ORIGINAL RESEARCH
published: 06 July 2017
doi: 10.3389/fpls.2017.01189

Genetic Diversity, Population Structure, and Linkage Disequilibrium of an Association-Mapping Panel Revealed by Genome-Wide SNP Markers in Sesame

Chengqi Cui[1†], Hongxian Mei[2,3,4†], Yanyang Liu[2,3,4], Haiyang Zhang[2,3,4] and Yongzhan Zheng[2,3,4*]*

[1] *National Key Laboratory of Crop Genetics and Germplasm Enhancement, Cotton Research Institute, Nanjing Agricultural University, Nanjing, China,* [2] *Henan Sesame Research Center, Henan Academy of Agricultural Sciences, Zhengzhou, China,* [3] *Key Laboratory of Oil Crops in Huanghuaihai Plain, Ministry of Agriculture, Zhengzhou, China,* [4] *Henan Provincial Key Laboratory for Oil Crops Improvement, Zhengzhou, China*

OPEN ACCESS

Edited by:
Marion S. Röder,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

Reviewed by:
Yuepeng Han,
Wuhan Botanical Garden of Chinese Academy of Sciences, China
Robert Henry,
The University of Queensland, Australia

***Correspondence:**
Haiyang Zhang
haaszhy@yahoo.com
Yongzhan Zheng
sesame168@163.com

[†] *These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Plant Genetics and Genomics, a section of the journal Frontiers in Plant Science

Received: *25 May 2017*
Accepted: *22 June 2017*
Published: *06 July 2017*

Citation:
Cui C, Mei H, Liu Y, Zhang H and Zheng Y (2017) Genetic Diversity, Population Structure, and Linkage Disequilibrium of an Association-Mapping Panel Revealed by Genome-Wide SNP Markers in Sesame. Front. Plant Sci. 8:1189. doi: 10.3389/fpls.2017.01189

The characterization of genetic diversity and population structure can be used in tandem to detect reliable phenotype–genotype associations. In the present study, we genotyped a set of 366 sesame germplasm accessions by using 89,924 single-nucleotide polymorphisms (SNPs). The number of SNPs on each chromosome was consistent with the physical length of the respective chromosome, and the average marker density was approximately 2.67 kb/SNP. The genetic diversity analysis showed that the average nucleotide diversity of the panel was 1.1×10^{-3}, with averages of 1.0×10^{-4}, 2.7×10^{-4}, and 3.6×10^{-4} obtained, respectively for three identified subgroups of the panel: Pop 1, Pop 2, and the Mixed. The genetic structure analysis revealed that these sesame germplasm accessions were structured primarily along the basis of their geographic collection, and that an extensive admixture occurred in the panel. The genome-wide linkage disequilibrium (LD) analysis showed that an average LD extended up to ∼99 kb. The genetic diversity and population structure revealed in this study should provide guidance to the future design of association studies and the systematic utilization of the genetic variation characterizing the sesame panel.

Keywords: *Sesamum indicum* L., SNPs, genetic diversity, population structure, linkage disequilibrium

INTRODUCTION

Sesame (*Sesamum indicum* L., $2n = 26$), a member of the Pedaliaceae family, is one of the most ancient oil crops and it is grown widely in both tropical and subtropical areas (Bedigian and Harlan, 1986; Ashri, 1998). Sesame seed is now widely used in food, nutraceutical, pharmaceutical, and industry in many countries. Compared with peanut (*Arachis hypogaea*), soybean (*Glycine max*), oilseed rape (*Brassica napus*), sunflower (*Helianthus annuus* L.) and other oilseed crops, sesame seeds innately contain a higher oil content (∼55% of dry seed), in which the oleic acid (18:1) (32∼46%) and linoleic acid (18:2) (42∼59%) components follow a near 1:1 ratio

(Uzun et al., 2008). In addition, sesame oil contains rich antioxidant lignans, such as sesamin and sesamolin, which are known to play important roles in food processing and human healthcare (Namiki, 1995; Anilakumar et al., 2010).

The cultivation history of sesame dates back to between 5,000 and 5,500 years ago in the Harappa Valley of the Indian subcontinent (Bedigian and Harlan, 1986). Under long-term natural and artificial selection, coupled to a wide geographic distribution, many diverse variations have accumulated in sesame, which could serve as important resources for genetic investigations and crop breeding objectives (Wei et al., 2015). In the past 19 years, great effort has been devoted to collecting and preserving sesame in India, China, and Korea (Zhang H. et al., 2012). Bisht et al. (1998) reported that 6,658 accessions were conserved in the Indian collection alone. In South Korea, the number of germplasm accessions has reached 7,698 (Park et al., 2015). In China, two sets of sesame germplasm collections are preserved by the Oil Crops Research Institute, Chinese Academy of Agricultural Sciences (OCRI-CAAS), and by the Henan Sesame Research Center, Henan Academy of Agricultural Sciences (HSRC-HAAS). This OCRI-CAAS collection contains 4,251 accessions, most of which were sent to the National Genebank for long-term storage, and a core collection of 453 representative samples that was established in 2000 (Zhang et al., 2000). With support from the China Agriculture Research System (CARS-15), there has been an extensive collecting effort conducted in the past 7 years: to date more than 5,200 accessions are now preserved in the HSRC-HAAS collection, and a core collection of 501 representative samples was amassed by Liu et al. (2017).

Knowledge of the genetic diversity and relationships among germplasm accessions is of vital importance for improving the sesame varieties. Molecular markers reflect the actual level of genetic variation existing among genotypes at the DNA level; hence, they provide a more accurate estimate of such variation than do either phenotypic or pedigree information (Pham et al., 2009). In sesame, the DNA markers such as amplified fragment length polymorphism (AFLP) (Laurentin and Karlovsky, 2006), sequence-related amplified polymorphisms (SRAP) (Zhang et al., 2010; Zhang H. et al., 2012), and inter-simple sequence repeat (ISSR) (Kumar et al., 2012) have been used for the analysis of germplasm genetic diversity and in cultivar fingerprinting. Recently, the rapid development and application of sequence-specific markers such as genomic simple sequence repeats (SSR; Cho et al., 2011), expressed sequence tag (EST)-SSR (Wei et al., 2008; Zhang Y. et al., 2012), insertions and deletions (InDels) (Wu et al., 2014) were also reported for sesame. However, a limited number of selected markers used in these studies might provide biased estimates of genetic variability (Queirós et al., 2015).

