Genetic Diversity, Population Structure Analysis Using Ultra-High Throughput Diversity array Technology (DArTseq) in Different Origin Sesame (Sesamum indicum L)

TEWODROS TESFAYE NEGASH (tesfaye.tewodros@yahoo.com)
Amhara Agricultural Research Institute https://orcid.org/0000-0001-7285-250X
KASSAHUN TESFAYE
Addis Ababa University College of Natural Sciences
GEMECHU KENENI WAKEYO
Ethiopian Institute of Agricultural Research
CATHRINE ZIYOMO
International livestock research institute

Research article

Keywords: DArTseq, silicoDArT, SNP, Genetic diversity, and Population structure

Posted Date: November 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-103763/v1

Abstract

Background

Sesame is an important oil crop widely cultivated in Africa and Asia continent. Characterization of genetic diversity and population structure of sesame genotypes in these continents can be used to designing breeding methods. In the present study, 300 sesame genotypes comprising 209 local, and 75 exotic collection, and 16 released varieties provided from the Ethiopian Biodiversity Institute and research centers were used in the present study.

Results

The panel was genotyped using two ultra-high-throughput diversity array technology (DArT) markers (silicoDArT and SNP). Both markers were used to identify the genetic diversity and population structure of sesame germplasm. A total of 6115 silicoDArT and 6474 SNP markers were reported, of which 5002 silicoDArT and 4638 SNP markers were screening with quality control parameters. The average polymorphic information content values of silicoDArT and SNP markers were 0.07 and 0.08, respectively. For further analysis, the allele frequency for each SNP site was calculated and purified with MAF < 0.01 and left 2997 high-quality SNPs evenly distributed across the whole genome that could be used for subsequent analysis. All genotypes used in this study were descended from eight 8 geographical origins. The genetic diversity analysis showed that the average nucleotide diversity of the panel was 0.14. Considering the genotypes based on their geographical origin, Africa collections (0.21) as a whole without Ethiopian collection was more diverse than Asia and when further portioned Africa, North Africa (0.23) collection was more diverse than others, but at the continent level, Asia (0.17) was more diverse than Africa (0.14). The genetic distance among the sesame populations was ranged from 0.015 to 0.394, with an average of 0.165. The sesame populations was clustered into four groups. The structure analysis divided the panel into four subgroups and 21 genotypes were clustered as an admixture. These indicates genotypes from the same origin didn’t classify properly on the premise of the country of origin.

Conclusions

The genetic diversity and population structure revealed in this study should guide the future research work to design association studies and the systematic utilization of the genetic variation characterizing the sesame panel.

Background

Sesame (*Sesamum indicum* L., 2n = 26), a member of the Pedaliaceae family, is one of the most

ancient oil crops are grown widely in both tropical and subtropical areas since the time immemorial [1, 2]. Archeological findings revealed that the cultivated sesame traces back its progenitor to the wild populations native to South Asia [3]. However, there's still some controversy as to the center of domestication. Some claimed that sesame has been cultivated in South Asia since the time of the

Harappan civilization from where it was spread west to the Mesopotamia before 2000 B.C. [3]. Others believed that the crop was first cultivated in Africa and later taken to India at a very early date [4, 5]. Still, others proposed that sesame was the main oil crop grown by the Indus Valley Civilization from where it was likely transferred to the Mesopotamia around 2500 B.C [6].

Sesame is produced in numerous parts of the world for various purposes but more than 96% of the world sesame seed production is covered by Africa and Asia, India, China, Burma (Myanmar), Sudan, Nigeria, The United Republic of Tanzania, Ethiopia, and Uganda being the key contributors [7]. Sesame seeds are good sources of fat, protein, carbohydrates, fiber, and essential minerals. Seeds are chemically composed of 44–57% oil, 18–25% protein, and 13–14% carbohydrates [8]. Sesame, also referred to as “queen of oilseeds”, is employed in sweets such as sesame bars and halva (dessert), and bakery products or milled to get high-grade edible oil [9].

Despite the nutritional and economic importance in several parts of the world, however, little focus is given to sesame research be it at the national or the international levels [1, 10–12].

In Ethiopia, sesame is among the foremost important oil crops both in terms of area coverage and total national annual production [13]. However, the farm level productivity of sesame in Ethiopia isn’t only far below the genetic potential of the crop yield of 2 t ha^{-1} [14] but also when compared with productivity in other countries like Egypt (1.29 t ha^{-1}), Nigeria (1.1 t ha^{-1}) Tanzania (1 t ha^{-1}), and chain (1.4 t ha^{-1}) [15]. Improved varieties released in Ethiopia are reported to yields ranging 0.3 to 1.3 t ha^{-1} under rain fed and 1 to 2.4 t ha^{-1} under irrigation on research stations [16] and 0.4–1.3 t ha^{-1} on farmers’ fields as reported by Ministry of Agriculture and Rural Development (MoARD) relesed book from 2010 to 2017 [17]. Nevertheless, the national average yield is low (0.68t ha^{-1}) [18]. On the other hand, Ethiopia is also considered as one of the centers of genetic diversity an immense wealth of genetic diversity in the germplasm collections for potential exploitation through genetic improvement in future breeding [19]. The effective utilization in breeding programs of this germplasm, by and large, depends, among others, on systematic genetic characterization to unveil the magnitude and pattern of genetic diversity available in the germplasm, enables the determination of useful genes and the possible progress that can be made through future breeding. Despite the huge amount of both locally collected and introduced germplasm held in the Ethiopian gene bank and in breeders’ stock, information on the use of molecular markers for the characterization of genetic diversity is limited [16]. It is well established since the time of N.I. Vavilov [20] that Ethiopian sesame landraces have valuable genetic diversity at the morphological level [21–25]. The studies conducted in the past at the molecular level, the eco-geographic distribution and microcenters of the diversity have several limitations as they are based on a few old markers and/or a limited number of genotypes [25–27].

In sesame, different molecular marker systems such as amplified fragment length polymorphism [28], sequence-related amplified polymorphisms [29, 30], inter-simple sequence repeat [31], simple sequence repeats [32], expressed sequence tag [30, 33] and insertions and deletions [34] have been used elsewhere for the analysis of germplasm genetic diversity.

More recently high-thorough put marker systems particularly single-nucleotide polymorphisms and Diversity Arrays Technology (DArT) markers have become the genetic markers of choice for genetic analyses including characterization of germplasm because of several comparative advantages like abundant in the genome, efficiency, low cost, and speed [35–39].

Over the last decade, DArT has generated two types of markers, namely silicoDArT and SNP markers. SilicoDArT markers are dominant microarray markers and scored for the presence or absence of a single allele, whereas DArTseq based SNPs are co-dominant markers, both of them being successfully applied in genetic diversity [40–44] and population structure [45, 46] study of several crop species. The present study, which is the first of its kind to utilize the DArT platforms in sesame, was designed to determine the magnitude and pattern of genetic diversity in a collection of 300 sesame germplasm thereby generate information on the eco-geographic distribution and microcenter of genetic diversity in Ethiopia.

Results

Marker discovery by DArTseq and quality analysis

Through the application of the complexity reduction method, a total of 6115 polymorphic silicoDArT (Table 1) markers were generated of which 5065 were aligned with the Reference sesame genome obtained from (NCBI) and 326 were scaffold, and 724 were unknown markers. Based on updating genome assembly and annotation available at http://ocri-genomics.org/Sinbase_v2.0 (genome assembly) and http://ocri-genomics.org/Sin_SNP_430RIL.tar.gz (SNP information). 5065 silicoDArT markers were distributed on all 13 chromosomes of sesame (Fig. 1 **and** Table 1) with an average of 389.62 silicoDArT markers per chromosomes; the maximum number of silicoDArT (643) was found on chromosome 6. The average number of silicoDArT markers/Mbp on all chromosomes was 19.26; chromosome 6 showed the maximum number (24.80), while chromosome 7 had the minimum (13.90).

Table 1
Distribution of DArTseq markers on different sesame chromosomes.

Chromosome no	Chromosome size (Mb) *	Number of silicoDArT markers	silicoDArT marker/Mbp	No of SNP markers	SNP marker/Mbp
1	20.26	391	19.35	539	26.60
2	18.42	407	22.10	483	26.22
3	25.85	545	21.08	733	28.36
4	20.59	321	15.59	378	18.36
5	16.58	397	23.94	381	22.98
6	25.97	643	24.80	662	25.49
7	16.76	234	13.90	238	14.20
8	26.18	492	18.79	539	20.59
9	22.85	505	22.10	533	23.33
10	19.49	332	17.03	387	19.86
11	14.05	301	21.42	399	28.40
12	16.28	255	15.60	315	19.35
13	16.47	242	14.69	234	14.21
Scaffold	-	724	-	305	-
Unknown position	-	326	-	348	-
Total	259.73	6115	23.54	6474	24.93

*Indicates the chromosomal size taken from the reference genome published by Wang et al

All of the markers (6115) showed ≥ 95% reproducibility (Fig. 3B) and had a call rate value ≥ of 80% (Fig. 4B) with an average value of 90.55 (Table 1). However, low-frequency markers can affect statistical analysis [47]. As such, 1113 markers with extremely low one ratio (<0.05), scaffold as well as unknown markers were not considered in the analysis. In total, 5002 silicoDArT markers cleared all the quality parameters and were selected for the study. Among the 5002 informative markers, around 0.18% were observed in PIC class 0.45 to 0.50 and 71.57% in 0 to 0.05 classes (Fig. 5). PIC values of the remaining markers were distributed (0.2–8.72%) across the rest of the marker group. Therefore, the median (0.027) was located far to the average PIC value of 0.07 and the data exhibited more on one direction distribution.

A total of 6474 SNP markers (Table 1) were also identified of which 5821 were aligned with the updated genome assembly and annotation available in http://ocri genomics.org/Sinbase_v2.0 (genome assembly) and http://ocri-genomics.org/Sin_SNP_430RIL.tar.gz (SNP information). The other 305 were scaffold and 348 were unknown markers. Based on updated genome assembly and annotation 5821 SNP markers were distributed on all 13 chromosomes of sesame (Fig. 2 **and** Table 1) with an average of 447.7 SNP markers per chromosomes; the maximum number of SNP (733) was found on chromosome 3. The average number of SNP markers/Mbp on all chromosomes was 22.15; chromosome 3 showed the maximum number (28.36), while chromosome 7 had the minimum (14.20).

SNP markers had an average of 98% reproducibility (Fig. 3A) and an 86% call rate (Fig. 4). 100% SNP markers had ≥ 94% reproducibility, of which 3,036 were found to be 100% reproducible (Fig. 3A). The call rate exhibited variance ranging from 40–100%. Around 21.44% of SNP markers displayed a <75% call rate (Fig. 4A) and were therefore not considered for this study. Similar to the above silicoDArT markers those nucleotide polymorphisms that had missing rates >0.25 and one ratio ≤ 0.5 had removed and a set of 4638 SNPs was generated. These markers were determined to be highly informative with an average PIC value of 0.08, and 0.07 median. Around 40.40% of markers were in the range (0–0.05) PIC value and only three markers in the highest PIC value range (0.45 to 0.50) (Fig. 5). The remaining PIC value groups increased from the highest range towards the lowest range and these ranging from 0.23 to 23.87% each.

