth, and late and long maturity (Bedigian 2003). These traits impair yield and hinder mechanized harvesting (Day 2000; Uzun et al. 2003; Uzun and Çağırgan 2009). In addition, not much progress has been made in the breeding of high-yielding cultivars with improved seed dispersal, growth habit, and disease resistance traits. Because molecular genetic research in sesame only recently accelerated, it has not been feasible to adopt molecular marker technologies in sesame breeding. Indeed, only a few studies were performed regarding understanding the genetic basis of agronomically important traits in sesame. In such a work, an AFLP (amplified fragment length polymorphism) locus was identified that is linked to the *closed capsule* mutation (Uzun et al. 2003) and an ISSR (inter simple sequence repeat) marker was introduced that is linked to the determinate growth mutant trait (Uzun and Çağırgan 2009). More recently, association analysis of oil, protein, oleic acid, and linoleic acid contents (Wei et al. 2013), and QTL (quantitative trait locus) analyses of seed coat color (Zhang et al. 2013a) and yield-related traits (Wu et al. 2014) were reported. Limited availability of sesame-specific molecular markers has been the primary obstacle in the advancement of molecular genetics in the crop.

Next-generation sequencing (NGS) approaches have recently been used for fast, cost-efficient development of sesame-specific DNA markers. Following NGS-based transcriptome sequencing approaches, Wei et al. (2011) and Zhang et al. (2012) identified 7,702 and 2,164 potential SSR (simple sequence repeat) loci, respectively. In another recent study, we identified 19,816 non-redundant SSRs in the sesame genome using a pyrosequencing approach and made a total of 933 experimentally validated markers publicly available (Uncu et al. 2015). Sequencing of the genome also served for SSR marker development in sesame. In 2014, the draft genome assembly comprising 16 pseudomolecules that encompassed approximately 80 % of the genome was published (Wang et al. 2014). The draft assembly was surveyed for SSRs (Wei et al. 2014), resulting in the identification of 23,438 potential SSR loci, 218 of which were experimentally validated as genomic SSR markers.

Continuous progress in next-generation sequencing technologies has made high-throughput SNP identification and simultaneous genotyping by sample multiplexing a fast and cost-effective route for generating thousands of species-specific markers and large quantities of genotypic data. However, the complexity of plant genomes with many repetitive sequences presents a challenge to sequence alignment and SNP identification (Zou et al. 2014). Fortunately, a variety of protocols such as RRS (reduced representation shotgun sequencing) (Altshuler et al. 2000), CRoPS (complexity reduction of polymorphic sequences) (Van Orsouw et al. 2007), RAD (restriction-site associated DNA) tag sequencing (Baird et al. 2008), GBS (genotyping by sequencing) (Elshire et al. 2011), and SLAF-seq (specific length amplified fragment-sequencing) (Sun et al. 2013), have been established that enable the reduction of genome complexity. All protocols take advantage of restriction enzymes for avoiding repeat-rich sequences in genomes and increasing the abundance of low copy regions in sequencing libraries. As a result, the risks of slippage while aligning repeat-bearing sequences and false alignment of non-homologous loci based on repeat identity are minimized, improving the accuracy of sequence alignment and polymorphism identification.

Genotyping by sequencing (GBS) is a refined version of RAD tag sequencing. This approach was first demonstrated by Elshire et al. (2011) on maize (*Zea mays* L.) and barley (*Hordeum vulgare* L.). Compared to RAD sequencing, GBS has fewer sample preparation steps and library preparation does not involve a fragment size selection procedure. Sample multiplexing is highly simplified by simultaneously ligating barcode and common adapters prior to sample pooling. Because GBS introduced simplicity and cost-efficiency into multiplex reduced-representation sequencing protocols, the approach was readily adopted by the plant genetics and breeding community (Poland and Rife 2012; He et al. 2014).

In the present work, a GBS approach was used for high-throughput SNP identification and genotyping in an intraspecific sesame RIL population. The presence of a genome assembly enabled identification of the SNP loci in their sequence context. Sesame-specific SSR markers were also located to the assembly sequences. Linkage analysis was carried out in order to construct a linkage map for further QTL analyses in the RIL population. SNP alleles and locations are publicly available and are a valuable resource for sesame molecular genetics and breeding.

Materials and methods

Plant material and DNA isolation

An intraspecific recombinant inbred line population (F_6-RILs) derived from the cross *S. indicum* (Acc. No. 95-223) × *S. indicum* (Acc. No. 92-3091) was used as plant material for GBS analysis. The geographical origins of the parental accessions 95-223 and 92-3091 were Africa and Korea, respectively. Parental accessions were obtained from Centro Nacional de Investigaciones Agropecuarias (CENIAP) Germplasm Bank (Venezuela) and the recombinant inbred line population was generated in Georg-August-University Göttingen (Germany).

Three seeds per F_6-RIL were planted and grown in soil containing peat moss, perlite, and natural fertilizer in the greenhouse facility at Izmir Institute of Technology. Genomic DNA from 91 RILs and parental accessions was isolated from liquid nitrogen-frozen ground leaf tissue using the NucleoSpin Plant II Kit (Macherey Nagel, Duren, Germany), according to the manufacturer's instructions.