Single-nucleotide polymorphisms (SNPs) based on next generation sequencing (NGS) are more useful than conventional markers. This is because SNPs are extremely abundant, and the number of unbiased SNPs can nowadays easily been obtained for non-model species (Fischer et al., 2017). With the advent of NGS technologies and a substantial reduction in the costs of sequencing, SNPs are being discovered and genotyped in a high-throughput way (Davey et al., 2011). The specific-locus amplified fragment sequencing (SLAF-seq) method, which combines NGS with restriction enzyme digestions to reduce the complexity of the target genomes, has several positive characteristics: namely, a low cost, high efficiency, and enhanced accuracy in genome-wide marker development and genotyping (Sun et al., 2013). Not surprisingly, the SLAF-seq approach is increasingly being used in high-density genetic map constructions and genome-wide association studies (GWAS; Qi et al., 2014; Zhao et al., 2015; Zhen et al., 2016; Mei et al., 2017).

In this study, 366 elite germplasm accessions were selected from the HSRC-HAAS primary core collection and genotyped using the SLAF-seq method to evaluate their usability as an association-mapping panel. Our objectives were as follows: (a) to assess the genetic diversity as represented by the 366 sesame germplasm accessions; (b) to calculate the characteristics of the population structure; and (c) to estimate the linkage disequilibrium (LD) patterns occurring in the sample of sesame germplasm.

MATERIALS AND METHODS

Sesame Plant Material

To assemble the association-mapping panel consisting of 366 germplasm accessions (taken from the HSRC-HAAS collection), we used 329 accessions coming from 18 provinces in China, plus another 37 accessions that were introduced from 11 countries (Supplementary Table 1). All of these accessions are able to flower and ripen under natural conditions in temperate regions, which would allow us to precisely evaluate their traits in field trials. Their wide geographic distribution and phenotype variation makes these germplasms an ideal model for exploring the genetic diversity of sesame.

DNA Extractions and SLAF-seq

Genomic DNA was extracted from the young leaves of each accession by using the CTAB method (Paterson et al., 1999) with slight modifications to the CTAB buffer components to eliminate any ultra-plentiful polysaccharides in the sesame leaves (Mei et al., 2017). Crude DNA samples were purified using a DNeasy Kit (Qiagen, Valencia, United States), then assessed by electrophoresis on 0.8% agarose gel and quantified using spectrophotometry (NanoDrop 8000, Thermo Scientific, United States). The SLAF libraries were constructed following the procedure described by Sun et al. (2013), except that here two restriction enzymes, *Hae*III [recognition site 5′-GG/CC-3′, New England Biolabs (NEB), United States) and *Hpy*166II (5′-GTN/NAC-3′, NEB), were used to digest the genomic DNA. Pooled samples were separated by 2% agarose gel electrophoresis, and those fragments ranging from 264 to 364 base pairs (with indexes and adaptors) in size were excised and purified using a QIAquick gel extraction kit (Qiagen, Hilden, Germany). The gel-purified products were diluted and subjected to pair-end sequencing on an Illumina HiSeq 2000 platform (Illumina Inc., San Diego, United States) at Beijing Biomarker Technologies Corporation[1].

SLAF-tag Grouping and SNP Calling

The high-quality paired-end reads were aligned with an improved Zhongzi No. 13 genome assembly (Cui et al., submitted) coupled to Burrows–Wheeler Aligner (BWA) software (Li and Durbin, 2009) that was set to the default parameters. The SLAF-tag groups were generated by reads that were mapped onto the same position. The SNP detection was performed using a Genome Analysis Toolkit (GATK, McKenna et al., 2010) and SAMtools (Li, 2011). The process went as follows (Zhou et al., 2015): (1) after the BWA alignment, the reads around the InDels were realigned by the GATK; (2) the SNPs were called at a population level with the GATK and SAMtools. For GATK, the SNP confidence score was set as >30, and the parameter "-stand_call_conf" to a value of 30. The same realigned BAM files were used in the SNP calls with the SAMtools "mpileup" package; (3) in the filter step, the common sites identified by the GATK and SAMtools were chosen using the SelectVariants package.

Data Analysis

To reduce the influence of a strong LD on the assessment of population stratification, a subset of 12,178 SNPs were selected by setting an LD (r^2) threshold to 0.1 for the SNP pairs in a sliding window of 50 SNPs—by using PLINK v.1.07 (Purcell et al., 2007). The parameter was: "plink –file data –indep-pairwise 50 10 0.1." Population structure was estimated with 12,178 SNPs by using a Bayesian model-based program implemented in STRUCTURE 2.3.4 (Pritchard et al., 2000). Three runs were performed for each number of populations (K) set from 1 to 10. The burn-in time and the Bayesian Markov Chain Monte Carlo (MCMC) replication number were both set to 100,000 for each run. The most likely K value was determined by the log likelihood of the data [LnP(D)] and an *ad hoc* statistic, ΔK, in the program Structure Harvester (Evanno et al., 2005). Lines with a probability of membership $\geq$70% were assigned to a subgroup, whereas those with $<$70% were assigned to a "Mixed" subgroup (Liu et al., 2003).

A principal components analysis (PCA) of the selected SNPs was performed with the EIGENSOFT software (Price et al., 2006). To construct the neighbor-joining tree, the PHYLIP software was used (Plotree and Plotgram, 1989). An analysis of molecular variance (AMOVA) and the population pair-wise F statistics (F_{ST}) were calculated by using the software Arlequin v.3.52 (Excoffier and Lischer, 2010). According to the standard described by Del Carpio et al. (2011), there is no differentiation between the subpopulations when $F_{ST} = 0$, but complete differentiation occurs between the subpopulations when $F_{ST} = 1$. Populations were considered to have little differentiation when $F_{ST} \leq 0.05$, moderate differentiation when $0.05 < F_{ST} \leq 0.15$, strong differentiation when $0.15 < F_{ST} \leq 0.25$, and very strong differentiation when $F_{ST} > 0.25$ (Hartl, 1980; Mohammadi and Prasanna, 2003).

Genome-wide LD was estimated in the total panel and for each subgroup (as determined by the population structure) by pairwise comparisons among the 44,109 SNP markers [missing rates <0.30 and minor allele frequency (MAF) ≥ 0.05] using r^2. For all pairs of SNPs, the r^2 was calculated using the PopLDdecay v.3.26 program[2].

RESULTS

SLAF-Seq Genotyping

A total of 81.2 Gb of sequence data, including 902.36 million pair-end reads, were generated by the SLAF-seq when applied to the 366 sesame germplasm accessions. The Q30 ratio and guanine-cytosine (GC) content were 84.82% and 39.47%, respectively (Supplementary Table 2). High quality reads were aligned to the improved Zhongzi No. 13 genome assembly, and a total of 202,603 SLAFs evenly distributed across the whole genome were identified. SLAFs that were used for calling the SNPs had an average depth of 15.32-folds per individual. A total of 722,824 SNPs were initially called for these accessions. After removing those nucleotide polymorphisms that had missing rates >0.30, a set of 138,029 SNPs was generated. The allele frequency for each SNP site was then calculated: the MAF of the SNPs varied from 0.30 to 49.86%, with an average of 7.76%, and ~68.04% of the SNPs had a low frequency (MAF < 0.05) across the 366 accessions (Supplementary Figure S1). After excluding the SNPs with a MAF < 0.01, there were left 89,924 (~65.15%) high-quality SNPs (Supplementary Table 3) evenly distributed across the whole genome that could be used for further analysis.