For further analysis the allele frequency for each SNP site was calculated and purified with MAF: the MAF of the SNPs varied from 0 to 49.6%, with an average of 5.1%, and ~61.29% of the SNPs had a low frequency (MAF <0.05) across the 300 accessions. After excluding the SNPs with a MAF <0.01, there were left 2997 (~64.61%) high-quality SNPs (**Additional file 2: Table S2**) evenly distributed across the whole genome that could be used for subsequent analysis.

Analysis of genetic diversity

The number of accessions, number of alleles, genetic diversity, heterozygosity, and the polymorphism information content (PIC), and major allele frequency of the eight populations are shown in Table 2. The mean PIC values for each SNP locus in sesame collections and introductions from Africa, Amhara, Asia, BG, Improved, Oromia, SNNP and Tigray were 0.18, 0.09, 0.14, 0.06, 0.1, 0.06, 0.06 and 0.12, respectively.

Table 2
Summary of the genetic diversity of the 300 sesame accessions based on their different geographical regions

Group	No of accessions	Allele No	Gene Diversity(GD)	Heterozygosity	PIC	Major Allele Frequency	GD with 3% Missing and 0.05MAF
Introduced from Asia	7	1.6	0.17	0.19	0.14	0.88	0.24
Introduced from different Africa countries except Ethiopia	68	1.97	0.21	0.15	0.18	0.86	0.26
Amhara	56	1.82	0.10	0.09	0.09	0.94	0.15
BG	38	1.68	0.07	0.07	0.06	0.96	0.11
Improved	16	1.59	0.12	0.12	0.10	0.93	0.18
Oromia	52	1.7	0.06	0.06	0.06	0.96	0.096
SNNP	3	1.24	0.08	0.08	0.06	0.95	0.10
Tigray	60	1.85	0.13	0.12	0.12	0.92	0.19
Continents							
Asia	7	1.6	0.17	0.19	0.14	0.88	
Africa	293	2.0	0.14	0.10	0.12	0.92	
By partition Africa							
E. Africa/ Ethiopia	231	1.97	0.10	0.09	0.09	0.94	
N. Africa	27	1.86	0.23	0.19	0.19	0.83	
S. Africa	18	1.75	0.15	0.13	0.13	0.9	
W. Africa	17	1.7	0.14	0.11	0.12	0.9	
Mean	300	2.0	0.14	0.11	0.12	0.92	

The mean number of alleles for each population was 1.97, 1.82, 1.6, 1.68, 1.59, 1.7, 1.24

and 1.85 respectively. The tendencies of the mean number of alleles were in the order

Africa > Tigray > Amhara > Oromia > BG > Asia > Improved > SNNP respectively. The average gene diversity of the panel was 0.14; the highest was 0.5 on LG11, followed by LG4 (0.498), with the lowest value found

(0.01) on LG6, LG4 and LG8 (**Table 4.2** and **Additional file 4: Table S4**). The ranges for the estimates of gene diversity in Africa, Amhara, Asia, BG, Improved, Oromia, SNNP, and Tigray were 0–0.5, and their corresponding averages were 0.21, 0.10, 0.17, 0.07, 0.12, 0.06, 0.08, 0.13, and 0.13 respectively. Heterozygosity values ranged from 0.06 (BG) to 0.19 (Asia), with an average of 0.1. Sesame population from Africa without Ethiopian had the highest level of PIC, gene diversity, and mean number of alleles, but the lowest level of major allele frequency (0.86). Sesame population from BG, Oromia, and SNNP had the lowest level of PIC and gene diversity but the highest level of major allele frequency (0.96, 0.96, and 0.95).

When we consider at the continent level, Asia had the highest level of Heterozygosity, PIC, and gene diversity than Africa, but had less allele no and major allele frequency.

Based on the size of the sample and the result of Africa we further portioned into four geographical origins based on direction (North, South, East and West Africa) and compared with Asia collections, then North Africa had the highest level of Heterozygosity, PIC, and gene diversity than other three directions of Africa geographical origin and Asia. Relatively East Africa (Ethiopian) collection had the lowest level of Heterozygosity, PIC, and gene diversity and had the highest level of allele no and major allele frequency.

Genetic relationships among Germplasm

The average genetic distance among 300 sesame accessions and improved varieties was

0.165 and the highest genetic distance (0.394) was calculated between two Oromia landraces

“Najjoo-68 (gabaa kamijaa)” and “17712”, while the minimum genetic distance (0.015)

was found between the Tigray landrace “9694” and one of the African country Egypt landrace “227888” (**Additional file 3: Table S3**). Genetic diversity among eight pre-grouped populations based on the magnitude of the Nei genetic distance moderate differentiations was revealed between populations from the geographical origin of Asia with Amhara (0.052), BG (0.056), Oromia (0.057) and SNNP(0.073) and the rest pairwise Nei genetic distance between populations from geographical origin revealed lower differentiations (< 0.05) (Table 3).

Table 3
Pairwise population Nei's genetic distance showing the magnitude of genetic differentiation between sesame populations from different sources

	AFRICA	AMHARA	ASIA	BG	IMPROVED	Oromia	SNNP	Tigray
AFRICA	0.000							
AMHARA	0.021	0.000						
ASIA	0.040	0.052	0.000					
BG	0.020	0.006	0.056	0.000				
IMPROVED	0.022	0.012	0.037	0.015	0.000			
Oromia	0.021	0.010	0.057	0.004	0.014	0.000		
SNNP	0.030	0.020	0.073	0.016	0.028	0.013	0.000	
Tigray	0.024	0.006	0.047	0.015	0.009	0.017	0.027	0.000

Cluster analysis of the 300 accessions derived from the eight different geographical origins was performed using the allele-sharing distance (ASD) method and the results allowed us to group them into four clusters. It also helps us to identify accessions and genotypes wrongly assigned to another geographical origin. The first cluster comprised majority from different countries of Africa (28), all accessions that were introduced from Asia (7) and the different regions of Ethiopia, Amhara (8), Benshangul-Gumz (4), Oromia (10), SNNP (1), Tigray (12), and 7 improved varieties. The second cluster constitutes the highest number of accessions that were collected from the different regions of Ethiopia, Amhara (40), Benshangul-Gumz (34), Oromia (41), SNNP (2), Tigray (23), and 4 improved varieties, the remaining 13 accessions were introduced from different Africa countries. The third cluster is comprised majority from the Tigray region (25) and a small number from Amhara (n = 8), Oromia (1), and 5 Improved varieties, the remaining 4 accessions were introduced from different countries of Africa. Cluster 4 comprised all the accessions that were introduced from one of the African country Egypt (23) only (Fig. 6). There were no relationships between cluster grouping and pedigree of accessions and genotypes, although most of the introduced sesame genotypes from Egypt grouped in cluster 4.

An analysis of molecular variance (AMOVA)

Analysis of molecular variance (AMOVA) among the 300 sesame germplasms indicated that 8.31% of the variance was due to genetic differentiation among the populations, 15.24% of the variance was accounted by genetic differentiation among individuals within populations, while the

remaining 76.44% of the variance was due to the differences within individuals (Table 4)

Table 4
AMOVA between different Geographical regions

Source of variation	df	SS	Variance components	Percentage variation
Among populations	7	9643.614	17.41032	8.31858
Among individuals Within populations	298	61461.430	31.89832	15.24089
Within individuals	300	46195.500	159.98566	76.44052
Total	599	117300.544	209.29431	
df = degree of freedom, SS = Sum of Square				

While in terms of continents 11.49% of the total molecular variation observed was due to differentiation between Asia and Africa, 19.45% of the variance was accounted by genetic differentiation among individuals within continents, while the remaining 69.06% of the variance was due to the differences within individuals (Table 5)

Table 5
AMOVA between the Continents

Source of variation	df	SS	Variance components	Percentage variation
Among continents	1	941.728	26.62029	11.49050
Among individuals Within continents	292	70163.316	45.06621	19.45258
Within individuals	300	46195.500	159.98566	69.05692
Total	599	117300.544	231.67217	

When we see further, In terms of population subdivision with different directions of Africa and Asia 22.17% of the total molecular variation observed was due to differentiation between different directions of Africa and Asia, 10.69% of the variance was accounted by genetic differentiation among individuals within different directions of Africa and Asia, while the remaining 67.12% of the variance was due to the differences within individuals (Table 6).

Table 6
AMOVA between the Different directions of Africa and Asia

Source of variation	df	SS	Variance components	Percentage variation
Among populations	4	12564.718	52.85675	22.17834
Among individuals Within populations	295	58540.327	25.48356	10.69273
Within individuals	300	46195.500	159.98566	67.12892
Total	599	117300.544	238.32598	

Population structure

A subset of 2997 SNPs was selected for analysis of the population structure. The hierarchical population structure was determined for the entire panel via the Bayesian model-based analysis using the STRUCTURE program. As *K* changed from 1 to 11 by inferring on Delta K of Evanno et al. [48], the log-likelihood value [LnP(D)] increased continuously and inflection was evident when *K* increased numerically from 1 to 4 (Fig. 7A). Thus, the most likely numerical value of *K* was 4. The number of subgroups (*K*) was further validated by the second-order statistics of ΔK. The ΔK value showed a peak at *K* = 4 (Fig. 7B), which supported the classification of the panel into four major subgroups (Fig. 7C). The genetic diversity within each population was explained through the estimation of the expected heterozygosity, which varied from 0.06 (POP2) to 0.31 (POP4). The expected heterozygosity of POP1 was 0.22 and that of POP3 was 0.18. The genetic divergence among the populations revealed by Nei's net nucleotide distance (D) indicated that a higher distance between POP3 and POP4 (0.22) and the genetic distance observed between POP1 and POP2 (D = 0.09) was the least among the pairs of populations. Mean fixation index of sub-populations ranged from 0.39 (POP4) to 0.77 (POP2) (Table 7).

Table 7
Genetic divergence among (net nucleotide distance) and within (expected heterozygosity) population, proportion of membership, and mean value of Fst observed from the study of the population structure of 300 sesame accessions and genotypes using DArTseq-SNP markers

Population	Net nucleotide distance			Expected	% of	Mean Fixation
	pop2	pop3	pop4	Heterozygosity	Membership	Index (Fst)
pop1	0.09	0.13	0.19	0.22	0.24	0.45
pop2		0.11	0.17	0.06	0.50	0.77
pop3			0.22	0.18	0.19	0.57
pop4				0.31	0.07	0.39

When using a probability of membership threshold of 50%, 54, 159, 43, and 23 accessions were respectively assigned into the four subgroups, Pop 1, Pop 2, Pop 3, and Pop 4, while the remaining 21

accessions were classified into a mixed subgroup (Mixed) with 50% levels of membership shared among the three genetic groups (**Additional file 5: Table S5**).

most accessions of Pop 1 introduced from different countries of Africa (27), 7 accessions from different Asia countries, while 18 accessions in total came from different regions of Ethiopia, Amhara (n = 4), Benshangul-Gumz (2), Oromia (5), Tigray (7) and 2 improved varieties. The accessions and genotypes of Pop 2 constitute the largest that was mainly collected from the different regions of Ethiopia, Amhara (n = 40), Benshangul-Gumz (35), Oromia (42), SNNP (2), Tigray (20), and 7 Improved varieties, the remaining 13 accessions were introduced from different Africa countries. The accessions of Pop 3 comprised mainly from three regions of Ethiopia, Amhara (n = 9), Oromia (1), Tigray (26), and 4 Improved varieties, the remaining 3 accessions were introduced from different countries of Africa. Pop 4 introduced from one of the African countries Egypt (23) only. For the Mixed group, 19 accessions were collected from different regions of Ethiopia and 2 accessions from two Africa countries.

