Published July 10, 2015

ORIGINAL RESEARCH

Genomic Simple Sequence Repeat Markers Reveal Patterns of Genetic Relatedness and Diversity in Sesame

Ayşe Özgür Uncu, Visam Gultekin, Jens Allmer, Anne Frary, and Sami Doganlar*

Abstract

Sesame (*Sesamum indicum* L. syn. *Sesamum orientale* L.) is an orphan crop species with most molecular genetic research work done in the last decade. In this study, we used a pyrosequencing approach for the development of genomic simple-sequence repeat (SSR) markers in sesame. Our approach proved successful in identifying 19,816 nonredundant SSRs, 5727 of which were identified in a contig assembly that covers 19.29% of the sesame genome. Mononucleotide repeats were the most abundant SSR type identified in the sesame genome (48.5% of all SSRs), followed by dinucleotide SSRs (45.0%). Adenine–thymine-rich motifs were predominant, representing 81.7, 51.7, 66.5, and 22.1% of the mononucleotide, dinucleotide, trinucleotide, and tetranucleotide SSRs, respectively. As a result of this work, we introduce 933 experimentally validated sesame specific markers, 849 of which are also applicable in *Sesamum mulayanum* (syn. *Sesamum orientale* var. *malabaricum* Nar.), the wild progenitor of cultivated sesame. Using a subset of the newly identified SSR markers, we analyzed molecular genetic diversity and population structure of a collection of world accessions. Results of the two analyses almost overlapped and suggested correlation between genetic similarity and geographical proximity. Indeed, a pattern of gene flow among sesame diversity centers was apparent, with levels of variability in some regions similar to that seen in the domestication origin of the crop. Taken together with the high rate of genomic marker transferability detected between *S. indicum* and *S. mulayanum*, our results represent additional molecular genetic evidence for designating the two taxa as cultivated and wild forms of the same species.

Published in The Plant Genome 8
doi: 10.3835/plantgenome2014.11.0087

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

SESAME is a member of the Pedaliaceae family and is one of the most ancient oil seed crops (Ashri, 1998; Bedigian, 2003). Sesame is cultivated in tropical, subtropical, and southern temperate regions of the world, but mainly in Asia, Africa, and South America (Anilakumar et al., 2010). Myanmar leads in sesame production with 890,000 tons, followed by India (636,000 tons), China (588,000 tons), Sudan (562,000 tons), and Tanzania (420,000 tons) (FAOSTAT, 2014). While its leaves are edible (Bedigian, 2003), sesame is mainly cultivated for its seeds. Sesame seeds are very nutritious with almost 50% oil and up to 25% protein content (Anilakumar et al., 2010). The species deserves its reputation as "queen of the oil seeds" (Bedigian and Harlan, 1986) because of its oil's resistance to oxidative deterioration and a high, unsaturated fatty acid content of nearly 85%. In addition, polyunsaturated fatty acids constitute more than half of the unsaturated fatty acid fraction in seeds (Anilakumar et al., 2010). The excellent stability of sesame oil is attributed to the presence of antioxidant lignans such as sesamin, sesamolin, and sesaminol (Abou-Gharbia et al., 2000). The health benefits of these compounds, including antioxidant, antiaging, antihypertensive, anticancer, cholesterol lowering, and antimutagenic properties are reported by several authors (Anilakumar et al., 2010).

Because sesame can set seed without significant yield loss under high temperature and drought, its cultivation is feasible and relatively easy in regions with such climatic conditions. However, sesame is the least productive oil seed crop and is mainly grown by small holders (Bhat et al., 1999; Bisht et al., 2004). Sesame productivity is

Izmir Inst. of Technology, Dep. of Molecular Biology & Genetics, Urla, Izmir 35430 Turkey. Received 25 Nov. 2014. Accepted 8 Jan. 2015. *Corresponding author (samidoganlar@iyte.edu.tr).