GBS library preparation and sequencing

Integrity of the DNA isolated from parental accessions and F_6-RILs was checked on a 1 % agarose gel. The concentration of DNA was measured using a Qubit 2.0 Fluorometer (Life Technologies, Thermo Fisher Scientific Inc., Waltham, MA, USA) with dsDNA BR Assay Kit (Life Technologies). All sample concentrations were adjusted to 10 ng/μL for GBS analysis. Next-generation sequencing library preparation procedure, including sample DNA digestion, common adapter and barcode adapter ligation, sample pooling and sample pool amplification was done as described in Elshire et al. (2011). Single-end sequencing of the 93-plex library (91 RILs and parental accessions) was done with a Genome Analyzer II device in a single flowcell channel (Illumina Inc., San Diego, CA, USA). Library preparation and sequencing were carried out at the University of Wisconsin-Madison Biotechnology Center.

Sequence alignment and SNP calling

Raw sequence processing, alignment, and SNP calling were performed at the University of Wisconsin-Madison Biotechnology Center. Raw sequence reads were converted to a FASTQ file by CASAVA 1.8.2 (Illumina Inc.) for further processing. To initiate data analysis with the GBS Discovery Pipeline (Glaubitz et al. 2014) of TASSEL Version 3.0 (Bradbury et al. 2007), the FASTQ file and barcode key file that listed the plate layout and barcodes for each sample were used as input files. Raw sequences and barcode key file can be accessed at https://figshare.com/articles/Sesame_GBS_sequence_and_marker_data/3168328. Using the FastqToTagCountPlugin of the pipeline, reads that began with the expected barcodes followed by an *Ape*KI cut site remnant (CWGC) were trimmed to 64 bases. Sequence reads with N (unidentified base) in the first 64 bases after the barcode were eliminated. Reads with an intact enzyme cut site or the beginning of the common adapter were truncated and padded to 64 bases with poly A. The reads were then sorted to merge the redundant reads into single tags and resultant tags were listed as a tagCount file by the plugin. MergeMultipleTagCountPlugin produced the merged tagCount file, the file was converted to FASTQ format by the TagCountToFastqPlugin to be used as the input file for tag alignment to the draft sesame genome assembly (Wang et al. 2014) by bowtie2 plugin. Pseudomolecule sequences of the assembly were downloaded from the Sinbase database (http://ocri-genomics.org/Sinbase/login.htm). The genome assembly comprised of 16 pseudomolecules of assembled scaffolds and one group of concatenated contigs with stretches of 100 N as the spacer. The output of the alignment was converted to a TOPM (Tags On Physical Map) file by SAMConverterPlugin for SNP calling from the alignment. Sequence reads sorted and demultiplexed according to their barcode adapters by the FastqToTBTPlugin were kept as a

TBT file (Tags By Taxa). TOPM and TBT files were used by the TagsToSNPByAlignmentPlugin for SNP calling. Non-default parameters used for SNP calling were: Minimum value of F (inbreeding coefficient = 1-Ho/He) [mnF]: 0.8, Minimum minor allele count (default: 10) [mnMAC]: 100000. SNPs that pass the mnMAC threshold were kept in the output HapMap file for each sesame pseudomolecule. Duplicate SNPs in the HapMap files were merged by the MergeDuplicateSNPsPlugin. In order to allow heterozygosity detection in SNP loci, callHets option of the plugin was switched to True. Threshold for genotypic mismatch rate (misMat) was set as 0.1.

HapMap files generated for the 16 sesame genome assembly pseudomolecules contained position information for each SNP. In addition to SNP loci, a total of 2,465 genomic SSR markers introduced in our previous work (Uncu et al. 2015) were incorporated in the HapMap files. In order to locate the SSR markers to their positions in the genome assembly, the FASTA file containing the assembly scaffolds (http://ocri-genomics.org/Sinbase/login.htm) was uploaded to BioEdit software Version 7.0.5.3 (Hall 1999) using the "Create a Local Nucleotide Database" option. Local BLAST tool of the software was used to align SSR primer sequences to the genome assembly. Alignments with a 100 % match score were accepted and incorporated into the HapMap files.

SSR Amplification

SSR markers to be used for linkage analysis were determined by performing a parental survey. SSR markers that were found to be polymorphic between the parental genotypes were genotyped in the F_6-RIL population. SSR alleles were amplified in 20-μL reaction mixtures containing 1X Colorless GoTaq Flexi PCR buffer, 1.5 mM $MgCl_2$, 0.25 mM of each deoxyribonucleotide triphosphate (dNTP) (Promega Corp., Madison, WI, USA), 1 U GoTaq G2 Flexi DNA Polymerase (Promega Corp.), 0.25 μM of each primer, and 50 ng template DNA. Thermal cycling conditions consisted of one cycle of initial denaturation for 10 min at 94 °C, followed by 35 cycles at 94 °C for 30 s, 55 °C for 30 s, 72 °C for 45 s, with a final extension step of 10 min at 72 °C. PCR products were then run on a Fragment Analyzer^{TM} (Advanced Analytical, Ames, IA, USA) capillary electrophoresis system using the DNF-900 dsDNA Reagent Kit (Advanced Analytical) according to the manufacturer's instructions. SSR alleles were visualized and scored using the PROSize 2.0^{TM} software version 1.2.1.1 (Advanced Analytical).