The PCA was done to further assess the population subdivisions, Principal component analysis (PCA) based on DArTseq - SNP markers revealed four distinct groups of sesame accessions and genotypes and two principal components, accounting for 93.7% of total variation (Fig. 8A). PC1 explained 84% of the genetic variation found, while PC2 explained 7.7% of the variation, respectively. However, some intermediate lines (admixture) made the grouping less than clear-cut. When considering these intermediate lines, the panel could be neatly divided into four clusters (Fig. 8B) corresponding to the four subgroups as inferred by using the STRUCTURE result.

Discussion

To develop sesame varieties with desirable traits, knowledge of the genetic diversity and relationships among germplasm accessions is vitally important. The actual level of genetic variation existing among genotypes at the DNA level reflected by Molecular markers; hence, they provide a more accurate estimate of variation than does either phenotypic or pedigree information [49].

This study based on the suitability of DArT platforms that applied for the genomic dissection of sesame. A total of 6115 silicoDArT markers were developed, of which 5002 markers provided robust information of the sesame genome in the absence of sequence information. On the other hand, DArTseq SNPs provided 6474 informative markers.

The average PIC values of silicoDArT were almost similar to that of SNP markers. The abundance of silicoDArT and SNP markers may achieve better genome coverage through the sampling of a greater number of points in the whole genome, as marker density has a high correlation with gene density [57, 58]. Therefore, both silicoDArT and SNP markers may better suit for genetic diversity studies, association/linkage mapping, and/or sequence-based physical mapping in sesame. Additionally, the co-dominant inheritance pattern of SNP markers may increase the utility of DArT platforms for genetic identity

and parentage analysis [59]. In comparison with the other existing marker technologies like microsatellite markers, DArT markers are pertinent to high-throughput work and have merits in terms of cost-effectiveness and time aspect [60]. The effectiveness of silicoDArT and SNP markers varies depending on the type of application. For genetic diversity and linkage mapping a large number of silicoDArT markers are suitable. However, for genetic identity and product quality testing, both markers can perform equally. Due to the opportunity to track alleles from parental genotypes, the co-dominant SNP markers are more suitable in plant identity and parentage analysis than silicoDArT.

Then, 2997 SNP markers were filtered with a call rate of 75%, and those having > 0.01 minor allele frequency were used for the analysis, The proportion of rare SNPs (i.e., MAF < 0.05) we examined amounted to ~61.29% %, which was similar to those reported for the genomes of sesame [38]. In our study, a high proportion of rare SNPs have two explanations. Firstly, since the SNPs were identified via DArTseq conducted by GBS technology, providing a broad genome coverage, they should be less prone to bias than would be low-coverage sequencing data[61]. Secondly, in following its recent program to conserve genetic resources, a significant number of minor sesame varieties have been collected and preserved by Ethiopian Biodiversity and research centers. The SNPs with a MAF < 0.05 were removed in several previous studies [62, 63]. However, rare SNPs might also have control over the expression of a particular phenotype [64]. Providing that the number of individuals with a specific genotype will 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 ability to check these rare alleles.

The average value of genetic diversity (0.14) was lower in the present study than in the earlier reports for the sesame collections analyzed with SNPs markers [29, 32, 38, 39] and SSR markers [65, 66]. However, with the use of 1022 SNP markers that were filtered with a call rate of 97% and > 0.05 MAF similar to the report on [38], the average value of genetic diversity (0.19) was higher than in the earlier reports for the sesame collections analyzed with different markers types [32, 38]. The broad range of variability among collections might be a source of the differences observed in genetic resources (such as landraces, advanced breeding lines, cultivars, etc.), data filtering methods, sampling approaches, and the number of markers [65]. The type of marker is also an important factor for the identification of gene diversity; In general, the genetic diversity estimated by SNPs may be lower than those estimated through SSR markers; however, the accurate consideration of genetic diversity reflected the number of loci instead of the number of alleles [38]. Therefore, sufficiently large numbers of next-generation-based SNPs are analyzed across the genome and are ready to estimate accurate genome-wide diversity in several crop species.

Considering the genotypes based on their geographical origin, Africa (0.21) without the different region of Ethiopian was more diverse than Asia and Different regions of Ethiopia collections, but when we compare at the continent level by including different regions of Ethiopia as Africa, Asia (0.17) was more diverse than Africa (0.14), even if the sample of Asia was little. This finding was expected because the geographical origin of crops generally shows a higher genetic diversity, as reported previously for cotton (Paterson A., 2009) and *Oryza* ssp. [61]. Laurentin and Karlovsky [28] also obtained higher genetic diversity in sesame accessions collected from Asia.

Based on the size of the sample and the result of Africa we further portioned into four geographical origins based on direction (North, South, East and West Africa) and compared with Asia collections, then North Africa collection (0.23) was more diverse than other three directions of Africa geographical origin and Asia also. East Africa (Ethiopian) collection was less diverse than the others. This indicates even if, Ethiopian sesame well known in international market and has its own taste and aroma, it needs a further breeding program to broaden genetic diversity with hybridization and the introduction of a highly diverse collection of North Africa and different countries of Asia.

Distribution of heterozygous sesame genotypes and SNP markers revealed low values of

heterozygosity, the average heterozygosity with in sesame panel was 0.1; this suggests that the accessions we used were close to being inbred lines. Hence, the accessions selected were suitable for investigating multiple phenotypic traits in a multi-plot field test over several years and to also carry out GWAS.

The genetic distance matrix among the sesame populations from 8 geographical origins was also used to construct the clustering tree (Fig. 6). The similarity coefficients ranged from 0.015 to 0.394, with an average of 0.165. The sesame populations could be clustered into four groups. The clustering Dendrogram based on the geographical distribution of accessions showed that the majority sesame accessions from the identical origin didn't classify properly on the premise of the country of origin except those accessions introduced from one of African country Egypt. Similar results were reported previously indifferent sesame germplasm [39, 68–70] and in other crops, including wheat [71], finger millet [72], and sorghum [73]. The explanation for this unequal distribution of sesame accessions based on the geographical origin may be associated with the gene flow among the various geographical areas due to migrations of people who traded with other regions for a century or who carried seeds for cultivation.

Similarly, Laurentin and Karlovsky [28] found no association between genetic diversity and accession origin, and they proposed that ecological and geographical factors have not played a significant role in the evolution of sesame. The present AMOVA analysis also supported the possibility of high rates of gene flow between regions, because the genetic variation among the geographical groups accounted for 8.3% of the total variation and in terms of continents, 11.49% of the total molecular variation among the continents (Table 3).

Most of the genotypes used in this study have been used as parental lines or have a similar genetic background, so a mixture of pedigree observed in all clusters. In our result, the genotypes in Cluster 2 and 3 were collected from different regions of Ethiopia that showed a tendency to cluster together and mostly originating from Ethiopia. This result matches the hypothesis that sesame seeds were dispersed to nearby countries by human activities. Subsequently, these distributed sesame genetic resources were later utilized in further breeding activities to a modern cultivars that were commercialized.

Cluster 1 contained accessions originating from two different continents (Africa and Asia), a close genetic relationship between accessions from East Africa, South Africa, North Africa, and West Africa to the accessions from Asia. This close genetic relationship observed might be due to the introduction of similar sesame genetic stock into many countries and material exchange among widely separated locations [74].

Moreover, the exchange of plant materials between Asia and East Africa dated back to a long time ago and is still occurring [75], with a gentle increase in annual exportation of raw sesame seeds mainly for industrial applications. The likelihood of crossover events between materials from different locations grown within the same area is high, knowing that cross-pollination in sesame has been reported to occur at a frequency between 5% and 60% [66]. This crossing could result the similarity of accessions from the eastern a part of Africa and Asia. Similar patterns have also been observed by other researchers [28, 69, 74]. Most of the genotypes used in this study have been used as parental lines or have a similar genetic backgrounds, so a mixture of pedigree observed in all clusters.

Cluster 4 indicates the possibility of genotypes from the same origin those were genotypes observed from one of the African countries Egypt were grouped together (Fig. 6).

Population Structure of the Association-Mapping Panel

The complex breeding history of the numerous important crops and also the limited gene flow in most wild plant populations have created complex structures within their germplasms [76]. Detailed knowledge about the population structure in an association panel is thus important to avoid any spurious associations [77]. An assessment of structure in sesame has been reported by using different populations. As an example, Ali et al., 2007[68] evaluated 96 sesame accessions, collected from different parts of the world and clustered into just two major groups that discriminated varieties as associated with their geographical origin. And [37] divided 705 sesame accessions into two clusters by employing a neighbor-joining tree. Recently, [38] with the *K* value of 2 was determined by both the LnP (D) and ΔK. By using a 70% probability of membership threshold, the 366 sesame germplasm was successfully divided into three subgroups (Pop 1, Pop 2, and the Mixed) and [39] divided 95 Mediterranean sesame core collection that contains agro-morphologically superior sesame accessions from geographically diverse regions in four continents (Asia, Europe, America, and Africa) into three groups ascertained using STRUCTURE with K = 3.

Similarly, in our study, the *K* value of 4 determined by both the LnP (D) and ΔK. By employing a 50% probability of membership threshold, the panel was successfully divided into four subgroups (Pop 1, Pop 2, Pop 3, and Pop 4) and the remaining 21 accessions were clustered as an admixture with varying levels of membership shared among the four genetic groups, based on structure analysis. The occurrence of some admixed/hybrid and introgressive hybrid genotypes indicated frequent hybridization and introgression events. Although the extent and significance of natural hybridization/introgression are unclear [79], new gene combinations between domestic cultivars and their wild or weedy relatives are important for the evolution of domesticated plant species [80].

The genetic diversity within each population was explained through the estimation of the expected heterozygosity (the average distances between each individual in the same cluster), which varied from 0.06 (POP2) to 0.31 (POP4). The expected heterozygosity of POP1 was 0.22 and that of POP3 was 0.18. The genetic divergence among the populations revealed by Nei's net nucleotide distance (D) indicated that a higher distance between POP3 and POP4 (0.22) and the genetic distance observed between POP1 and

POP2 (D = 0.09) was the least among the pairs of populations. Mean fixation index of sub-populations ranged from 0.39 (POP4) to 0.77 (POP2) (Table 5).

The population genetic structure reflects interactions among species with regard to their long-term evolutionary history, mutation and recombination, genetic drift, reproductive system, gene flow, and natural selection [81, 82]. Thus, an understanding of the extent and structure of the genetic diversity of a crop could be a prerequisite for the conservation and efficient use of the germplasm available for breeding [83]. The various approaches (STRUCTURE, PCA, and the clustering tree) used to analyze the structure and relation of the sesame germplasm appeared to provide complementary information. The neighbor-joining tree divided the sesame germplasm into four main clusters which are in complete concordance with the structure and PCA analysis results. These results suggest that the crossing among inter-cluster genotypes may develop cultivars with promising agronomic traits.

According to the AMOVA results, 8.3% and 11.49% of the marker variation was explained among the population from different geographical regions of the sesame panel and differentiation between Asia and Africa population respectively. This result suggests the absence of a complicated population structure in our association-mapping panel. Relatively, 22.17% of the marker variation was explained among the population from different directions of Africa and Asia, this suggests the presence of certain complicated between population structure in different directions of Africa and Asia association-mapping panel.