Abbreviations: AARI, Aegean Agricultural Research Institute; EST, expressed sequence tag; Mb, megabase; PCR, polymerase chain reaction; SSR, simple-sequence repeat.

constrained by growth-habit traits such as seed shattering, nonsynchronous ripening, and indeterminate growth, and by diseases caused by the bacterium *Pseudomonas syringae* pv. and the fungi *Cerocospora sesame* and *Alternaria sesame*. Wild relatives are proposed as a potential source of pest and microbe resistance alleles (Kawase, 2000). Thus, inclusion of wild accessions with reproductive compatibility into sesame breeding programs will be an effective strategy to improve biotic stress tolerance characters and broaden the breeding potential of sesame.

While integration of molecular marker technologies have significantly improved the speed and precision of modern plant breeding, molecular genetic research in sesame has lagged behind other crops, restricting the use of molecular breeding in the crop. The onset of DNA-based research in sesame is relatively recent with the first report describing the use of the random amplified polymorphic DNA technique for sesame germplasm characterization (Bhat et al., 1999). However, no sesame-specific markers were reported until the development of the first SSR markers by Dixit et al. (2005). According to one common definition, SSRs are iterations of 1 to 6 bp DNA motifs (Jones et al., 2009). However, this definition varies with some authors defining minimum motif length as 2 bp (Brown et al., 1996), and some maximum motif length as 5 (Powell et al., 1996), 7 (Jannati et al., 2009), or 8 bp (Wu et al., 2014). Simple sequence repeats are widely used in plant molecular genetic studies because they are hypervariable, reproducible, relatively abundant, and provide extensive genome coverage (Powell et al., 1996). Reports on SSR marker development in sesame are limited and most work involved the development of genic SSRs, either from expressed sequence tag (EST) sequences in public databases (Wei et al., 2008; Yepuri et al., 2013) or transcriptome sequencing (Wei et al., 2011; Zhang et al., 2012; Wu et al., 2014). Fewer authors described genomic SSRs in sesame (Dixit et al., 2005; Spandana et al., 2012; Wei et al., 2014; Surapaneni et al., 2014). While genic markers have their own advantages such as lower development costs and a high degree of interspecific conservation, an abundance of genomic markers is a prerequisite for the equal representation of both gene-rich and noncoding regions in linkage groups. In addition, genomic markers are extremely useful to establish core collections that sufficiently represent the extent of allelic diversity (Frary et al., 2014). Due to their highly polymorphic nature, genomic SSRs also allow distinguishing among closely related genotypes and are of great use in hybridity and purity tests.

The primary goal of this work was to develop genomic SSR markers in sesame using pyrosequencing technology. The developed markers were tested on cultivated sesame and its putative progenitor *S. mulayanum* (syn. *S. orientale* var. *malabaricum* Nar.) (Bedigian, 2003) to validate amplification efficiency and assess transferability. The SSR markers were also used to analyze the molecular genetic diversity and population structure of a sesame collection comprising 94 accessions from 38 countries. A total of 24 Turkish sesame accessions were included in the analyses to assess the genetic relationships between Turkish and foreign germplasm and test the potential of our markers to measure diversity in relatively narrow, local gene pools.

MATERIALS AND METHODS

Plant Material and DNA Isolation

A collection of *S. indicum* accessions, consisting of 78 landraces and 15 cultivars, was used as plant material (Table 1). In addition, the wild accession *S. mulayanum* was included in analyses. The majority of the accessions (74 accessions), were obtained from the USDA Plant Genetic Resources Conservation Unit, Griffin, GA, USA, and originated from 38 countries including Turkey. Among the remaining accessions, 17 were obtained from Aegean Agricultural Research Institute, Izmir, Turkey (AARI). All accessions provided by AARI were Turkish accessions, including five cultivars released by AARI (Cumhuriyet 99, Kepsut 99, Orhangazi 99, Osmanli 99, and Tan 99) and three landraces (Golmarmara, Muganli 57, and Ozberk) released by West Mediterranean Agricultural Research Institute (BATEM), Antalya, Turkey. In addition, two accessions from Africa (95-223) and Korea (92-3091) were contributed by Dr. Petr Karlovsky, University of Gottingen, Gottingen, Germany. *Sesamum mulayanum* acc. COL/INDIA/1992/MAFF/0161 seeds were obtained from Dr. Makoto Kawase, National Institute of Agrobiological Sciences, Japan.