The 89,924 high-quality SNPs covered all 13 linkage groups (LGs) (**Figure 1** and **Table 1**). The largest number of SNPs was found on LG5 (12,491 SNPs) followed by LG3 (11,351 SNPs), whereas the smallest number of SNPs occurred on LG4 (4,028 SNPs). The number of SNPs on each chromosome was consistent with the physical length of the respective chromosome. The average marker density was approximately 2.67 kb/SNP. LG6 had the lowest SNP marker density (3.96 kb/SNP), and LG3 had the highest marker density (1.70 kb/SNP).

Population Structure of the Association-Mapping Panel

A subset of 12,178 SNPs was selected for analysis of the population structure. The hierarchical population structure was determined for the entire panel via the model-based STRUCTURE program, but without providing any information per se on the population structure. As K changed from 1 to 10, the log likelihood value [LnP(D)] increased continuously and an inflection was evident when K increased numerically from 1 to 2 (**Figure 2A**). Thus, the most likely numerical value of K was 2. The number of subgroups (K) was further validated by the second-order statistics of ΔK. The ΔK value showed a peak at $K = 2$ (**Figure 2B**), which supported the classification of the panel into two major subgroups (**Figure 2C**).

[1] http://Biomarker.com.cn/

[2] https://github.com/BGI-shenzhen/PopLDdecay

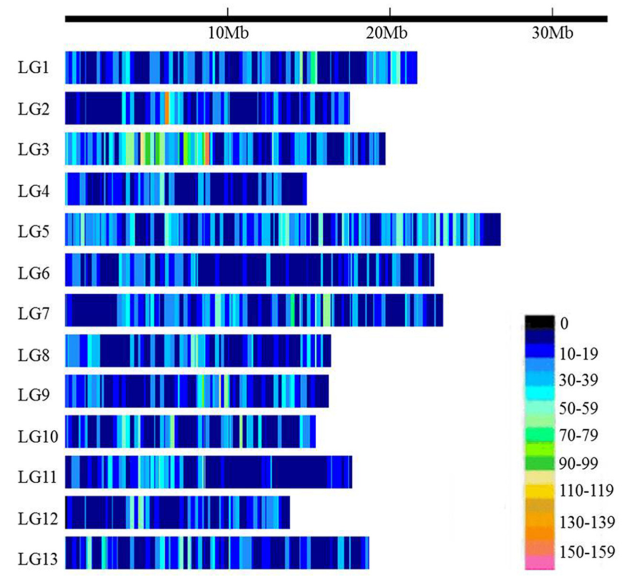

FIGURE 1 | Single-nucleotide polymorphism (SNP) distributions on the 13 linkage groups (LGs) of sesame. The horizontal axis shows the LG length; the 0∼159 legend insert depicts the SNP density (the number of SNPs per 50 kb window).

TABLE 1 | Summary of the number of SNPs mapped onto each linkage group (LG) and the gene diversity and LD decay estimated for each LG.

LG	Number of SNPs	Density of SNP (kb/SNP)	Gene diversity	LD decay (kb)
LG1	8,416	2.53	0.13	75
LG2	6,193	2.76	0.14	97
LG3	11,351	1.70	0.18	84
LG4	4,028	3.59	0.20	104
LG5	12,491	2.11	0.16	66
LG6	5,637	3.96	0.15	255
LG7	9,202	2.49	0.17	107
LG8	5,180	3.09	0.18	105
LG9	5,788	2.74	0.25	85
LG10	6,244	2.41	0.13	82
LG11	5,451	3.18	0.18	67
LG12	4,355	3.10	0.17	82
LG13	5,606	3.29	0.18	106
Total	89,924	2.67	0.17	99

When using a probability of membership threshold of 70%, 144 and 111 accessions were respectively assigned into the two subgroups, Pop 1 and Pop 2, while the remaining 111 accessions were classified into a mixed subgroup (Mixed) (Supplementary Table 1). Most accessions of Pop 1 came from the Southern areas in China, while 14 accessions in total came from Bangladesh ($n = 2$), Japan (1), India (2), Burma (2), Thailand (1), Arab Emirates (2), Mexico (1), Mozambique (2), and Guinea (1). The remainder of Pop 1 comprised 12 provinces in China (eight Southern provinces and four Northern provinces). The accessions of Pop 2 were mainly collected from the Northern areas in China; specifically, just three accessions came from South Korea, Japan, and United States, whereas 108 accessions came from nine provinces in China (eight Northern provinces and one Southern province). For the Mixed group, its accessions were collected from both the Southern (23 from five provinces) and Northern areas (68 from five provinces) in China, along with 20 accessions coming from Japan (12), Burma (3), Mozambique (1), Arab Emirates (1), United States (2), and Thailand (1).

The PCA was done to further assess the population subdivisions. PC1 explained 10.10% of the genetic variation found, while PC2 and PC3 explained 6.77 and 4.40% of the variation, respectively. There was a significant correlation between PC1 and distributional latitudes of the accessions (Pearson's correlation test, $r^2 = 0.37$, $P < 0.0001$). Therefore, the sesame panel along the PC1 axis could be separated into Type 1 (a Southern-area subgroup) and Type 2 (a Northern-area subgroup) subpopulations (**Figure 3A** and Supplementary Table 1), which corresponded to those accessions collected from the Southern and Northern areas in China, respectively. However, some intermediate lines made the grouping less than clear-cut. When considering these intermediate lines, the panel could be neatly divided into three clusters (**Figure 3B**) corresponding to the three subgroups (**Figure 2C**) as inferred by using STRUCTURE.

The neighbor-joining phylogenetic tree was built to search for genetic relationships among the sesame accessions in the panel. This panel was grouped into two recognizable clusters—blue vs. green/red clusters shown in **Figure 4**—of which the blue cluster tended to be from the Northern areas of China. Similarly, the accessions from the Southern areas tended to be clustered together (red, **Figure 4**) while those accessions belonging to the Mixed group were distributed across the whole phylogenetic tree (green, **Figure 4**).