Genetic linkage map construction

SNP genotype data obtained from GBS analysis were used for the construction of a genetic linkage map. SSR markers were used in conjunction with SNPs for map construction. Merged SNPs were filtered further with the GBSHapMapFiltersPlugin in order to select SNPs appropriate for linkage analysis in the RIL population. Non-default SNP filtering parameters were: Minimum taxon coverage [mnTCov]: 0.01, Minimum site coverage [mnSCov]: 0.5, filtering for SNPs in statistically significant LD (Linkage Disequilibrium) with at least one neighboring SNP [hLD]: True, Minimum R^2 value for the LD filter [-mnR2]: 0.2, Minimum Bonferroni-corrected p-value for the LD filter [-mnBonP]: 0.005. A genetic linkage map was constructed using the JoinMap 4.0 computer program (Van Ooijen 2006). Marker order was determined with the regression mapping algorithm using a maximum recombination frequency threshold of 0.40. Minimum logarithm of odds (LOD) threshold was set as 6 and a goodness-of-fit jump threshold for loci removal was set as 5. The ripple command was adjusted to confirm marker order after the addition of each locus. Map distances were calculated with the Kosambi mapping function (Kosambi 1943). Linkage groups were visualized with the MapChart 2.3 computer program (Voorrips 2002).

Results and discussion

A low rate of marker polymorphism has been consistently reported for sesame by several authors (Dixit et al. 2005; Wei et al. 2009; Wang et al. 2012; Zhang et al. 2013b). In addition, as sesame is a neglected crop species, the number of available sesame-specific markers is limited. The low marker polymorphism rate and the limited pool of genome-specific markers are obstacles for gene/QTL mapping, map-based cloning, and acceleration of molecular breeding in sesame. Therefore, if gene/QTL mapping is intended, it is necessary to utilize novel approaches that allow simultaneous polymorphism identification and genotyping in experimental sesame populations. GBS has been demonstrated as a versatile approach for large-scale SNP identification and genotyping with work on a wide range of plant taxa (Elshire et al. 2011; Poland et al. 2012a, 2012b; Huang et al. 2014; Jarquín et al. 2014; Russell et al. 2014; Hart and Griffiths 2015; Islam et al. 2015; Jaganathan et al. 2015). In this work, a GBS approach (Elshire et al. 2011) was employed for high-throughput SNP marker development and genotyping in a F_6-RIL mapping population.

The major advantage of using recombinant inbred lines in gene/QTL mapping studies is that RILs constitute permanent mapping populations. Because segregation is totally or almost complete in RILs, the propagation of genotypes can be unlimited. In trait mapping studies, RIL populations can be confidently evaluated over subsequent years and in many different environments. Therefore, RILs improve the precision of detecting the genetic component of variance while analyzing quantitative traits (Burr et al. 1988). The parents of the sesame F_6-RIL population used in this work were identified as genetically distinct genotypes by Laurentin and Karlovsky (2006). Moreover, the parental genotypes produced distinct profiles when evaluated for their secondary metabolite content (unpublished data), and, therefore, were used for generating a RIL population in order to study the inheritance of secondary metabolite accumulation in sesame. Because the high cost of manual harvesting restricts the economic profit of sesame cultivation (Uzun et al. 2003; Georgiev et al. 2008), suitability for mechanized harvesting is a new direction in sesame breeding (Georgiev et al. 2008). Preliminary evaluation of the RIL population under greenhouse conditions revealed segregation for traits such as height at first branch, branch number and length, and stem thickness, which are morphological indices of suitability for mechanized harvesting (Georgiev et al. 2008). Thus, the genotypic data obtained through GBS will be utilized in further work to identify the genetic control of traits related to secondary metabolite accumulation and suitability for mechanical harvesting in sesame.

Sequence filtering and tag alignment

A total of 343,970,622 raw sequence reads were obtained from sequencing of the GBS library representing 91 F_6-RILs and two parental accessions. The number of accepted reads containing the expected barcodes and the enzyme cut site remnant was 164,433,076, comprising 47.8 % of the total reads (Table S1). Out of the 91 RILs, two were excluded from the analyses due to a very low rate of representation in the pool of accepted reads with less than 200 sequences. Elshire et al. (2011) indicate that inter-sample variation is valid for all multiplex sequencing protocols and arises from accuracy problems with DNA quantification methods. The authors highlight the necessity of developing high-throughput methods with improved precision for quantifying high molecular weight DNA for multiplex sequencing analyses. However, in our work only two RILs had insufficient sequence data and were excluded from the analysis. The most highly represented genotype yielded 3,296,818 appropriately barcoded reads and the lowest number of accepted reads per genotype was 13,592. The average number of reads obtained from RILs was 1,758,127. The two parental accessions, *S. indicum* Acc. No. 95-223 and Acc. No. 92-3091, yielded 1,523,490 and 2,920,019 good quality reads, respectively. The merged tagCount file, generated by collapsing reads into a set of unique sequence tags called “merged tags”, contained 416,048 tags (Table S1).

Haploid chromosome number of the sesame genome is 13. The sesame draft genome assembly (Wang et al. 2014) comprises 16 pseudomolecules of assembled scaffolds that represent the 13 chromosomes and a 17th group of unanchored, concatenated contigs. The output of tag alignment to the draft assembly (TOPM file) contained 416,048 tags (merged tags) (Table S1) which represented a total of 157,667,909 tag sequences from the RIL population and the parental accessions. As a result of the alignment, 383,890 tags (92.27 % of the merged tags) were aligned to the assembly. Detailed statistics of tag aligment are provided in Table S1.