Ten seeds from each accession were planted and grown in soil containing peat moss, perlite, and natural fertilizer. Plants were grown in the growth chamber at 23 to 25°C, 18-h photoperiod, and approximately 35% humidity at Izmir Institute of Technology, Turkey. Genomic DNA from each accession was isolated from liquid nitrogen-frozen ground leaf tissue pooled from 10 plants harvested at the two- to four-leaf stage. DNA extraction was done using the Wizard Magnetic 96 Plant System (Promega Corp.) with the Biomek NX Workstation (Beckman Coulter) according to the manufacturer's instructions.

DNA Sequencing and Simple Sequence Repeat Validation

For SSR identification, total genomic DNA of *S. indicum* 'Muganli 57' was subjected to pyrosequencing. Pyrosequencing was done with a Roche 454 GS-FLX sequencer (Roche Diagnostics) and performed by 454 Lifesciences Corp. (Branford, CT, USA). Simple sequence repeat validation was done using the dye-terminator sequencing method. Polymerase chain reaction (PCR) products, purified with the DNA Clean & Concentrator–5 Kit (Zymo Research), were used as template in the dye-terminator sequencing reaction, prepared using GenomeLab DTCS Quick Start Kit (Beckman Coulter) according to the manufacturer's instructions. Sequencing reaction thermal cycling conditions were 30 cycles of

Table 1. Sesame material used in the study. Accessions are listed in order of analysis. Cluster assignments according to Structure and DARwin analyses are presented in the last four columns.

Genotype (source)†	Plant introduction	Origin	Landrace (cultivar)	Inferred ancestry subpopulation		Subpopulation assignment‡	Cluster assignment§
				1	2		
1 (US)	PI167115	Turkey, Adana	Landrace	0.843	0.157	1	A1
2 (US)	PI161385	Korea	Landrace (Kyorgii Do)	0.749	0.251	1	A2
3 (US)	PI154298	Mexico	Landrace	0.794	0.206	1	A2
4 (US)	PI250099	Egypt	Landrace (Simsim)	0.972	0.028	1	A1
5 (US)	PI543241	Bolivia	Landrace (Chinchilin negro)	0.026	0.974	2	B
6 (US)	PI229668	Argentina	Landrace	0.462	0.538	Admixed	B
7 (US)	PI263441	Japan, Honshu	Landrace (Ban-4)	0.666	0.334	1	A1
8 (US)	PI304259	Thailand	Landrace	0.015	0.985	2	B
9 (US)	PI207665	Morocco	Landrace	0.231	0.769	2	B
10 (US)	PI490024	Thailand	Landrace (Nga Khaw buk)	0.064	0.936	2	B
11 (US)	PI234427	China	Landrace (Tainan White No. 1)	0.771	0.229	1	A2