The genome-wide data set of 89,924 SNP markers with an MAF $\geq$ 0.01 was used to estimate the genetic differentiation for the panel. The AMOVA results indicated that 14.92% of the total genetic variation occurred among the subgroups. While a larger amount of variation (63.06%) was among individuals within the populations, 22.03% of variation was found within the individuals. The measure of population differentiation, F_{ST}, among the subgroups was 0.15 highly significant at $P < 0.0001$. The corresponding F_{ST} values between the subgroups ranged from moderate, for Pop 1 vs. Mixed (0.07, $P < 0.0001$) and Pop 2 vs. Mixed (0.11, $P < 0.0001$), to very strong (0.26, $P < 0.0001$) for that of Pop 1 vs. Pop 2.

Genetic Diversity Revealed by SNP Markers

A total of 89,924 SNPs were used to study the genetic diversity of the sesame panel. The average gene diversity of the panel was 0.17; the highest was 0.25 on LG9, followed by LG4 (0.20), with the lowest value found (0.13) on LG1 and LG10 (**Table 1**). The ranges for the estimates of gene diversity in Pop1, Pop2, and the Mixed were 0–0.66, 0–0.66, and 0–0.67, respectively, and their corresponding averages were 0.17, 0.11, and 0.17. On the basis of these SNP data, the sequence diversity (π) was estimated as 1.1×10^{-3} for all of the sesame accessions, and as 1.0×10^{-4},

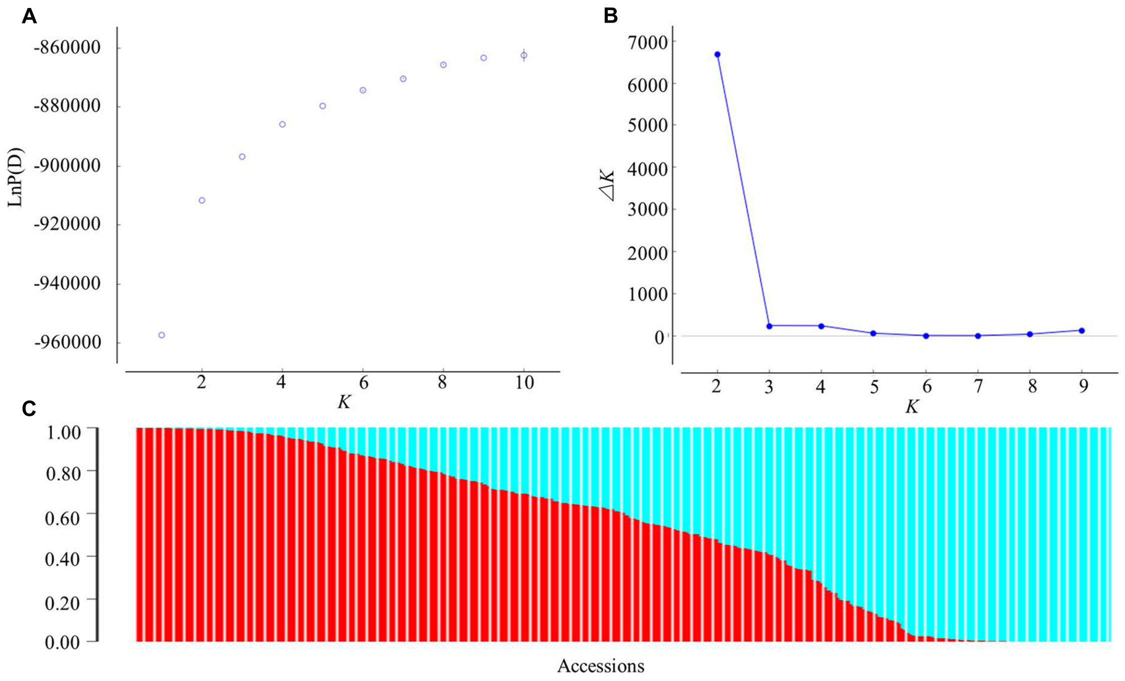

FIGURE 2 | Analysis of the population structure of the 366 sesame accessions using STRUCTURE. **(A)** Estimated LnP(D) of possible clusters (*K*) from 1 to 10; **(B)** Δ*K* based on the rate of change of LnP(D) between successive *K*; **(C)** population structure based on *K* = 2. In panel **(C)**, each individual is represented by a vertical bar partitioned into two colored segments, with their respective lengths representing the proportion of the individual's genome in a given group: red, for Pop 1; blue, for Pop 2.

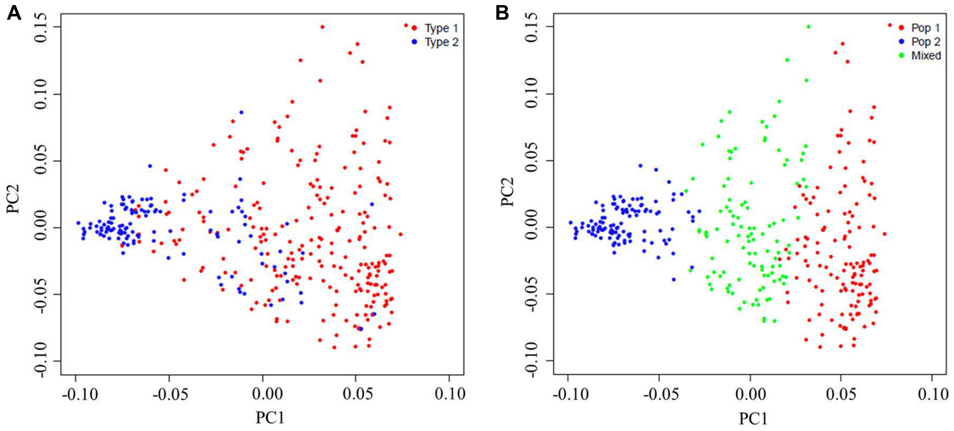

FIGURE 3 | Principal components analysis (PCA) of the population for 366 sesame accessions based on 12,178 single-nucleotide polymorphisms (SNPs). Each individual is represented by one dot, with its symbol color corresponding to the assigned subgroup classification. **(A)** PCA plots of the sesame germplasm collection based upon their geographic origins: Type 1, the Southern-area subgroup, and Type 2, the Northern-area subgroup. **(B)** PCA plots of the same sesame germplasm collection but now based on subgroups as identified by STRUCTURE.

2.7×10^{-4}, and 3.6×10^{-4} for Pop 1, Pop 2, and the Mixed group, respectively.

To reveal the genetic differences among the different groups of the sesame germplasm, a comparative analysis of their allele frequencies was performed. The Pop 1 accessions had the largest number of group-specific SNPs (n = 3,720 SNPs), followed by Pop 2 (290), while the Mixed accessions had the smallest number of group-specific SNPs (157). In the pairwise comparisons of Pop 1, Pop 2, and the Mixed group, the number of SNPs unique to the Pop 1 (31,203) vastly exceeded those unique to the Pop 2 (2,195), though the number of group-specific SNPs (4,833) in Pop1 was just a little more than twice that in the Mixed group (2,352), whereas the number of SNPs unique to the Mixed group (27,640) was far greater than those unique to the Pop 2 (1,403).