SNP identification

Sequence reads sorted by taxa were used in conjunction with the TOPM file for SNP calling from the tag alignment. As a result, a total of 16,705 raw SNPs were identified (Table 1). Raw SNPs were further processed by merging duplicates (redundant SNP loci identified in reads from both directions). When duplicates were merged, the resultant number of unique SNPs was 15,521, with 14,786 SNPs (95.26 %) located in the genome assembly and 735 SNPs (4.74 %) represented the 17th group consisting of unanchored contigs (Table 1). Transition mutations were predominant (58 %) among the identified SNPs. Location and allelic variant information for the merged SNPs can be accessed at https://figshare.com/articles/Sesame_GBS_sequence_and_marker_data/3168328. With 2,360 raw

Table 1 Single nucleotide polymorphism (SNP) identification statistics

LG[a]	Raw SNPs	Merged SNPs	Filtered SNPs
1	1,612	1,496	11
2	908	844	27
3	2,360	2,192	29
4	1,071	995	53
5	1,221	1,116	58
6	1,053	970	61
7	1,015	960	40
8	1,061	991	19
9	960	906	5
10	1,010	939	24
11	1,078	1,023	28
12	1,044	986	5
13	406	366	201
14	202	182	12
15	765	704	7
16	137	116	12
17	802	735	178
TOTAL	16,705	15,521	770
Located in assembly (%)	95.2	95.26	76.88
Located in contigs (%)	4.8	4.74	23.12

[a] LG (Linkage Group) stands for genome assembly pseudomolecules.

Pseudomolecule 17 corresponds to previously unanchored sequences

and 2,192 merged SNPs, the most SNPs were located to pseudomolecule 3 of the sesame genome assembly. Conversely, the lowest number of SNPs (137 raw and 116 merged SNPs) were located to pseudomolecule 16 of the assembly (Table 1).

The availability of a reference genome is an important advantage for GBS analysis. Tag alignment to a reference genome prior to SNP calling enables determination of the precise order of the identified markers relative to physical sequences (Poland and Rife 2012). Thus, the presence of a sesame genome assembly was highly beneficial, because it complemented an important shortcoming of GBS analysis. The GBS protocol (Elshire et al. 2011) produces sequence tags of maximum 64 bp length, and the library preparation protocol does not yield overlapping fragments which would enable sequence assembly. Therefore, unless SNP markers identified through GBS are located in a sequence context, they cannot be further utilized for genotyping in other populations and their utilization remains exclusive to the population genotyped in GBS analysis. In this work, it was feasible to locate almost 15,000 SNPs along physical chromosomes, which will enable primer/probe design for the further utilization of this large number of SNP markers for genotyping any sesame population. Moreover, the reference assembly allowed saturation of the pseudomolecules with SSRs, in addition to the SNP markers. In a previous study, we identified 5,727 genomic SSR loci in a contig assembly generated through pyrosequencing of the sesame genome (Uncu et al. 2015). Out of the 5,727 loci, 2,465 allowed successful primer design with sufficient length of flanking sequences. In this work, a total of 1,854 SSR markers displayed perfect alignment with the genome assembly sequences, 1,436 (77.45 %) of which were located along the 16 pseudomolecules and 418 (22.55 %) aligned with the unassembled contig sequences (Table 2). Incorporation of the SSR location data to the HapMap files resulted in a total of 16,222 markers with precise physical location information in the sesame genome assembly.

The reference assembly (Wang et al. 2014) encompasses a physical distance of 233.23 megabases (Mb). SNP and SSR markers covered almost the entire physical chromosomes, encompassing 230.73 Mb, corresponding to 99 % of the physical distance covered by the assembly. Marker positions along physical chromosomes can be accessed at https://figshare.com/articles/Sesame_GBS_sequence_and_marker_data/3168328.

The highest coverage was obtained for Pseudomolecule 10 (99.71 %) with a total of 1,054 markers (Table 2). Conversely, Pseudomolecule 13 had the lowest coverage rate of 95.25 % with 394 markers (Table 2). Average density of the markers located along the pseudomolecules was 14.71 kilobases (kb)/marker interval. Overall, the distance between adjacent markers was smaller than 10 kb for the majority of marker intervals (78 %), thus the combination of SNP and SSR loci provided a good resolution to saturate the genome assembly with markers.

Construction of a genetic linkage map

In order to determine a core set of SNPs appropriate for use in segregation analyses for gene/QTL mapping in the genotyped F_6-RIL population, merged SNPs were

Table 2 Statistics of marker localization in the sesame genome assembly

LG[a]	Located SSRs	Total number of located markers	Pseudomolecule size (Mb)	Distance encompassed (Mb)	Coverage (%)	Marker density (kb)
1	114	1,610	18.58	18.43	99.2	11.45
2	109	953	18.5	18.4	99.5	19.31
3	140	2,332	24.93	24.84	99.64	10.65
4	99	1,094	17.36	16.86	97.12	15.41
5	99	1,215	18.9	18.81	99.52	15.48
6	153	1,123	25.29	25	98.9	22.27
7	60	1,020	11.73	11.66	99.4	11.43
8	149	1,140	21.52	21.39	99.44	18.77
9	71	977	12.41	12.27	98.9	12.56
10	115	1,054	17.25	17.2	99.71	16.32
11	101	1,124	15.45	15.23	98.58	13.56
12	51	1,037	6.37	6.25	98.12	6.02
13	28	394	5.05	4.81	95.25	12.22
14	32	214	4.88	4.83	98.98	22.56
15	85	789	10.05	9.99	99.4	12.66
16	30	146	4.96	4.76	95.97	32.6
17	418	1,153				