12 (US)	PI433863	Nigeria	Landrace	0.481	0.519	Admixed	C
13 (US)	PI239001	Greece, Rhodes	Landrace	0.897	0.103	1	A1
14 (US)	PI323306	Pakistan	Landrace	0.03	0.97	2	B
15 (US)	PI251294	Jordan	Landrace	0.962	0.038	1	A1
16 (US)	PI254698	South America	Landrace	0.933	0.067	1	A1
17 (US)	PI198158	Former USSR	Landrace (Roussi)	0.275	0.725	2	B
18 (US)	PI179485	Iraq	Landrace	0.369	0.631	2	B
19 (US)	PI158769	Venezuela	Landrace (Venezuela-51)	0.687	0.313	1	A2
20 (US)	PI226567	Ethiopia	Landrace	0.336	0.664	2	B
21 (US)	PI601234	United States	Cultivar (Sesaco 7)	0.792	0.208	1	A2
22 (US)	PI198156	Iraq	Landrace (Mahalli No. 8)	0.786	0.214	1	A1
23 (US)	PI561704	Mexico	Cultivar (Ostimuri 89)	0.655	0.345	1	A1
24 (US)	PI200428	Pakistan	Landrace	0.028	0.972	2	B
25 (US)	PI490114	Sudan	Cultivar (Zira 37)	0.533	0.467	Admixed	C
26 (US)	PI186511	Nigeria	Landrace	0.588	0.412	Admixed	C
27 (US)	PI211627	Afghanistan	Landrace	0.957	0.043	1	A1
28 (US)	PI231033	Mozambique	Landrace (Almadnagor White)	0.01	0.99	2	B
29 (US)	PI164142	India	Landrace (Til)	0.054	0.946	2	B
30 (US)	PI184671	Liberia	Landrace (Dumboy)	0.264	0.736	2	B
31 (US)	PI306695	India	Landrace	0.268	0.732	2	B
32 (US)	PI207664	Morocco	Landrace	0.551	0.449	Admixed	B
33 (US)	PI250029	Iran	Landrace	0.884	0.116	1	A1
34 (US)	PI229667	Argentina	Landrace	0.376	0.624	2	B
35 (US)	PI250030	Iran	Landrace	0.872	0.128	1	A1
36 (US)	PI153509	Venezuela	Landrace (Criollo)	0.513	0.487	Admixed	B
37 (US)	PI158038	China	Landrace	0.626	0.374	1	A2
38 (US)	PI203150	Jordan	Landrace	0.935	0.065	1	A2
39 (US)	PI189082	Cameroon	Landrace	0.434	0.566	Admixed	C
40 (US)	PI643459	Tajikistan	Landrace	0.873	0.127	1	A1
41 (US)	PI258372	Former USSR	Cultivar	0.742	0.258	1	A1
42 (US)	PI200427	Pakistan	Landrace	0.051	0.949	2	B
43 (US)	PI209965	Ethiopia	Landrace	0.616	0.384	1	A
44 (US)	PI599444	United States	Cultivar (Calinda)	0.823	0.177	1	A2
45 (US)	PI234424	China	Landrace (Tainan Black No. 1)	0.46	0.54	Admixed	B
46 (US)	PI195122	China	Landrace (Par-wang-pien)	0.015	0.985	2	B
47 (US)	PI254705	United States	Landrace	0.954	0.046	1	A1
48 (US)	PI157155	India	Landrace	0.736	0.264	1	B
49 (US)	PI207667	Morocco	Landrace	0.328	0.672	2	B
50 (US)	PI198155	Egypt	Landrace (Giza No. 10)	0.987	0.013	1	A1
51 (US)	PI211088	Afghanistan	Landrace	0.851	0.149	1	A1