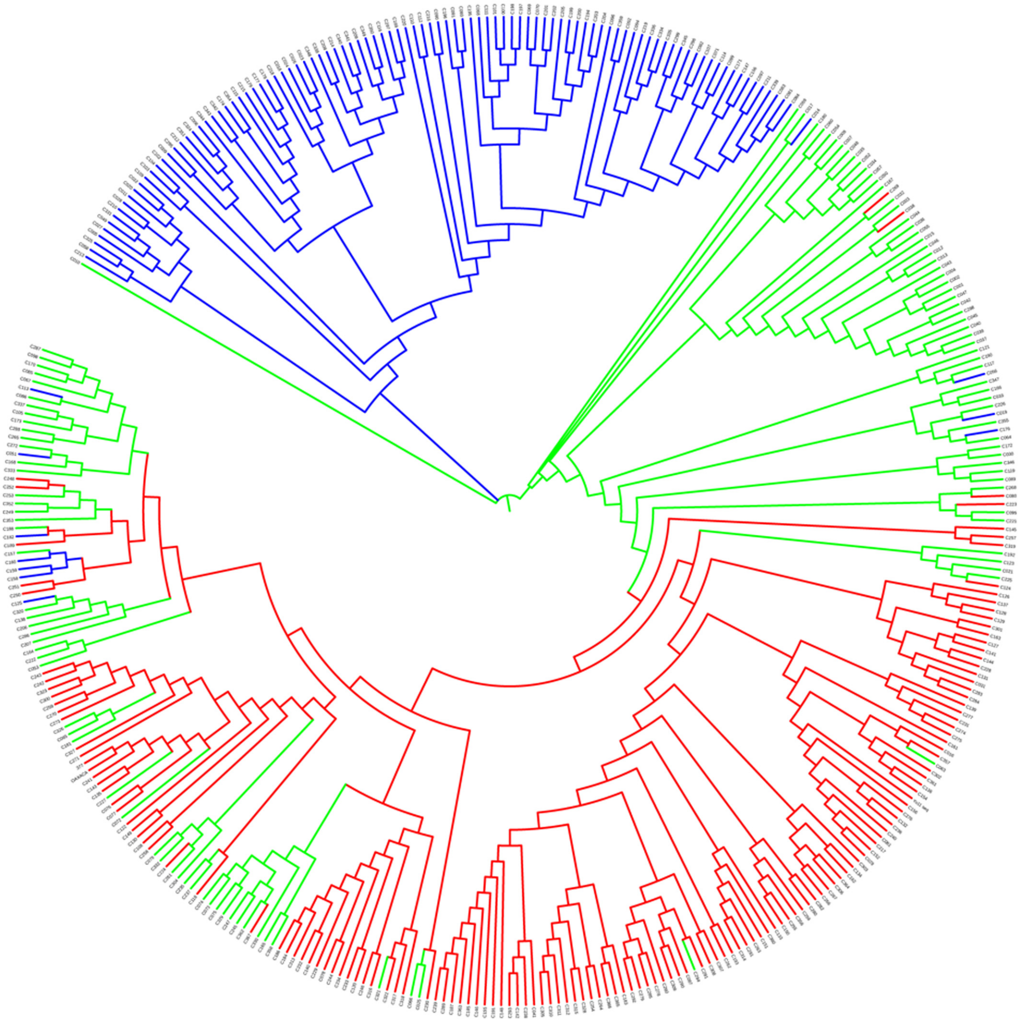

FIGURE 4 | Neighbor-joining phylogenetic tree of the 366 sesame accessions. Red indicates Pop 1, blue indicates Pop 2, and green indicates the Mixed group.

The low level of heterozygous genotypes (heterozygosity = 0.05) was consistent with the inbreeding nature of sesame. The average heterozygosity was higher in Pop 2 (= 0.05) than in Pop 1 (= 0.04). The genetic distance among the 366 sesame accessions averaged 0.17, with a range of 0.01–0.42. Nonetheless, Pop 2, which included accessions collected from the Northern areas in China, had the lowest average genetic distance, at 0.11, as well as the smallest range, from 0.01 (C020 and C022) to 0.16 (C027 and C125). By contrast, Pop 1 and the Mixed group had a similar average genetic distance (at 0.17 and 0.16, respectively), which ranged from 0.01 (C150 and C255) to 0.40 (C278 and 377), and from 0.05 (C245 and C247) to 0.30 (C076 and C181), respectively.

Linkage Disequilibrium across Whole Sesame Genome

In this panel, the genome-wide LD decay at which r^2 decreased to 0.37 from the initial value of 0.74 was ~99 kb, such that the r^2 dropped to 0.2 at ~234 kb and to 0.1 at ~467 kb (**Figure 5**). To obtain more details of this LD behavior, the chromosome-wise LD between the SNP pairs was also calculated. The highest maximum average LD ($r^2 < 0.80$) was found for LG1, and the lowest maximum average LD ($r^2 = 0.66$) was found for LG7. Except for the latter, all of the other LGs showed a maximum average LD that exceeded 0.70. The highest LD decay was ~255 kb for LG6, followed by that seen for LG7 (~107 kb), whereas the lowest LD decay (~66 kb) was observed for LG5 (Figure S2). When the population-wise LD was calculated, the maximum average LD ($r^2 = 0.75$) occurred in Pop 2, for which the LD decay was ~83 kb. However, the minimum average LD was found to be very low (0.04) for Pop 2. The maximum average LD for the Mixed group was comparatively lower (0.72), with the decay to half to its initial value occurring at ~86 kb. The average minimum LD was also very low (0.04) for the Mixed subgroup. The maximum and minimum average LD values for Pop 1 were 0.74 and 0.06, respectively. Compared with the LD decays of the two other subgroups, Pop 1 had the highest rate of LD decay, at ~100 kb (**Figure 5**).