In this study, most collections (225) were from Ethiopia and a specific collection was from West, South, and North Africa and seven collections were from 4 Asia countries. Ethiopian sesame has useful characteristics, and often branded as ‘Humera’, ‘Gondar’ and ‘Welega’ types, well known in the world market by their white color, sweet taste and aroma. The Humera and Gondar sesame seeds are suitable for bakery and confectionary purposes and the high oil content of the Welega sesame seed gives a major advantage for edible oil production[84]. Collections that were introduced from a different direction of Africa and Asia were accustomed to compare the degree of genetic relationship and differentiation among genetic resources of Ethiopian collection, which broadens genetic diversity can also be used to combine alleles for valuable agricultural traits [86].The SNPs obtained from this collection could benefit future breeding and association mapping work in sesame. Our diversity analysis of this collection revealed genetic relationships among the accessions that may be valuable for parental selection in sesame improvement research. Therefore, the identification of genetically distant accessions (such as Najjoo-68 (gabaa kamijaa) and 17712) for hybridization in sesame breeding programs has the potential to lead to the development of elite varieties. Even based on economical traits and the distance we got from SNPs, we can further select a number of accessions for the different breeding programs.

Conclusions

The present research showed the effectiveness of DArTseq in characterizing the genetic diversity

and population structure of sesame collection. The gene diversity values calculated based on the 2997 SNPs and 300 genotypes suggest that among continents, Asia, out of the different directions of Africa,

North Africa is relatively genetically diverse. And even if Ethiopian sesame has useful characteristics and has its own aroma and tests, the collection of East Africa /Ethiopia is less diverse and need further crossing and introduction of germplasm for creating variability that favor improvement for different biotic and abiotic stress. The local and exotic collection provide useful genetic data for future molecular-based studies. This study also supports the idea; ecological and geographical factors less effective in the evolution of sesame. This finding provides guidance to the systematic utilization and conservation of the genetic resource and indicates the further collection of sesame genotypes from these different origins.

Therefore, our next objective is to identify sesame genotypes with desirable traits and to

conduct association mapping studies focusing on resistance to biotic stresses and seed yield under environmental stresses like drought and water lodging.

Methods

Plant materials

A set of 209 Ethiopian local collections, 75 exotic collections and 16 released varieties which is a total of 300 genotypes were used in this study. The 209 Ethiopian local collection collected between 1931 and 2008 a.m.s.l in different climate zones and regions Amhara (56), Benshangul-Gumz (BG) (38), Oromia (52), SNNP (3), and Tigray (60) and 16 well popularized and currently released sesame varieties between 1942 and 2014. Among 75 exotic collections, 68 were introduced in seven African countries in four directional geographical origins; North Africa (27), South Africa (18), West Africa (17), and the remaining 6 were in East Africa without including the Ethiopian collection. The other 7 were collection introduced in 4 Asian countries. These were kindly provided by the Ethiopian Biodiversity Institute (EBI) and regional and federal research centers. The sampling sites covered a wide range of natural eco-geographical locations and the description presented in (Fig. 9 and **Additional file 1: Table S1**).

DNA extraction

The seed yield of each sesame genotype was harvested when the plants got matured. Seeds of each genotype were placed in the 2 mm deep wheel tube together with 1 steel ball that had 30 mm diameter and crushed with Geno/Grinder followed by mixing 800 ul Lysis buffer to the sample of each genotype powder for tan bead DNA extraction process. Those samples were incubated for 1 hour at 65°c then centrifuge for 5 minutes to remove plant tissue debris. The lysate was taken and load on column #1 and the nucleic acid of the samples was extracted with Automated Nucleic acid Extractor (Maelstrom series). During the process, the silicon dioxide layer coated on the magnetic beads adsorb nucleic acid from samples, remove contaminants with wash Buffer, and elute purified genomic DNA by Elution Buffer. At the end of the program, collected Nucleic acid was found in column #6 with a clean tube. DNA quality was evaluated on 0.8% agarose gels and it was adjusted to 50 ng/µl for GBS analysis.

GBS library preparation and sequencing

DArTseq combines genome complexity reduction methods and next-generation sequencing platforms [87–90]. Therefore, DArTseq represents a new implementation of the sequencing of complexity-reduced representations [91] and more recent applications of this concept on the next-generation sequencing platforms [92, 93]. DArTseq libraries (96-plex) were prepared for the 312 accessions using 50 ng of DNA each. Briefly, DNA samples were digested individually with *Pst*I- *Mse*I restriction enzymes. In this technology, the *Pst*I-based complexity reduction method [40] was applied for the enrichment of genomic representation with single-copy sequences. This method involved the digestion of DNA samples with a rare cutting enzyme *Pst*I, paired with a set of secondary frequently cutting restriction enzyme *Mse*I, ligation with site-specific adapters, and amplification of adapter-ligated fragments. Post digestion with a *Pst*I-*Mse*I pair, a *Pst*I overhang compatible oligonucleotide adapter (5′-CAC GAT GGA TCC AGT GCA-3′ annealed with 5′ -CTG GAT CCA TCG TGC A-3′) was ligated, and the adapter-ligated fragments were amplified in adherence to the prescribed standard procedures [40]. To develop SNP and silicoDArT markers, the DArTseq technology was optimized using replacing a single *PstI*-compatible adapter with two different adapters corresponding to two different restriction enzymes (RE) overhangs. The *PstI*-compatible adapter was designed to include Illumina flowcell attachment sequence, sequencing primer sequence, and staggered varying length barcode regions. The reverse adapter contained the flowcell attachment region and *MseI*-compatible overhang sequence. Only “mixed fragments” (*PstI*–*MseI*) were effectively amplified in 30 rounds of PCR using the following reaction conditions: 1 min at 94◦C for initial denaturation; 30 cycles each consisting of 20 s at 94◦C for denaturation, 30 s at 58◦C for annealing and 45 s at 72◦C for the extension, and finally a 7 min extension step at 72◦C. The genomic representations were generated following the procedures described by Kilian et al [94].

Post-PCR, cluster generation was carried out in cBOT (Illumina) according to the procedures described by the manufacturer. Briefly: 10 nM DNA of each library is denatured, diluted in hybridization buffer, loaded into the machine, and clusters are generated in the flow cell by cBOT with use of the set cBOT reagents (Bridge Amplification). During cluster generation, the molecules of each library were attached to the flow cell surface and amplified to form clonal clusters.

Next-generation sequencing technology was implemented using the sequencer HiSeq2500 (Illumina, USA) to detect SNPs and silicoDArT markers. The flow cell with clusters generated in the previous step (cBOT) is loaded to the HiSeq 2500 together with the sequencing reagents. HiSeq 2500 performed sequencing according to user-selected sequencing parameters. All amplicons were sequenced in a single lane. The single-read sequencing was run for 77 cycles.

Real-time Analysis (RTA) happened simultaneously to the sequencing run and RTA data were outputted to a server. The main sequence outputted data were base calling files *.bcl files. These files were the input files for downstream data conversion. The primary workflow was a custom build software for downstream processing of *.bcl files. The first step was a conversion of *.bcl files which was done by Illumina bc12fastq software embedded in the primary workflow, the second step performed two functions at the same time: first using target definition from DArTdb the software splits sequencing reads according to the barcode sequence (demultiplexing), Secondly, it removed reads below quality filters. Two quality filters

were applied: more stringent for barcode sequence and less stringent for the remaining part of the sequencing read.

Finally fold compression of the sequence tags was copied to DArTdb (Diversity Arrays Technology data base, Australia) for permanent storage. We extracted compressed sequence tags from DArTdb and load them to DArTsoft14 for marker data extraction. DArTsoft14 extracts two types of marker data: SNPs and SilicoDArTs. SilicoDArTs represent dominant markers and is scored in a binary format "1"= Presence and "0"= Absence of restriction fragment with the marker sequence in the genomic representation of the sample. **"-"** represents calls with non-zero counts but too low to score confidently as "1" (often representing heterozygotes). Single Nucleotide Polymorphism (SNPs) can be defined as a variation in the base composition of a single nucleotide position within a specific locus of a single chromosome of the haploid set. In standard format, SNPs markers were presented for reference and SNP alleles for each marker and genotype. This format of SNPs can be converted to other formats if required. The report was prepared as binary or read counts file, or both depending on the order specifications. Two technical replicates of the DNA samples of each of 21 accessions were genotyped to calculate the reproducibility of the marker data. Thereafter, the SNPs and SilicoDArTs obtained were run against the sesame reference genome database (http://ocri genomics.org/Sinbase/login.htm.) to understand on which chromosomes of sesame the SNPs and SilicoDArTs were located.

Quality analysis of marker data

The markers were tested for reproducibility (%), call rate (%), polymorphism information content (PIC) and, one ratio. Scoring of reproducibility involved the proportion of technical replicate assay pairs for which the marker score exhibited consistency. The call rate determined the success of reading the marker sequence across the samples and was estimated from the percentage of samples for which the score was either '0' or '1'. PIC is the degree of diversity of the marker in the population and showed the usefulness of the marker for linkage analysis. One ratio constitutes the proportion of the samples for which genotype scores equaled '1'.

Data analysis

DArTseq markers were mapped using the consensus map version 4.0 (www.diversityarrays.com)

developed by DArT Pty. Ltd., Australia, and the updated genome assembly and annotation issued from the Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, available online at (http://ocri-genomics.org/Sinbase_v2.0 (genome assembly) and http://ocri-genomics.org/Sin_SNP_430RIL.tar.gz (SNP information).

DArTseq raw data were filtered according to markers criterion; minor allele frequency > 0.01% and missing data ≤ 25%. The summary statistics of the filtered DArTseq markers such as the expected heterozygosity (He) or genetic diversity (GD), minor allele frequency (MAF), and the polymorphic information content (PIC), were calculated using Power Marker v 3.25 [95]. PIC was estimated based on the probability of finding

polymorphisms between any two random samples while gene diversity defined as the probability of two randomly chosen alleles from the population is different.

For genetic differentiation and relationships, Genetic distances between each pair of accessions and between pre-grouped populations were measured based on both the shared-allele distances (ASD) and Nei genetic distance [96] between populations or individuals by using KD compute plugin system.

For cluster analysis of the collection, the allele-sharing distance matrix was computed as described by Goa et al. [97]. Classification of the individuals into groups was performed using the allele sharing matrix and Ward's minimum variance algorithm [98]. The clustering algorithm and Principal component analysis (PCA) for the genetic relationships among individuals was implemented in version KD compute plugin system.

The software Arlequin V3.5 [99] was used to calculate the genetic variation between and within geographical groups with an analysis of molecular variance (AMOVA).

Population structure

The genetic structure of the genotype was analyzed using STRUCTURE v.2.3.4 [100]. The number of hypothetical subpopulations (K) was estimated with the STRUCTURE software through the application of a Bayesian clustering approach for the organization of genetically similar cultivars into the same subgroups. Five individual Markov Chain Monte Carlo (MCMC) simulations were conducted for each K-value from 1 to 11 with a burnin length of 50,000, followed by 100,000 iterations. The admixture model was applied without using any prior population information and correlated allele frequencies. The log-likelihood of the observed data for each K value was calculated and compared across the range of K values. The best K-value was estimated based on the membership coefficient (Q) for each individual in each cluster. The Q values indicate the level of relatedness of each accession to various subgroups.