[a] LG (Linkage Group) stands for genome assembly pseudomolecules. Pseudomolecule 17 corresponds to previously unanchored sequences

SSR simple sequence repeat

filtered for minimum taxa coverage, minimum locus coverage, and LD. Because GBS analysis yielded an excess of SNPs, a sufficient number of SNPs (770 SNPs) were still retained as a result of filtering and used in linkage analysis for the construction of a genetic linkage map of the sesame genome. The first report of high-throughput SNP identification and genotyping in sesame describes the application of the SLAF-seq approach (Zhang et al. 2013b). More recently, RAD tag sequencing was applied on a sesame RIL population (Wu et al. 2014). These two multiplex reduced representation sequencing approaches yielded a total of 3,673 and 3,769 polymorphic loci, respectively. In this work, the percentage of SNP markers used for linkage analysis in the RIL population was 4.96 % (770 markers), lower than those reported by Zhang et al. (2013b) and Wu et al. (2014), who incorporated 34.63 % (1,272 markers) and 35.21 % (1,327 markers) of the markers they identified into their linkage analyses, respectively. Presumably, this was the result of the high stringency SNP filtering protocol applied in this work, which included a high LD filter for recombinant inbred populations that eliminates SNPs with a high rate of falsely genotyped individuals. This is reflected by a lack of surrounding SNPs that are in LD with the locus (Glaubitz et al. 2014).

A total of 820 markers (770 SNPs and 50 SSRs) were used to construct the linkage map (Fig. 1). Out of the 820 markers, 432 (420 SNPs, 12 SSRs) were mapped into 13 linkage groups (LGs) that span a total genetic distance of 914 cM. Average marker density of the map was 2.61 cM per marker interval. Number of markers ranged from 10 to 101 and linkage groups ranged from 5.9 to 188.19 cM long (Table 3). The linkage map had a total of 15 gaps larger than 10 cM located in linkage groups 2, 3, 5, 6, 8, 12 and 13 (Fig. 1). In parallel with the results of this work, Wu et al. (2014) reported 16 gaps (>10 cM) in their sesame genetic linkage map constructed through a RAD tag sequencing approach. When reduced-representation sequencing protocols are employed, it is reasonable to expect that certain regions of the genome are under-represented. Indeed, the degree of uniformity in genome representation is unknown for sequence tags generated through GBS analysis (Poland and Rife 2012). Among mapped markers, 112 (25.93 %) were identified in the unanchored contigs of the draft