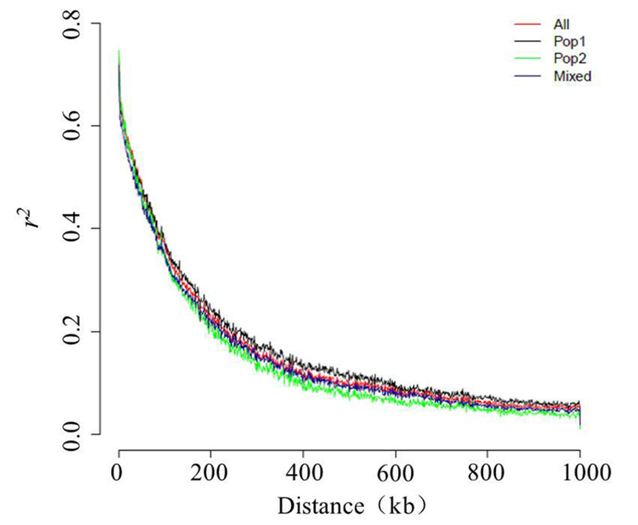

FIGURE 5 | Genome-wide decay of the linkage disequilibrium (LD) in the sesame genome.

DISCUSSION

Genetic Diversity Revealed by SNP Markers

The genetic diversity housed in germplasm collections is an important foundation for crop improvement and a key component of plant conservation and breeding strategies (Thomson et al., 2007). For sesame, its genetic diversity has been assessed in several previous studies. For example, Zhang et al. (2010) reported an average gene diversity of 0.24 in 404 accessions, as based on SRAPs and SSRs, while Cho et al. (2011) found a lower gene diversity, of 0.17, in a sesame panel based solely on SSRs. In general, the level of genetic diversity estimated by a limited number of SNPs tends to be lower than that estimated through the use of SSR markers (Jones et al., 2007; Van Inghelandt et al., 2010). However, in considering the criteria for genetic diversity, more weight should have been given to the number of loci instead of the number of alleles (Lu et al., 2009). So long as a sufficiently large number of unbiased NGS-based SNPs are analyzed across the genome, the SNP estimates will accurately reflect the genome-wide diversity occurring in natural populations (Fischer et al., 2017).

In our study, the 89,924 SNPs identified by SLAF-seq were evenly distributed across the whole genome of sesame. Compared to the SSR markers used in prior studies, these SNPs derived via SLAF-seq apparently could provide a better coverage of the genome and unbiased estimates of its diversity. By using the number of SNPs, the average gene diversity of 0.17 was calculated for the panel. In the present study, the average nucleotide diversity was 1.1×10^{-3}, a value similar to that reported by both Wei et al. (2015) and Wang et al. (2014); this indicates that the present panel had a modest level of nucleotide diversity and so it could be considered as a representative sample to conduct GWAS. However, as cultivars derived from a single taxon (Bedigian and Harlan, 1986; Bedigian, 2010), the average nucleotide diversity (1.1×10^{-3}) of sesame was lower than that of rice (Huang et al., 2010), yet similar to that of the soybean landraces which are considered to originate from a single domestication event (Zhou et al., 2015).

The proportion of rare SNPs (i.e., MAF < 0.05) we examined amounted to ~68.04%, which was similar to those reported for the genomes of Arabidopsis and Alfalfa (Clark et al., 2007; Zhang et al., 2015). The high proportion of rare SNPs in our study may have two explanations. Firstly, since the SNPs were identified via SLAFs evenly distributed across the whole genome, they should be less prone to bias than would be low-coverage sequencing data (Huang et al., 2012). Secondly, in following its recent program to conserve genetic resources, a significant number of minor sesame varieties have been collected and preserved by HSRC-HAAS. For the LD-based mapping, markers with rare alleles are thus more likely to result in spurious findings. The SNPs with a MAF < 0.05 were removed in several previous studies (Huang et al., 2010; Yano et al., 2016). However, rare SNPs may also have an effect on the expression of a specific phenotype (Song et al., 2015). Given that the number of individuals with a specific genotype can be very small, the effect of rare alleles on genome mapping could extend beyond the effect of just small population sizes. In such cases, increasing the number of individuals with rare alleles could improve the power to test these rare alleles. Fortunately, a biparental population-based mapping approach, especially as done through the Nested Association Mapping (NAM), is able to use alleles occurring at a low frequency in natural populations by designing crosses to create artificial populations that have inflated frequencies of those alleles (Lu et al., 2010; Tian et al., 2011).

Sesame is a mostly self-pollinated plant, and hence it is likely that all of the sesame accessions in the present study had been held for many generations via self-pollination. Their genomes are thus expected to be mostly homozygous. In line with this expectation, the average heterozygosity in the sesame panel was 0.046; this suggests that the accessions we used were quite close to being inbred lines. Hence, the accessions selected from the HSRC-HAAS core collection are useful and suitable for investigating multiple phenotypic traits in a multi-plot field test over several years and to also carry out GWAS.

Population Structure of the Association-Mapping Panel

The complex breeding history of many important crops and the limited gene flow in most wild plant populations have created complex structures within their germplasms (Sharbel et al., 2000). Detailed knowledge about the population structure in an association panel is thus important to avoid any spurious associations (Flint-Garcia et al., 2005). An assessment of structure in sesame has been reported by using different populations. For instance, Ali et al. (2007) found that 96 sesame accessions, collected from different parts of the world, could be separated into just two major groups that discriminated varieties as related to their geographical origin. Recently, Wei et al. (2015) divided 705 sesame accessions into two clusters by using a neighbor-joining tree. Similarly, in our study, the *K* value of 2 was determined by both the LnP(D) and ΔK. By using a 70% probability of membership threshold, the panel was successfully divided into three subgroups (Pop 1, Pop 2, and the Mixed). Most of the accessions belonging to Pop 1 came from Southern areas whereas the accessions of Pop 2 came from Northern areas in China. The remaining 111 accessions were all collected from both Southern and Northern areas, thus showing the substantial exchange of germplasm that has occurred in sesame.

The PCA was performed to examine the population structure of the sesame panel. The significant correlation found between PC1 and the latitudinal distribution of the sesame accessions—similar to findings by Wei et al. (2015)—indicates that the sesame germplasm accessions were structured on the basis of their geographic collection. According to the PC1 axis, the panel divided into two major groups, consistent with the prior structure analysis at $K = 2$. However, the PCA was unable to capture much variance in the PC1 axis, which suggests that the sesame panel contained a number of admixed lines and a low layer of population structure.

According to the AMOVA results, 14.92% of the marker variation was explained by the population structure of the sesame panel. This result suggests the absence of a complicated population structure in our association-mapping panel. The differentiation between the subgroups was further validated by the F_{ST} value, calculated here as 0.15, which demonstrates the moderate genetic differentiation characterizing this panel. The latter finding should favor the detection of gene effects on the power of structure-based association studies.