The STRUCTURE results were subsequently analyzed by the STRUCTURE HARVESTER application [101] (http://taylor0.biology.ucla.edu/structureHarvester/) to identify a distinct peak in the change of likelihood (ΔK) at the true value of K. CLUMPAK: "a program for identifying clustering modes and packaging population structure inferences across K" (CLUMPAK server) was used (Kopelman *et al.*, accepted to *Molecular Ecology Resources*). Each sesame accession was then assigned to a cluster (Q) based on a probability determined by STRUCTURE V2.3.4, which provided clustering for the genotypes. The cut-off probability for assignment to a cluster was 0.50 for the clusters.

Abbreviations

AMOVA: Analysis of molecular variance; BG: Benshangul-Gumz; CSA: CENTRAL STATISTICAL AGENCY; DArTseq: Diversity Array Technology Sequencing; EBI: Ethiopian Biodiversity Institute; FAO: Food and Agriculture Organization of the United Nations: Statistical Databases; GBS: Genotype by Sequencing; GD: Genetic Diversity; He: Heterozygosity; MAF: Minor Allele Frequency; MCMC: Markov Chain Monte Carlo; MoARD: Ministry of Agriculture and Rural Development ; NCBI: *National Center for Biotechnology*

Information; NGS: Next-generation Sequencing; PIC: Polymorphic Information; PCA: Principal Component Analysis; RE: Restriction Enzyme; RTA: Real-time Analysis; SNNP: Southern Nations, Nationalities, and Peoples'; SNP: Single Nucleotide Polymorphism.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data sets supporting the results of this article are included in this manuscript and its additional information files.

Competing interests

The authors declare that they have no competing interests.

Funding

Fingerprinting of genotype was made possible with the financial support from Amhara Agricultural Research Institute of (ARARI) and Gondar Agricultural Research Center through BENEFIT-SBN project. The role of the funding bodies is limited to direct funding of the fingerprinting of genotypes that result in this manuscript because of the project faze.

Authors' contributions

TT, KT, and GK conceived and designed the study. TT and CZ assembled the panel and participated in genotyping. TT prepared the manuscript and carried out the data analysis. All co-authors participated in interpreting the data, revising and editing the manuscript and approved the final version of the manuscript.

Corresponding author

Correspondence to Tewodros Tesfaye

Email; -tesfaye.tewodros@yahoo.com

Acknowledgements

The authors are grateful to Ethiopian Biodiversity Institute (EBI) and Assosa, Bako, Gondar, Humera, and, Werer Agricultural Research Centers for their generous provision of collection and improved varieties used in this study. The financial support of Amhara Agricultural Research Institute of (ARARI) and Gondar

Agricultural Research Center through BENEFIT-SBN project, is appreciated. The first author further obliged to Addis Ababa University and IGSS program in BecA-ILRI Hub, Nairobi-Kenya for the training provided.

Authors' information

[1]Amhara Agricultural Research Institute, Gondar Research Center, P. O. Box 1337, Gondar, Ethiopia,

Tewodros Tesfaye, Email; - tesfaye.tewodros@yahoo.com

[2]Addis Ababa University, College of Natural Sciences, Department of Microbial, Cellular and Molecular Biology, P. O. Box 1176, Addis Ababa, Ethiopia, Email;-

Kassahun Tesfaye, Email; - kassahuntesfaye@yahoo.com

[3]Ethiopian Institute of Agricultural Research, Holeta Research Center, P. O. Box 2003, Holeta, Ethiopia,

Gemechu Keneni, Email; - gemechukeneni@yahoo.com

[4]International Livestock Research Institute, Box 30709, 00100 Nairobi, Kenya, Email;-

Cathrine Ziyomo, Email; - c.ziyomo@cgiar.org

Corresponding author: tesfaye.tewodros@yahoo.com

References

1. Bedigian D, Harlan JR. Evidence for cultivation of sesame in the ancient world. Econ Bot. 1986.
2. Ashri A. Sesame breeding. Plant Breed Rev. 1998;16:179–228.
3. Fuller DQ. Further Evidence on the Prehistory of Sesame. Asian Agrihist. 2003;7:127–37.
4. Alegbejo MD, Iwo GA, Abo ME, Idowu AA. Sesame: A potential industrial and export oilseed crop in Nigeria. J Sustain Agric. 2003.
5. Simmonds NW, Purseglove JW. Tropical Crops. Dicotyledons. J Ecol. 1969.
6. Tunde-Akintude T, Oke MO, Akintunde BO Sesame Seed, oil seeds. In Tech; 2012.
7. Food and Agriculture Organization of the United Nations. FAOSTAT. Food and agriculture data. FAOSTAT Data- base Gateway. 2017.
8. Borchani C, Besbes S, Blecker C, Attia H. Chemical Characteristics and Oxidative Stability of Sesame Seed, Sesame Paste, and Olive Oils. JAgrSciTech. 2010.
9. Dorothea Bedigian. History and Lore of sesame in Southwest Asia. Econ Bot. 2004.
10. Bhat KV, Babrekar PP, Lakhanpaul S. Study of genetic diversity in Indian and exotic sesame (Sesamum indicum L.) germplasm using random amplified polymorphic DNA (RAPD) markers. Euphytica. 1999.
11. IPGRI and NBPGR (International Plant Genetic Resources Institute and National Bureau of Plant Genetic Resources). Descriptors for sesame (Sesamum spp.). New Delhi: National Bureau of Plant

Genetic Resources; 2004.
12. Bedigian D. Characterization of sesame (Sesamum indicum L.) germplasm : a critique. 2010;:641–7.
13. CSA (CENTRAL STATISTICAL AGENCY). AGRICULTURAL SAMPLE SURVEY 2015/2016. Report on Area and Production of Major Crops(PRIVATE PEASANT HOLDINGS, MEHER SEASON). Stat Bull. 2016.
14. Mkamilo GS, Bedigian D In PROTA (Plant Resources of Tropical Africa/Ressources végétales de l'Afrique tropicale), Wageningen, Netherlands. 2007.
15. Sharaby N, Butovchenko A. Cultivation technology of sesame seeds and its production in the world and in Egypt. IOP Conf Ser Earth Environ Sci. 2019;403.
16. Gebremichael DE. Sesame (Sesamum indicum L.) Breeding in Ethiopia. Int J Nov Res Life Sci. 2017;4:1–11. file:///home/rnsarma/Downloads/SesameBreedinginEthiopia-924.pdf.
17. MoARD (Ministry of Agriculture and Rural Development). Crop Variety Register Book. Animal and Plant health Regulatory Directorate. Addis Abeba, Ethiopia.
18. CSA. Sample survey area and production of major crops (private peasant holdings, meher season). 2019;I:1–58.
19. Institute of Biodiversity Conservation (IBC). Ethiopia: Third country report on the state of plant genetic resources for food and agriculture. Addis Ababa, Ethiopia; 2012.
20. Vavilov NI. The origin, variation, immunity and breeding of cultivated plants. 1951;72:482.
21. Sileshi AM. Genetic Divergence and Correlation Study in Sesame (Sesamumindicum L.) Genotypes. M.Sc thesis. Addis Ababa University, Ethiopia.; 2008.
22. Gidey YT, Kebede SA, Gashawbeza GT. Extent and pattern of genetic diversity for morpho-agronomic traits in Ethiopian sesame landraces (Sesamum indicum L.). Asian J Agric Res. 2012;6:118–28.
23. Yirgalem TGSA. and GT. Assessment of genetic variability, genetic advance, correlation and path analysis for morphological traits in sesame genotypes. Int J Plant Breed Genet. 2013;7:21–34.
24. Desawi Hdru Teklu. Alamerew Kebede S. DEG. Assessment of genetic variability, genetic advance, correlation and path analysis for morphological traits in sesame genotypes. Asian J Agric Res. 2014;8:181–94.
25. Abate M, Mekbib F. Assesment of Genetic Variability and Character Association in Ethiopian Low-Altitude Sesame (Sesamum Indicum L.) Genotypes. J Adv Stud Agric Biol Environ Sci. 2015;2:55–66.
26. Daniel EG, Parzies HK. Genetic variability among landraces of sesame in Ethiopia. African Crop Sci J. 2011;19:1–13.
27. Dagmawi TW, Kassahun T, Endashaw B. Genetic diversity of sesame germplasm collection (SESAMUM INDICUM L.): implication for conservation, improvement and use. Int J Biotechnol Mol Biol Res. 2015;6:7–18.
28. Laurentin HE, Karlovsky P. Genetic relationship and diversity in a sesame (Sesamum indicum L.) germplasm collection using amplified fragment length. 2006;10:1–10.
29. Zhang YX, Zhang XR, Hua W, Wang LH, Che Z. Analysis of genetic diversity among indigenous landraces from sesame (Sesamum indicum L.) core collection in China as revealed by SRAP and SSR markers. Genes and Genomics. 2010.

30. Zhang Y, Zhang X, Che Z, Wang L, Wei W, Li D. Genetic diversity assessment of sesame core collection in China by phenotype and molecular markers and extraction of a mini-core collection. BMC Genet. 2012;13:1. doi:10.1186/1471-2156-13-102.
31. Kumar H, Kaur G, Banga S. Molecular Characterization and Assessment of Genetic Diversity in Sesame (Sesamum indicum L.) Germplasm Collection Using ISSR Markers. J Crop Improv. 2012.
32. Park YCJ, Ra CLW, Lee JCJ. Evaluation of the genetic diversity and population structure of sesame (Sesamum indicum L.) using microsatellite markers. 2011; April.
33. Farshadfar M, Farshadfar E. Genetic variability and path analysis of chickpea (Cicer arientinum L.) landraces and lines. J Appl Sci. 2008;8:3951–6.
34. Wu K, Liu H, Yang M, Tao Y, Ma H, Wu W, et al. High-density genetic map construction and QTLs analysis of grain yield-related traits in Sesame (Sesamum indicum L.) based on RAD-Seq techonology. 2014;:1–14.
35. Gupta PK, Roy JK, Prasad M. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. 2001; February.
36. Gupta PK, Rustgi S, Mir RR. Array-based high-throughput DNA markers for crop improvement. Heredity. 2008.
37. Wei X, Liu K, Zhang Y, Feng Q, Wang L, Zhao Y, et al. Genetic discovery for oil production and quality in sesame. Nat Commun. 2015; May.
38. Cui C, Mei H, Liu Y, Zhang H, Henry R. Genetic Diversity. Population Structure, and Linkage Disequilibrium of an Association-Mapping Panel Revealed by Genome-Wide SNP Markers in Sesame. 2017;8 July:1–10.
39. Basak M, Uzun B, Id EY. Genetic diversity and population structure of the Mediterranean sesame core collection with use of genome-wide SNPs developed by double digest RAD-SEq. 2019;:1–15. doi:10.1371/journal.pone.0223757.
40. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, et al. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A. 2004.
41. Yang S, Pang W, Ash G, Harper J, Carling J, Wenzl P, et al. Low level of genetic diversity in cultivated Pigeonpea compared to its wild relatives is revealed by diversity arrays technology. Theor Appl Genet. 2006.
42. Bolibok-Bragoszewska H, Heller-Uszyńska K, Wenzl P, Uszyński G, Kilian A, Rakoczy-Trojanowska M. DArT markers for the rye genome - genetic diversity and mapping. BMC Genomics. 2009.
43. Sánchez-sevilla JF, Horvath A, Botella MA. Diversity Arrays Technology (DArT) Marker Platforms for Diversity Analysis and Diversity Arrays Technology (DArT) Marker Platforms for Diversity Analysis and Linkage Mapping in a Complex Crop, the Octoploid Cultivated Strawberry (Fragaria × ananassa. 2015; December.
44. Tang J, Daroch M, Kilian A, Jeżowski S, Pogrzeba M, Mos M. DArT-based characterisation of genetic diversity in a Miscanthus collection from Poland. Planta. 2015.