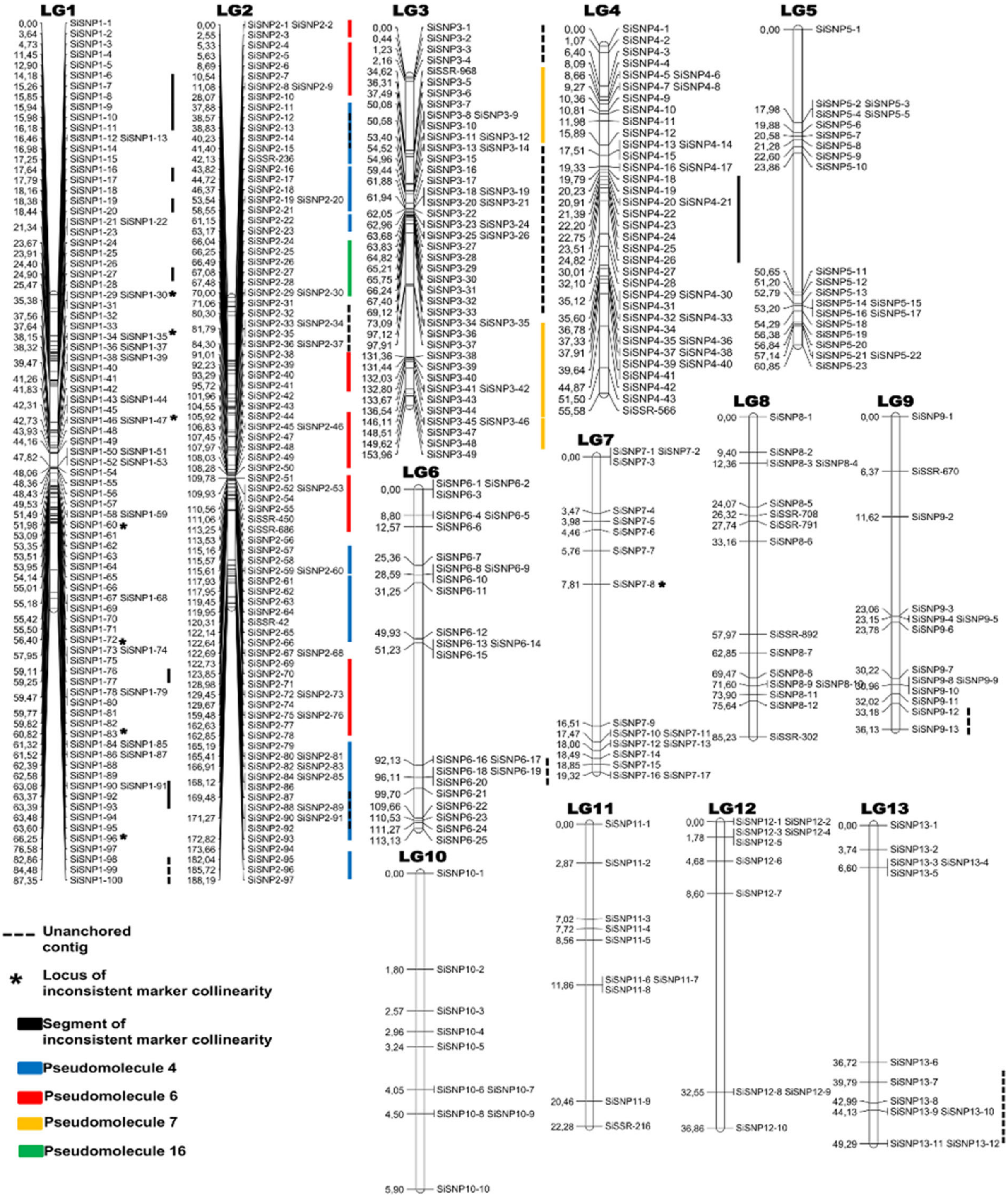

Fig. 1 Genetic linkage map of the sesame genome constructed by genotyping by sequencing (GBS) analysis. Marker locations (cM) are displayed on the left side of each linkage group. Color-coded and dashed bars depict marker blocks from different pseudomolecules and unanchored sequences, respectively. Overlapping color-coded and dashed bars indicate marker clusters from unanchored sequences that interrupt pseudomolecule segments

Table 3 Distribution of markers in linkage groups

LG[a]	Number of markers	Represented pseudomolecule[b]	Size (cM)	Marker density (cM per marker interval)
1	100	13, 17	87.35	0.87
2	101	4, 6, 16, 17	188.19	1.86
3	50	7, 17	153.96	3.08
4	44	5, 17	55.58	1.26
5	23	17	60.85	2.65
6	25	17	113.13	4.52
7	17	2	19.32	1.14
8	16	3	85.23	5.33
9	14	3, 17	36.13	2.58
10	10	17	5.9	0.59
11	10	10	22.28	2.23
12	10	10	36.86	3.69
13	12	1, 17	49.29	4.11
Total	432		914.07	2.61

[a, b] Single nucleotide polymorphism (SNP) locations in draft genome assembly are given as pseudomolecule numbers. LG (Linkage Group) stands for genome assembly pseudomolecules. Pseudomolecule 17 corresponds to previously unanchored sequences

genome assembly and the rest (74.07 %) were identified in the pseudomolecule sequences. Marker clusters representing the unanchored sequences encompassed a genetic distance of 124,86 cM, corresponding to 13.66 % of the total map distance. Two pairs of linkage groups, designated as LG8 & 9, and LG11 & 12, corresponded to pseudomolecules 3 and 10, respectively (Table 3). Thus, the two pseudomolecules segregated as distinct linkage groups according to our analysis. In addition, segments from three distinct pseudomolecules (4, 6, and 16) were identified as a single linkage group in our analysis. Nucleotide locations of mapped SNP and SSR markers on assembly pseudomolecules and primer sequences of polymorphic SSRs are available at https://figshare.com/articles/Sesame_GBS_sequence_and_marker_data/3168328.

LG1 consisted of 100 SNPs, with the vast majority (97 SNPs) representing a 3.82 Mb interval on the pseudomolecule 13 of the genome assembly and three SNPs located in unanchored sequences (Fig. 1). Small intervals and single loci of contrasting collinearity with the pseudomolecule sequence are shown in Fig. 1. Markers that represent segments from the pseudomolecules 4, 6, and 16 co-segregated into a single linkage group (LG2) (Fig. 1). Seven segments from pseudomolecule 4 (0.844–0.846, 2.64–2.92, 4.95–8.23, 8.29–9.39, 9.45–9.47, 11.18–11.46, and 12.06–12.11 Mb) were represented in LG2. Marker order was consistently inverted in the linkage group with respect to pseudomolecule 4, except for the marker cluster between 117.93 and 122.64 cM, corresponding to the 4.95–8.23 Mb sequence interval. A total of six pseudomolecule 6 segments (3.68–3.73, 4.47–4.69, 19.34–21.63, 22.15–22.19, 22.26– 22.37, and 23.21–23.45 Mb) were represented in LG2. Marker collinearity was conserved with respect to the pseudomolecule sequence except for the two marker clusters located at 105.92–108.28 and 122.73–162.85 cM intervals, representing the 22.15–22.19 and 19.34–21.63 Mb sequence intervals, respectively. A 4 cM region in the linkage group consisted of markers located to the pseudomolecule 16 at an interval between 2.30 and 2.77 Mb. Four SSR markers were also mapped to LG2. Marker clusters that correspond to segments from the unanchored contigs were interspersed in the linkage group. LG3 corresponded to co-segregating segments from pseudomolecule 7 and unanchored contigs (Fig. 1). Three segments of 2.46, 3.25, and 0.05 Mb length from pseudomolecule 7 corresponded to map intervals of 18.78, 63.45, and 3.52 cM, respectively. A single SSR marker located to pseudomolecule 7 was also mapped to the linkage group. LG4 consisted solely of markers representing pseudomolecule 5, except for two individual SNP loci from the unanchored sequences. Two marker clusters at the extremes of the linkage group represented 3.69 and 0.98 Mb segments from pseudomolecule 5 in an inverted order. At the center of the linkage group was a 5 cM region that disrupted marker collinearity between the map and the genome assembly. LG5 and LG10 consisted of markers representing previously unanchored contig sequences (Fig. 1). SNP loci, representing a total of 4.3 Mb sequence from pseudomolecule 11, were located in LG6 (Fig. 1). A small marker block that represents unanchored sequences was also mapped to the linkage group. LG7 consisted of markers that represent two sequence blocks from pseudomolecule 2 (Fig. 1). The two marker blocks represented a total of 0.54 Mb sequence in an inverted order with respect to the pseudomolecule. Linkage groups 8 and 9 consisted of markers located at the pseudomolecule 3 of the genome assembly (Fig. 1). A 8.93 Mb portion of pseudomolecule 3 was represented in LG8. In LG9, a 2.53 Mb pseudomolecule