Linkage Disequilibrium in Sesame

The decay of LD over a known genetic distance is an important parameter for determining the number and density of molecular markers deemed appropriate for GWAS and selection strategies (Mather et al., 2007). Sesame, as a predominantly self-pollinated species, albeit one that readily outcrosses if able, is expected to have a higher level of LD than that found in cross-pollinated crops. Wang et al. (2014) had reported that the LD decay was ∼150 kb in 29 sesame accessions. Later, an LD decay of ∼88 kb was reported in 705 sesame accessions (Wei et al., 2015). In the present study, the LD decay was estimated at ∼99 kb—the point at which the r^2 dropped to 0.37 from its initial value of 0.74—and much lower than that found by Wang et al. (2014). This discrepancy may reflect the effect of a small population size, as reported by Wang et al. (2014), wherein genetic drift drives a consistent loss of rare allelic combinations, thereby increasing its LD level. Because the LD decay in our study was closer to that reported by Wei et al. (2015), it suggests that our sesame panel had an adequate number of accessions and a rich genetic diversity for GWAS.

CONCLUSION

The sesame accessions used in this study were selected from the HSRC-HAAS core collection and so they mainly encompassed landraces and cultivars. All the accessions displayed good adaptation to most of the current growing conditions in China; this means that field experiments could be robustly conducted to test multiple phenotypic traits in different places throughout China. With the aim of evaluating the potential of a panel for an association analysis, we studied the population diversity and structure of a sesame panel consisting of 366 accessions. Our results show that this sesame panel is a representative sample, and one therefore suitable for further association mapping, given the modest level of nucleotide diversity and the slight population stratification that favors GWAS and a false association control.

AUTHOR CONTRIBUTIONS

HM and YL developed the association-mapping panel; CC and HM performed the data analysis; CC wrote the manuscript, HM revised the manuscript; HZ and YZ designed and supervised the study. All authors read and approved the final manuscript.

FUNDING

This study was supported by the China Agriculture Research System (CARS-15), the Henan Fundamental and Frontier Research Fund (162300410164), and the Independent Innovation Project (2016ZC69) from the Henan Academy of Agricultural Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.01189/full#supplementary-material

REFERENCES

Ali, G. M., Yasumoto, S., and Seki-Katsuta, M. (2007). Assessment of genetic diversity in sesame (*Sesamum indicum* L.) detected by Amplified Fragment Length Polymorphism markers. *Electron. J. Biotechnol.* 10, 12–23. doi: 10.1186/1471-2164-9-250

Anilakumar, K. R., Pal, A., Khanum, F., and Bawa, A. S. (2010). Nutritional, medicinal and industrial uses of sesame (*Sesamum indicum* L.) seeds – an overview. *Agric. Conspec. Sci.* 75, 159–168.

Ashri, A. (1998). "Sesame breeding," in *Plant Breeding Reviews*, ed. J. Janick (New York, NY: John Wiley & Sons, Inc.), 179–228.

Bedigian, D. (2010). Characterization of sesame (*Sesamum indicum* L.) germplasm: a critique. *Genet. Resour. Crop Evol.* 57, 641–647. doi: 10.1007/s10722-010-9552-x

Bedigian, D., and Harlan, R. (1986). Evidence for cultivation on sesame in the ancient world. *Econ. Bot.* 40, 137–154. doi: 10.1007/BF02859136

Bisht, I. S., Mahajan, R. K., Loknathan, T. R., and Agrawal, R. C. (1998). Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. *Genet. Resour. Crop Evol.* 45, 325–335. doi: 10.1023/A:1008652420477

Cho, Y. I., Park, J. H., Lee, C. W., Ra, W. H., Chung, J. W., Lee, J. R., et al. (2011). Evaluation of the genetic diversity and population structure of sesame (*Sesamum indicum* L.) using microsatellite markers. *Genes Genomics* 33, 187–195. doi: 10.1007/s13258-010-0130-6

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., et al. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317, 338–342. doi: 10.1126/science.1138632

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

Del Carpio, D. P., Basnet, R. K., De Vos, R. C. H., Maliepaard, C., Visser, R., and Bonnema, G. (2011). The patterns of population differentiation in a *Brassica rapa* core collection. *Theor. Appl. Genet.* 122, 1105–1118. doi: 10.1007/s00122-010-1516-1

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., et al. (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics* 18:69. doi: 10.1186/s12864-016-3459-7

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064. doi: 10.1111/j.1365-313X.2005.02591.x

Hartl, D. L. (1980). *Principles of Population Genetics*, 1st Edn. Sunderland: Sinauer Associates.

Huang, X. H., Wei, X. H., Sang, T., Zhao, Q. A., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–976. doi: 10.1038/ng.695

Huang, X. H., Zhao, Y., Wei, X. H., Li, C. Y., Wang, A. H., Zhao, Q., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44, 32–39. doi: 10.1038/ng.1018

Jones, E. S., Sullivan, H., Bhattramakki, D., and Smith, J. S. C. (2007). A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.). *Theor. Appl. Genet.* 115, 361–371. doi: 10.1007/s00122-007-0570-9

Kumar, H., Kaur, G., and Banga, S. (2012). Molecular characterization and assessment of genetic diversity in sesame (*Sesamum indicum* L.) germplasm collection using ISSR markers. *J. Crop Improv.* 26, 540–557. doi: 10.1080/15427528.2012.66056

Laurentin, H. E., and 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. doi: 10.1186/1471-2156-7-10

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Liu, K., Goodman, M., Muse, S., Smith, J. S., Buckler, E., and Doebley, J. (2003). Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165, 2117–2128.

Liu, Y. Y., Mei, H. X., Du, Z. W., Wu, K., Zheng, Y. Z., Cui, X. H., et al. (2017). Constructing core collection of sesame based on phenotype and molecular markers. *Sci. Agric. Sin.* 50.

Lu, Y. L., Yan, J. B., Guimarães, C. T., Taba, S., Hao, Z. F., Gao, S. B., et al. (2009). Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor. Appl. Genet.* 120, 93–115. doi: 10.1007/s00122-009-1162-7

Lu, Y. L., Zhang, S. H., Shah, T., Xie, C. X., Hao, Z. F., Li, X. H., et al. (2010). Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19585–19590. doi: 10.1073/pnas.1006105107

Mather, K. A., Caicedo, A. L., Polato, N. R., Olsen, K. M., McCouch, S., and Purugganan, M. D. (2007). The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* 177, 2223–2232. doi: 10.1534/genetics.107.079616

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mei, H., Liu, Y., Du, Z., Wu, K., Cui, C., Jiang, X., et al. (2017). High-density genetic map construction and gene mapping of basal branching habit and flowers per leaf axil in sesame. *Front. Plant Sci.* 8:636. doi: 10.3389/fpls.2017.00636

Mohammadi, S. A., and Prasanna, B. M. (2003). Analysis of genetic diversity in crop plants-salient statistical tools and considerations. *Crop Sci.* 43, 1235–1248. doi: 10.2135/cropsci2003.1235