45. Matthies IE, van Hintum T, Weise S, Röder MS. Population structure revealed by different marker types (SSR or DArT) has an impact on the results of genome-wide association mapping in European barley cultivars. Mol Breed. 2012.
46. Laidò G, Mangini G, Taranto F, Gadaleta A, Blanco A, Cattivelli L, et al. Genetic Diversity and Population Structure of Tetraploid Wheats (Triticum turgidum L.) Estimated by SSR, DArT and Pedigree Data. PLoS One. 2013.
47. Tabangin ME, Woo JG, Martin LJ. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc. 2009.
48. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol. 2005.
49. Pham TD, Bui TM, Werlemark G, Bui TC, Merker A, Carlsson AS. A study of genetic diversity of sesame (Sesamum indicum L.) in Vietnam and Cambodia estimated by RAPD markers. Genet Resour Crop Evol. 2009;56:679–90.
50. Seegler CJ. Oil plants in Ethiopia,their taxonomy and agricultural significance. 1983.
51. Bedigian D. Evolution of sesame revisited: Domestication, diversity and prospects. Genetic Resources and Crop Evolution. 2003.
52. Pathak N, Rai AK, Kumari R, Thapa A, Bhat KV. Sesame Crop: An Underexploited Oilseed Holds Tremendous Potential for Enhanced Food Value. Agric Sci. 2014.
53. Valdisser PAMR, Pereira WJ, Filho JEA, Müller BSF, Coelho GRC, Menezes IPP, De, et al. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. 2017;:1–19.
54. Ndjiondjop M, Semagn K, Gouda AC, Kpeki SB. Genetic Variation and Population Structure of Oryza glaberrima and Development of a Mini-Core Collection Using DArTsEq. 2017;8 October:1–19.
55. Zaitoun SYA, Jamous RM, Shtaya MJ, Mallah OB, Eid IS, Ali-shtayeh MS. Characterizing Palestinian snake melon (Cucumis melo var. flexuosus) germplasm diversity and structure using SNP and DArTseq markers. 2018;:1–12.
56. Uncu AO, Frary A, Karlovsky P, Doganlar S. High-throughput single nucleotide polymorphism (SNP) identification and mapping in the sesame (Sesamum indicum L.) genome with genotyping by sequencing (GBS) analysis. Mol Breed. 2016;36.
57. Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, et al. Diversity arrays technology: A generic genome profiling technology on open platforms. Methods Mol Biol. 2012.
58. Dierig D, Ray DT. New crops breeding: Lesquerella. Oil Crop. 2009;:507–16.
59. Alam M, Neal J, Connor KO, Kilian A, Topp B. Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. 2018;:1–20.
60. Kilian A, Huttner E, Wenzl P, Jaccoud D, Carling J, Caig V, et al. The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. Proc Int Congr “In Wake Double Helix From Green Revolut to Gene Revolution”, 27–31 May. 2003; May 2003:443–61.

61. Wei X, Qiao WH, Chen YT, Wang RS, Cao LR, Zhang WX, et al. Domestication and geographic origin of Oryza sativa in China: Insights from multilocus analysis of nucleotide variation of O. sativa and O. rufipogon. Mol Ecol. 2012.
62. Huang XH, Wei XH, Sang T, Zhao QA, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;:42:961–76.
63. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016;:48:927–34.
64. Song XJ, Kuroha T, Ayano M, Furuta T, Nagai K, Komeda N, et al. Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. Proc Natl Acad Sci USA. 2015;:112:76–81.
65. Dossa K, Wei X, Zhang Y, Fonceka D, Yang W, Diouf D. Analysis of Genetic Diversity and Population Structure of Sesame Accessions from Africa and Asia as Major Centers of Its Cultivation. 2016;:1–13.
66. Wei X, Wang L, Zhang Y, Qi X, Wang X, Ding X, et al. Development of Simple Sequence Repeat (SSR) Markers of Sesame (Sesamum indicum) from a Genome Survey. 2014;:5150–62.
67. Genetics. and Genomics of Cotton. 2009.
68. Ali GM. Assessment of genetic diversity in sesame (Sesamum indicum L.) detected by Amplified Fragment Length Polymorphism markers. 2007;10:1–12.
69. Cho Y, Il, Park JH, Lee CW, Ra WH, Chung JW, Lee JR, et al. Evaluation of the genetic diversity and population structure of sesame (Sesamum indicum L.) using microsatellite markers. Genes and Genomics. 2011.
70. Ercan AG, Taskin M, Turgut K. Analysis of genetic diversity in Turkish sesame (Sesamum indicum L.) populations using RAPD markers ⋆ Analysis of genetic diversity in Turkish sesame (Sesamum indicum L.) populations using RAPD markers à¬. 2004; September.
71. Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomics. 2003.
72. Kumar A, Sharma D, Tiwari A, Jaiswal JP, Singh NK, Sood S. Genotyping-by-Sequencing Analysis for Determining Population Structure of Finger Millet Germplasm of Diverse Origins. Plant Genome. 2016.
73. Agrama HA, Tuinstra MR. Phylogenetic diversity and relationships among sorghum accessions using SSRs and RAPDs. African J Biotechnol. 2003.
74. Kim DH, Zur G, Danin-Poleg Y, Lee SW, Shim KB, Kang CW, et al. Genetic relationships of sesame germplasm collection as revealed by inter-simple sequence repeats. Plant Breed. 2002.
75. Zohary D, Hopf M, Weiss E. Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin. 2012.
76. Sharbel TF, Haubold B, Mitchell-Olds T. Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol. 2000;9:2109–18.
77. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 2005;44:1054–

1:1054–1064.

78. Ali GM, Yasumoto S, Seki-Katsuta M. Assessment of genetic diversity in sesame (Sesamum indicum L.) detected by amplified fragment length polymorphism markers. Electron J Biotechnol. 2007;10:12–23.
79. Jarvis DI, Hodgkin T. Wild relatives and crop cultivars: conserving the connection. In: In The Proceedings of an International Symposium on in situ conservation of plant genetic diversity. George Allen & Unwin, London, UK; 1998. p. (pp. 163–179).
80. Jarvis DI, Hodgkin T. (1999). Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems. Mol Ecol. 1999;8:S159–73.
81. Slatkin M. Gene flow and the geographic structure of natural populations. Science (80-). 1987.
82. Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA. Phylogeographic studies in plants: Problems and prospects. Mol Ecol. 1998.
83. Mangini G, Taranto F, Gadaleta A, Blanco A, Cattivelli L, Marone D, et al. Genetic Diversity and Population Structure of Tetraploid Wheats (Triticum turgidum L.) Estimated by SSR, DArT and Pedigree Data. 2013;8.
84. Wijnands J, Biersteker J and HR. Oilseeds Business Opportunities in Ethiopia Survey report, Ministry of Agriculture, Nature and Food Quality. The Netherlands, The Hague. 2007;:8–20.
85. KINDIE AYSHESHM. SESAME MARKET CHAIN ANALYSIS: THE CASE OF METEMA WOREDA, NORTH GONDAR ZONE, AMHARA NATIONAL REGIONAL REGIONAL STATE. Haramaya University; 2007.
86. Wang Y, Abdul M, Rashid R, Li X, Yao C, Lu L, et al. Collection and Evaluation of Genetic Diversity and Population Structure of Potato Landraces and Varieties in China. 2019;10 February:0–10.
87. Courtois B, Audebert A, Dardou A, Roques S, Ghneim-Herrera T, Droc G, et al. Genome-wide association mapping of root traits in a japonica rice panel. PLoS One. 2013.
88. Cruz VMV, Kilian A, Dierig DA. Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop Lesquerella and Related Species. PLoS One. 2013.
89. Raman H, Raman R, Kilian A, Detering F, Carling J, Coombes N, et al. Genome-wide delineation of natural variation for pod shatter resistance in Brassica napus. PLoS One. 2014.
90. Kilian A, Sanewski G, Ko L. The application of DArTseq technology to pineapple. In: Acta Horticulturae. 2016.
91. Huang YF, Poland JA, Wight CP, Jackson EW, Tinker NA. Using Genotyping-By-Sequencing (GBS) for genomic discovery in cultivated oat. PLoS One. 2014.
92. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, et al. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS One. 2013.
93. Bastien M, Sonah H, Belzile F. Genome Wide Association Mapping of Sclerotinia sclerotiorum Resistance in Soybean with a Genotyping-by-Sequencing Approach. Plant Genome. 2014.

94. Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, et al. Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms BT - Data Production and Analysis in Population Genomics: Methods and Protocols. In: Methods in Molecular Biology. 2012.
95. Liu J, Muse SV, PowerMarker. Integrated analysis environment for genetic marker data. Bioinformatics. 2005.
96. Nei M. Genetic distance between populations. Am Nat. 1972;106:283–92.
97. Gao X, Martin ER. Using allele sharing distance for detecting human population stratification. Hum Hered. 2009.
98. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963.
99. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010;10:564–7. Evol Bioinform Online.
100. Pritchard JK, Stephens M, Donnelly P. Structure Software for Population Genetics Inference (V.2.3.4). Pritchard Lab, Stanford University. 2000.
101. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012.