3 sequence was represented by 13 SNPs and one SSR. Marker order was inverted with respect to the pseudomolecule in the linkage group. Linkage groups 11 and 12 consisted of markers representing pseudomolecule 10 of the genome assembly (Fig. 1). A 1.92 Mb sequence was represented in two linkage groups, encompassing a total genetic distance of 5,912 cM. LG13 consisted of markers representing two co-segregating portions of sequence from pseudomolecule 1 and unanchored contigs (Fig. 1).

In the present study, the GBS approach, which enables simultaneous high-throughput SNP marker development and genotyping, was applied for the first time in sesame. GBS enabled the identification of more than 15,000 SNP loci in the sesame genome, proving superiority over SLAF-seq and RAD taq sequencing approaches in high-throughput polymorphism discovery. Because GBS was initially intended for marker discovery useful for trait mapping and genome wide association studies, the TASSEL GBS Pipeline is optimized for identifying as many markers as possible to ensure that some would segregate with traits of interest (Glaubitz et al. 2014). Therefore, GBS seems to be the optimal choice for the identification of a sufficient number of candidate polymorphisms, especially when working with relatively low-diversity species. In addition to the default properties of the GBS data analysis protocol, successful polymorphism discovery is also dependent on the underlying SNP density that reflects the level of existing diversity in the population (Glaubitz et al. 2014). While the average number of reads obtained from each genotype was roughly comparable in the present study (1,758,127 reads) and other work that utilized SLAF-seq (1,416,287 reads) (Zhang et al. 2013b) and RAD tag sequencing (1,644,718 reads) (Wu et al. 2014) approaches, GBS yielded over four times as many SNPs than the other two approaches. This result can be attributed to population choice, since the studies by Zhang et al. (2013b) and Wu et al. (2014) both employed mapping populations derived from parental accessions with similar geographical origins (China). In contrast, the present work used parents from diverse geographical locations (Africa and Northeast Asia). Indeed, Wu et al. (2014) indicated that the number of discovered SNPs (3769) was much less than expected and attributed it to the low level of genetic dissimilarity between the parents of the mapping population.

Taking advantage of the genome assembly (Wang et al. 2014), it was feasible to precisely locate sesame-specific SSR markers as well as newly identified SNP loci in a sequence context, resulting in a reserve of more than 16,000 markers with potential to be utilized for genotyping any sesame population. Thus, the outcomes of this work represent a significant contribution to the existing molecular genetic tools specific for the sesame genome and we anticipate that researchers will benefit greatly from the large number of markers as a resource for candidate polymorphic loci with defined genome sequence contexts.

Despite the high stringency linkage disequilibrium filter applied prior to linkage analysis, it was feasible to construct a genetic linkage map with the expected number of linkage groups that corresponded to the haploid chromosome number of sesame (n = 13). The genetic linkage map constructed using the RIL population constitutes the basis for future work that will involve mapping metabolic and agronomic traits in the sesame genome. Because RILs are immortal genotypes that can be propagated infinitely, once a linkage map is established, it can continuously be improved by reanalyzing preexisting genotypic data in conjunction with data from new analyses. Since segregation is totally or almost completely absent in RILs, it will be feasible to evaluate the population by field trials over subsequent years for metabolic traits and traits that are associated with yield and suitability for mechanized harvesting. Therefore, further utilization of the outcomes of this work will be of importance in facilitating the development of new sesame cultivars with improved yield and superior oil quality due to elevated concentrations of health beneficial antioxidant compounds.

Compliance with ethical standards

Funding This study was supported by grant 108O478 from the Scientific and Technological Research Council of Turkey (TUBITAK).

Conflict of interest The authors declare that they have no conflicts of interest.

References

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L et al (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516

Anilakumar KR, Pal A, Khanum F, Bawa AS (2010) Nutritional, medicinal and industrial uses of sesame (*Sesamum indicum* L.) seeds - An overview. Agric Conspec Sci 75:159–168

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376

Bedigian D (2003) Evolution of sesame revisited: domestication, diversity and prospects. Genet Resour Crop Ev 50:779–787

Bedigian D, Harlan JR (1986) Evidence for the cultivation of sesame in the ancient world. Econ Bot 40:137–154

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635

Burr B, Burr FA, Thompson KH, Albertson MC, Stuber CW (1988) Gene mapping with recombinant inbreds in maize. Genetics 118:519–526

Day JS (2000) Development and maturation of sesame seeds and capsules. Field Crops Res 67:1–9

Dixit A, Jin MH, Chung JW, Yu JW, Chung HK, Ma KH et al (2005) Development of polymorphic microsatellite markers in sesame (*Sesamum indicum* L.). Mol Ecol Notes 5:736–738

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al (2011) A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379