Namiki, M. (1995). The chemistry and physiological functions of sesame. *Food Rev. Int.* 11, 281–329. doi: 10.1080/87559129509541043

Park, J. H., Suresh, S., Raveendar, S., Baek, H. J., Kim, C. K., Lee, S., et al. (2015). Development and evaluation of core collection using qualitative and quantitative trait descriptor in sesame (*Sesamum indicum* L.) germplasm. *Korean J. Crop Sci.* 60, 75–84. doi: 10.7740/kjcs.2014.60.1.075

Paterson, A. H., Brubaker, C., and Wendel, J. F. (1999). A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol.* 11, 122–127. doi: 10.1007/BF02670470

Pham, T. D., Bui, T. M., Werlemark, G., Bui, T. C., Merker, A., and Carlsson, A. S. (2009). A study of genetic diversity of sesame (*Sesamum indicum* L.) in Vietnam and Cambodia estimated by RAPD markers. *Genet. Resour. Crop Evol.* 56, 679–690. doi: 10.1007/s10722-008-9393-z

Plotree, D., and Plotgram, D. (1989). PHYLIP-Phylogeny inference package (version 3.2). *Cladistics* 5, 163–166.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Qi, Z., Huang, L., Zhu, R., Xin, D., Liu, C., Han, X., et al. (2014). A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS ONE* 9:e104871. doi: 10.1371/journal.pone.0104871

Queirós, J., Godinho, R., Lopes, S., Gortazar, C., de la Fuente, J., and Alves, P. C. (2015). Effect of microsatellite selection on individual and population genetic inferences: an empirical study using cross-specific and species-specific amplifications. *Mol. Ecol. Resour.* 15, 747–760. doi: 10.1111/1755-0998.12349

Sharbel, T. F., Haubold, B., and Mitchellolds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* 9, 2109–2118. doi: 10.1046/j.1365-294X.2000.01122.x

Song, X. J., Kuroha, T., Ayano, M., Furuta, T., Nagai, K., Komeda, N., et al. (2015). Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. *Proc. Natl. Acad. Sci. U.S.A.* 112, 76–81. doi: 10.1073/pnas.1421127112

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. doi: 10.1371/journal.pone.0058700

Thomson, M. J., Septiningsih, E. M., Suwardjo, F., Santoso, T. J., Silitonga, T. S., and McCouch, S. R. (2007). Genetic diversity analysis of traditional and improved Indonesian rice (*Oryza sativa* L.) germplasm using microsatellite markers. *Theor. Appl. Genet.* 114, 559–568. doi: 10.1007/s00122-006-0457-1

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, 159–162. doi: 10.1038/ng.746

Uzun, B., Arslan, C., and Furat, S. (2008). Variation in fatty acid compositions, oil content and oil yield in a germplasm collection of sesame (*Sesamum indicum* L.). *J. Am. Oil Chem. Soc.* 85, 1135–1142. doi: 10.1007/s11746-008-1304-0

Van Inghelandt, D., Melchinger, A. E., Lebreton, C., and Stich, B. (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theor. Appl. Genet.* 120, 1289–1299. doi: 10.1007/s00122-009-1256-2

Wang, L. H., Han, X. L., Zhang, Y. X., Li, D. H., Wei, X., Ding, X., et al. (2014). Deep resequencing reveals allelic variation in *Sesamum indicum*. *BMC Plant Biol.* 14:225. doi: 10.1186/s12870-014-0225-3

Wei, L. B., Zhang, H. Y., Zheng, Y. Z., Guo, W. Z., and Zhang, T. Z. (2008). Developing EST-Derived microsatellites in sesame (*Sesamum indicum* L.). *Acta Agron. Sin.* 34, 2077–2084. doi: 10.1016/S1875-2780(09)60019-5

Wei, X., Liu, K. Y., Zhang, Y. X., Feng, Q., Wang, L. H., Zhao, Y., et al. (2015). Genetic discovery for oil production and quality in sesame. *Nat. Commun.* 6:8609. doi: 10.1038/ncomms9609

Wu, K., Yang, M. M., Liu, H. Y., Tao, Y., Mei, J., and Zhao, Y. Z. (2014). Genetic analysis and molecular characterization of Chinese sesame (*Sesamum indicum* L.) cultivars using insertion-deletion (InDel) and simple sequence repeat (SSR) markers. *BMC Genet.* 15:35. doi: 10.1186/1471-2156-15-35

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48, 927–934. doi: 10.1038/ng.3596

Zhang, H., Wei, L., Miao, H., Zhang, T., and Wang, C. (2012). Development and validation of genic-SSR markers in sesame by RNA-seq. *BMC Genomics* 13:316. doi: 10.1186/1471-2164-13-316

Zhang, T., Yu, L. X., Zheng, P., Li, Y., Rivera, M., Main, D., et al. (2015). Identification of loci associated with drought resistance traits in heterozygous autotetraploid Alfalfa (*Medicago sativa* L.) using genome-wide association studies with genotyping by sequencing. *PLoS ONE* 10:e0138931. doi: 10.1371/journal.pone.0138931

Zhang, X. R., Zhao, Y. Z., Cheng, Y., Feng, X. Y., Guo, Q. Y., Zhou, M. D., et al. (2000). Establishment of sesame germplasm core collection in China. *Genet. Resour. Crop Evol.* 47, 273–279. doi: 10.1023/A:1008767307675

Zhang, Y., Zhang, X., Che, Z., Wang, L., Wei, W., and Li, D. (2012). Genetic diversity assessment of sesame core collection in China by phenotype and molecular markers and extraction of a mini-core collection. *BMC Genet.* 13:102. doi: 10.1186/1471-2156-13-102

Zhang, Y. X., Zhang, X. R., Hua, W., Wang, L. H., and Che, Z. (2010). Analysis of genetic diversity among indigenous landraces from sesame (*Sesamum indicum* L.) core collection in China as revealed by SRAP and SSR markers. *Genes Genomics* 32, 207–215. doi: 10.1007/s13258-009-0888-6

Zhao, X., Han, Y., Li, Y., Liu, D., Sun, M., Zhao, Y., et al. (2015). Loci and candidate gene identification for resistance to *Sclerotinia sclerotiorum* in soybean (*Glycine max* L. Merr.) via association and linkage maps. *Plant J.* 82, 245–255. doi: 10.1111/tpj.12810

Zhen, Z., Shang, H., Shi, Y., Long, H., Li, J., Ge, Q., et al. (2016). Construction of a high-density genetic map by specific locus amplified fragment sequencing (SLAF-seq) and its application to Quantitative Trait Loci (QTL) analysis for boll weight in upland cotton (*Gossypium hirsutum*.). *BMC Plant Biol.* 16:79. doi: 10.1186/s12870-016-0741-4

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Re-sequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.