Figures

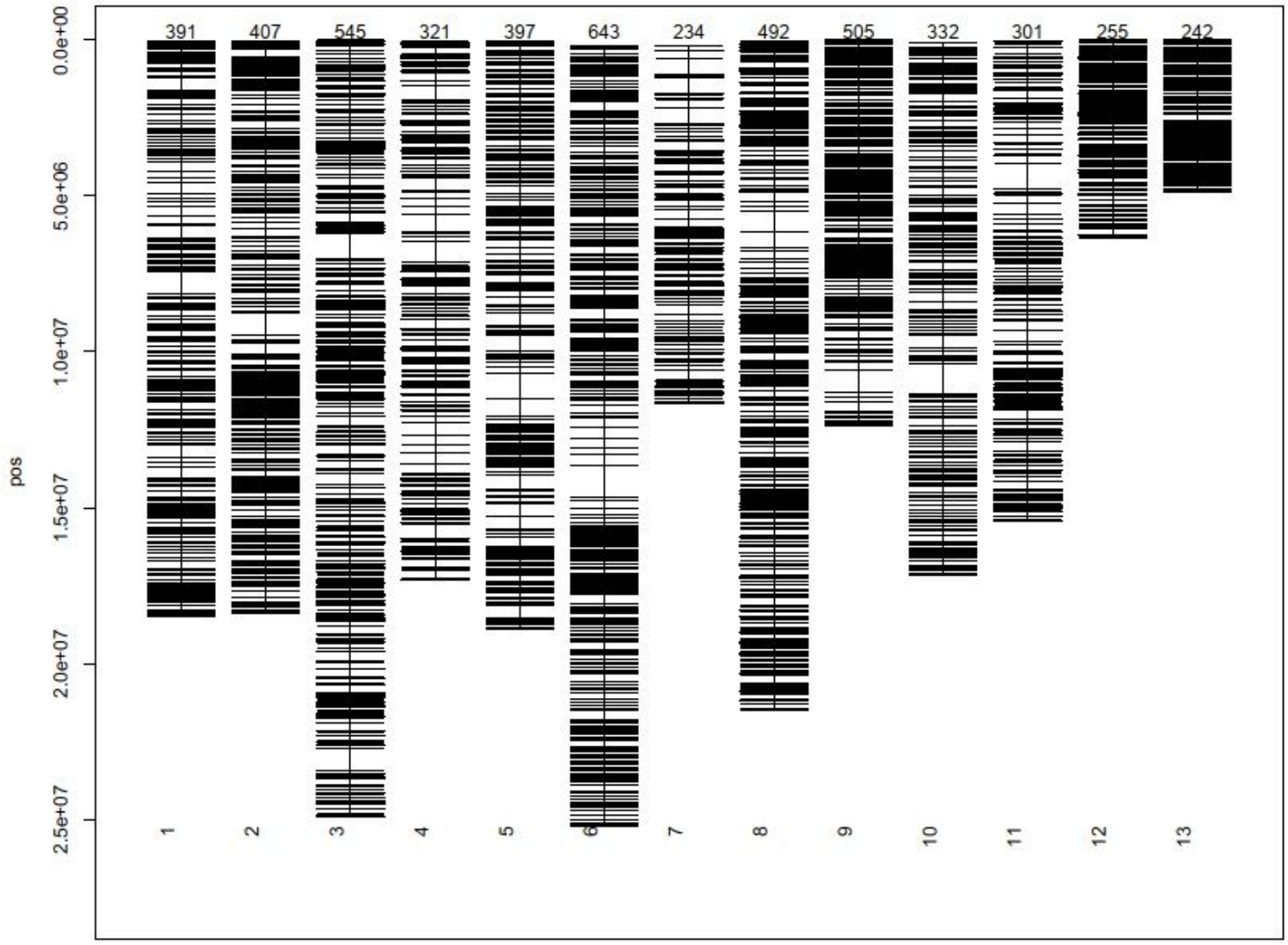

Figure 1

Distribution of DArTseq silicoDArT markers on different chromosomes of sesame

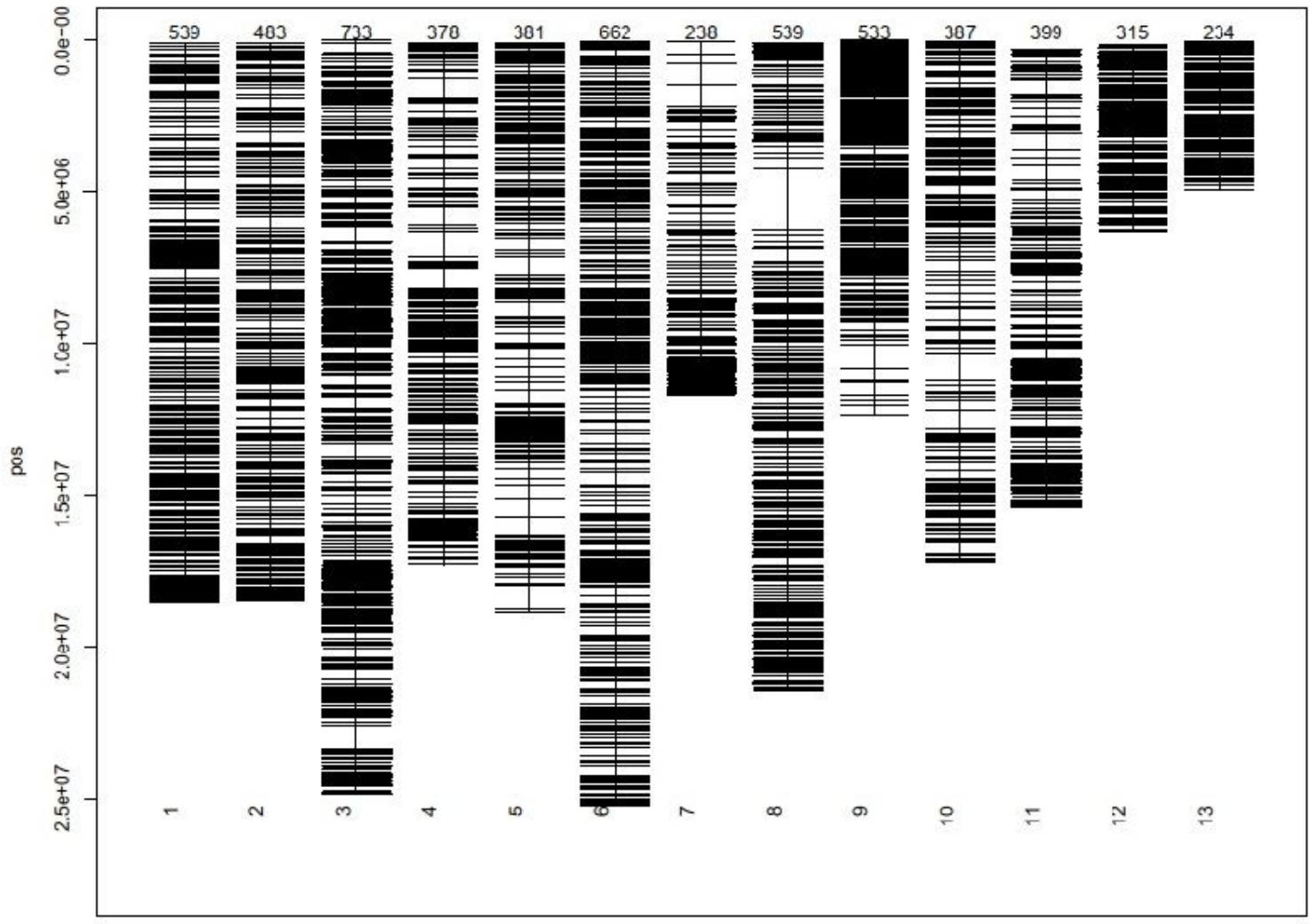

Figure 2

Distribution of DArTseq SNP markers on different chromosomes of sesame

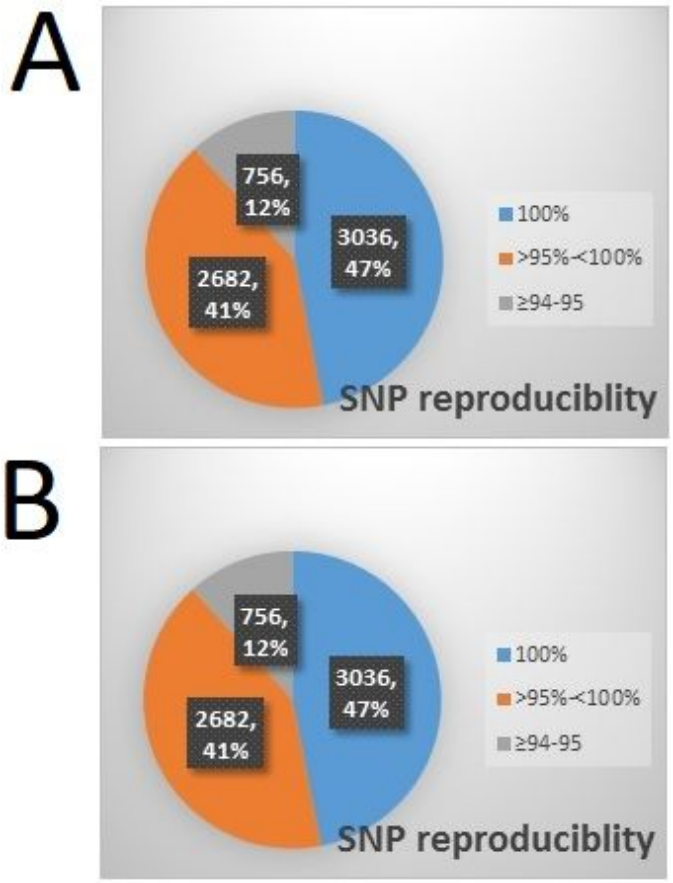

Figure 3

A. Distribution of Reproducibility values of SNP used for genomic studies in Sesame B. Distribution of Reproducibility values of silicoDArT markers used for genomic studies in Sesame

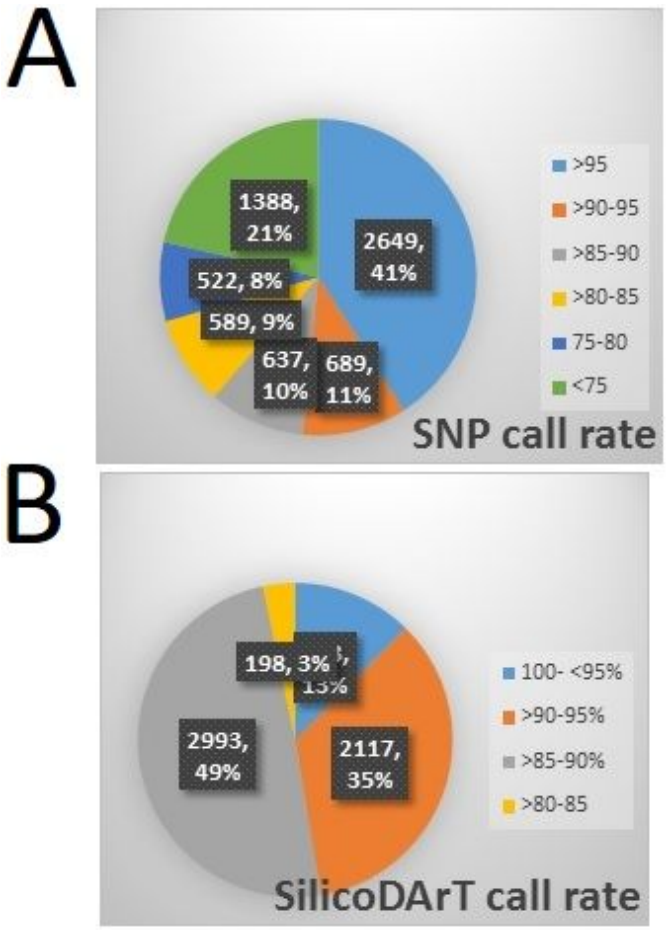

Figure 4

A. Distribution of call rate values of SNP markers used for genomic studies in Sesame. B. Distribution of call rate values of silicoDArT markers used for genomic studies in Sesame.