Georgiev S, Stamatov S, Deshev M (2008) Requirements to sesame (*Sesamum indicum* L.) cultivars breeding for mechanized harvesting. Bulg J Agric Sci 14:616–620

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9:e90346

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acid S 41:95–98

Hart JP, Griffiths PD (2015) Genotyping-by-Sequencing enabled mapping and marker development for the By-2 Potyvirus resistance allele in common bean. The Plant Genome. doi:10.3835/plantgenome2014.09.0058

He J, Zhao X, Laroche A, Lu ZX, Liu HK, Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelarate plant breeding. Front Plant Sci 5:484

Huang YF, Poland JA, Wight CP, Jackson EW, Tinker NA (2014) Using Genotyping-By-Sequencing (GBS) for Genomic Discovery in Cultivated Oat. PLoS ONE 9:e102448

Islam MS, Thyssen GN, Jenkins JN, Fang DD (2015) Detection, validation, and application of Genotyping-by-Sequencing based single nucleotide polymorphisms in Upland cotton. The Plant Genome. doi:10.3835/plantgenome2014.07.0034

Jaganathan D, Thudi M, Kale S, Azam S, Roorkiwal M, Gaur PM et al (2015) Genotyping-by-sequencing based intra-specific genetic map refines a "QTL-hotspot" region for drought tolerance in chickpea. Mol Genet Genomics 290:559–571

Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G et al (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 215:740

Kosambi DD (1943) The estimation of map distances from recombination values. Ann Hum Genet 12:172–175

Laurentin HE, Karlovsky P (2006) Genetic relationship and diversity in a sesame (*Sesamum indicum* L.) germplasm collection using amplified fragment length polymorphism (AFLP). BMC Genet 7:10

Namiki M (2007) Nutraceutical functions of sesame: A review. Crit Rev Food Sci 47:651–673

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5:92–102

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012a) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y et al (2012b) Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome 5:103–113

Russell J, Hackett C, Hedley P, Liu H, Milne L, Bayer M et al (2014) The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. Mol Breeding 33:835–849

Sun X, Liu D, Zhang X, Li W, Liu H, Hong W et al (2013) SLAF-seq: an efficient method of large-scale De novo SNP discovery and genotyping using high-throughput sequencing. PLoS ONE 8:e58700

Uncu AO, Gultekin V, Allmer J, Frary A, Doganlar S (2015) Genomic simple sequence repeat markers reveal patterns of genetic relatedness and diversity in sesame. The Plant Genome. doi:10.3835/plantgenome2014.11.0087

Uzun B, Çağırgan MI (2009) Identification of molecular markers linked to determinate growth habit in sesame. Euphytica 166: 379–384

Uzun B, Lee D, Donini P, Çağırgan MI (2003) Identification of a molecular marker linked to closed capsule mutant trait in sesame using AFLP. Plant Breeding 122:95–95

Van Ooijen JW (2006) JoinMap® 4, Software for calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, Netherlands

Van Orsouw NJ, Hogers RCJ, Janssen A, Yalçın F, Snoeijers S, Verstege E et al (2007) Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 11:e1172

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Wang L, Yu J, Li D, Zhang X (2015) Sinbase: an integrated database to study genomics, genetics and comparative genomics in *Sesamum indicum*. Plant Cell Physiology 56:e2. doi:10.1093/pcp/pcu175

Wang L, Yu S, Tong C, Zhao Y, Liu Y, Song C et al (2014) Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. Genome Biol 15:R39

Wang L, Zhang Y, Qi X, Gao Y, Zhang X (2012) Development and characterization of 59 polymorphic cDNA-SSR markers for the edible oil crop *Sesamum indicum* (Pedaliaceae). Am J Bot 99:e394–e398

Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D et al (2011) Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics 12:451

Wei X, Wang L, Zhang Y, Qi X, Wang X, Ding X et al (2014) Development of simple sequence repeat (SSR) markers of sesame (*Sesamum indicum*) from a genome survey. Molecules 19:5150–5162

Wei W, Zhang Y, Lü H, Li D, Wang L, Zhang X (2013) Association analysis for quality traits in a diverse panel of Chinese sesame (*Sesamum indicum* L.) germplasm. J Integr Plant Biol 55:745–748

Wei LB, Zhang HY, Zheng YZ, Miao HM, Zhang TZ, Guo WZ (2009) A genetic linkage map construction for sesame (*Sesamum indicum* L.). Genes Genom 31:199–208

Wu K, Liu H, Yang M, Tao Y, Ma H, Wu W et al (2014) High-density genetic map construction and QTLs analysis of grain yield-related traits in sesame (*Sesamum indicum* L.) based on RAD-seq technology. BMC Plant Biol 14:274

Zhang H, Miao H, Wei L, Li C, Zhao R, Wang C (2013a) Genetic analysis and QTL mapping of seed coat color in sesame (*Sesamum indicum* L.). PLoS ONE 8:e63898

Zhang Y, Wang L, Xin H, Li D, Ma C, Ding X et al (2013b) Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. BMC Plant Biol 13:141

Zhang H, Wei L, Miao H, Zhang T, Wang C (2012) Development and validation of genic-SSR markers in sesame by RNA-seq. BMC Genomics 13:316

Zou X, Shi C, Austin RS, Merico D, Munholland S, Marsolais F et al (2014) Genome-wide single nucleotide polymorphism and insertion-deletion discovery through next-generation sequencing of reduced representation libraries in common bean. Mol Breeding 33:769–778