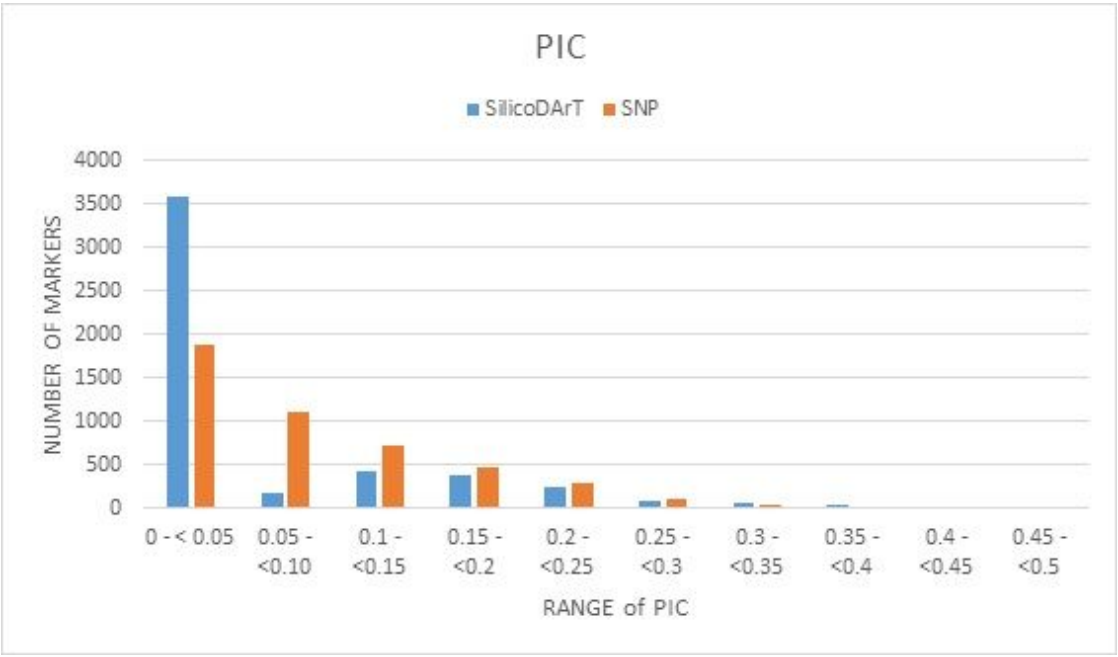

Figure 5

Distribution of PIC values of silicoDArT and SNP markers used for genomic studies in Sesame.

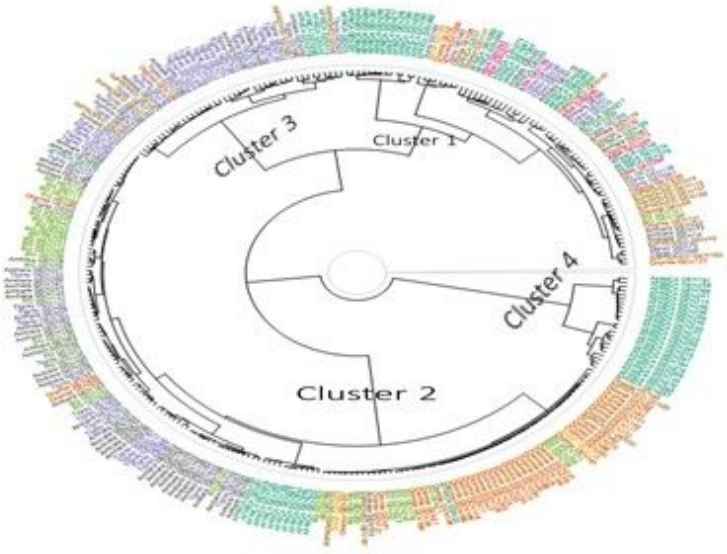

Figure 6

Cluster tree of 300 accessions and genotypes derived from eight geographical origin based on allele sharing genetic distance. Africa, Amhara, Asia, BG: Benshangul-Gumz; Improved, Oromia, SNNP: Southern Nations, Nationalities, and People's Region; Tigray.

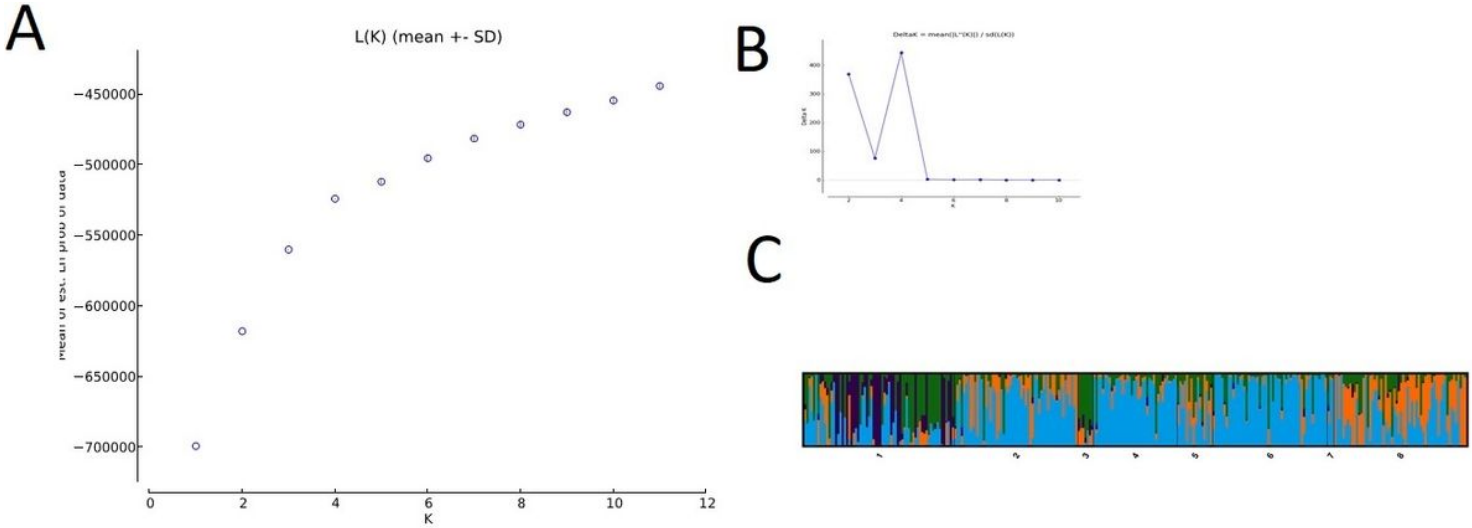

Figure 7

Analysis of the population structure of the 300 sesame accessions using STRUCTURE. (A) Estimated LnP(D) of possible clusters (K) from 1 to 12; (B) ΔK based on the rate of change of LnP(D) between successive K; (C) population structure based on K = 4. In panel (C), each individual is represented by a vertical bar partitioned into four colored segments, with their respective lengths representing the proportion of the individual's genome in a given group K=45/5, Mean (LnProb) = -524093.120, Mean (similarity score) = 0.998

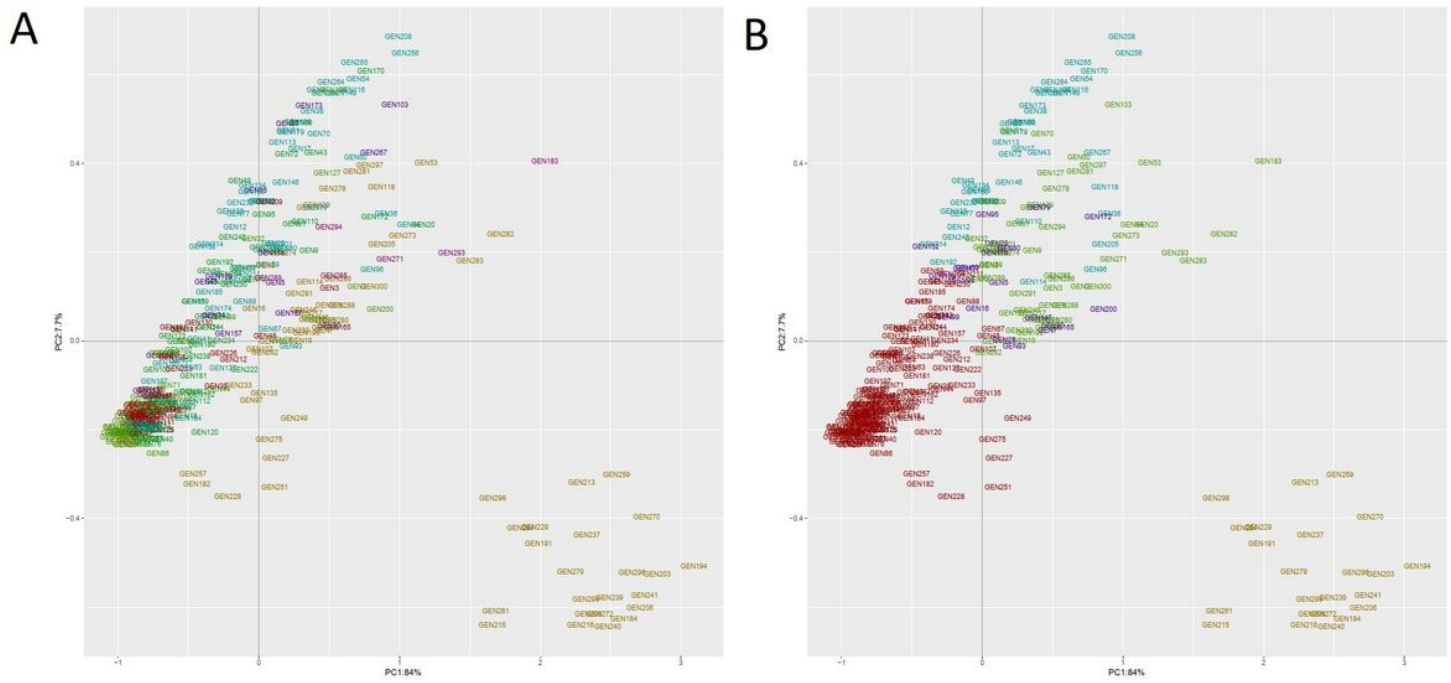

Figure 8

Principal components analysis (PCA) of the population for 300 sesame accessions based on 2997 single-nucleotide polymorphisms (SNPs). Each individual is represented by GEN number, with its symbol color corresponding to the assigned subgroup classification. (A) PCA plots of the sesame germplasm collection based upon their geographic origins. (B) PCA plots of the same sesame germplasm collection but now based on subgroups as identified by STRUCTURE.

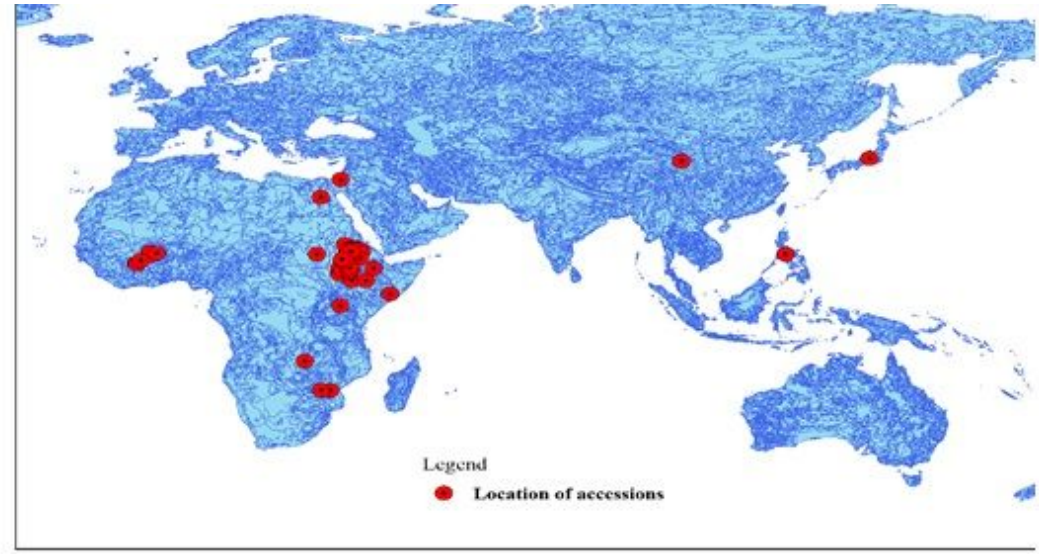

Figure 9

Map of collection areas of sesame genotypes Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research

Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Addationalfile5.xlsx