(cont'd)

Table 1. Continued.

Genotype (source)†	Plant introduction	Origin	Landrace (cultivar)	Inferred ancestry subpopulation		Subpopulation assignment‡	Cluster assignment§
				1	2		
52 (US)	PI231034	Mozambique	Landrace (Branco)	0.052	0.948	2	B
53 (US)	PI156618	China	Landrace	0.817	0.183	1	A1
54 (US)	PI490072	Korea, South	Cultivar (Suweon 33)	0.801	0.199	1	A2
55 (US)	PI253984	Syria	Landrace	0.904	0.096	1	A1
56 (US)	PI186509	Nigeria	Landrace	0.502	0.498	Admixed	C
57 (US)	PI210687	Somalia	Landrace	0.046	0.954	2	B
58 (US)	PI189081	Cameroon	Landrace	0.834	0.166	1	A1
59 (US)	PI238988	Greece, Rhodes	Landrace	0.971	0.029	1	A1
60 (US)	PI189229	Belgian Congo	Landrace	0.017	0.983	2	B
61 (US)	PI163595	Guatemala	Landrace	0.18	0.82	2	B
62 (US)	PI321096	Kenya	Landrace	0.467	0.533	Admixed	C
63 (US)	PI253424	Israel	Landrace	0.958	0.042	1	A1
64 (US)	PI251704	Former USSR	Landrace (Kohditersky 2058)	0.796	0.204	1	A1
65 (US)	PI224663	Libya	Landrace	0.476	0.524	Admixed	B
66 (US)	PI288852	Nepal	Landrace	0.03	0.97	2	B
67 (US)	PI238430	Turkey, Izmir	Landrace (Kirmizi Susam)	0.98	0.02	1	A1
68 (US)	PI200106	Myanmar	Landrace (Baktaung)	0.512	0.488	Admixed	B
69 (US)	PI254703	Venezuela	Landrace (Acarigua Selection)	0.287	0.713	2	B
70 (AA)	TR38356	Turkey, Tekirdag	Landrace	0.973	0.027	1	A1
71 (AA)	Golmarmara	Turkey	Registered landrace	0.933	0.067	1	A1
72 (AA)	Ozberk	Turkey	Registered landrace	0.964	0.036	1	A1
73 (AA)	Tan 99	Turkey	Cultivar	0.988	0.012	1	A1
74 (AA)	Cumhuriyet 99	Turkey	Cultivar	0.991	0.009	1	A1
75 (AA)	Osmanli 99	Turkey	Cultivar	0.953	0.047	1	A1
76 (AA)	Kepsut 99	Turkey	Cultivar	0.976	0.024	1	A1
77 (AA)	Orhangazi 99	Turkey	Cultivar	0.961	0.039	1	A1
78 (US)	PI177072	Turkey, Eskisehir	Landrace	0.978	0.022	1	A1
79 (US)	PI170753	Turkey, Canakkale	Landrace	0.962	0.038	1	A1
80 (AA)	PI238431	Turkey, Manisa	Landrace (Sari Susam)	0.652	0.348	1	A1
81 (US)	PI167248	Turkey, Adana	Landrace	0.979	0.021	1	A1
82 (US)	PI205229	Turkey, Izmir	Landrace	0.989	0.011	1	A1
83 (AA)	PI238481	Turkey, Adiyaman	Landrace	0.987	0.013	1	A1
84 (AA)	PI238420	Turkey, Izmir	Landrace	0.991	0.09	1	A1
85 (AA)	PI238445	Turkey, Manisa	Landrace	0.971	0.029	1	A1
86 (AA)	PI238450	Turkey, Manisa	Landrace (Beyaz Susam)	0.966	0.034	1	A1
87 (AA)	PI238433	Turkey, Mersin	Landrace	0.988	0.012	1	A1
88 (AA)	PI240844	Turkey, Mersin	Landrace	0.986	0.014	1	A1
89 (US)	PI205225	Turkey, Antalya	Landrace	0.98	0.02	1	A1
90 (AA)	PI238453	Turkey, Canakkale	Landrace	0.985	0.015	1	A1
91 (UG)	95-223	Africa	Landrace	0.361	0.639	2	B
92 (UG)	92-3091	Korea	Landrace	0.732	0.268	1	A
93 (AA)	Muganli 57	Turkey	Registered landrace	0.973	0.027	1	A1
94 (NI)	*S. mulayanum*	India	*S. mulayanum*	0.354	0.646	2	B

† Seed sources are coded: US, USDA; AA, Aegean Agricultural Research Institute; UG, University of Gottingen; NI, National Institute of Agrobiological Sciences.

‡ Accessions were assigned to subpopulations based on the proportion of inferred ancestry with a threshold of $\geq$0.60.

§ Cluster assignments based on the neighbor-joining dendrogram are displayed.

96°C 20 sec, 50°C 20 sec, 60°C 4 min. The reaction mixture for each SSR amplicon was then purified using ZR DNA Sequencing Clean-up Kit (Zymo Research), resuspended in 30 μL of sample loading solution (Beckman Coulter) and run on a Beckman CEQ8800 capillary electrophoresis device using the LFR-c method (injection voltage 2.0 kV for 10–15 sec, separation temperature 60°C, separation voltage 7.4 kV, separation time 45 min).

Pyrosequencing Data Processing and Sequence Assembly

Adaptor and linker sequences were removed from the raw sequence reads to facilitate genome assembly. Because most assembly tools cannot directly process standard flowgram format (SFF) files, SFF data were converted to separate FASTA (Lipman and Pearson, 1985) and quality files. The conversion was performed using an open source package of tools written in Python language (http://bioinf.comav.upv.es/seq_crumbs/download.html). The seq_crumbs tool from the package was used to perform the conversion with the default settings. The resulting FASTA and FASTQ format files were suitable for sequence assembly. MIRA version 3.4, a whole genome shotgun and EST sequence assembler (Chevreux et al., 2004), was used for sequence assembly. Assembly quality was based on various parameters, such as the weighted median of contig lengths (N50), a commonly used measure. The most successful assembly among more than 100 trials used nondefault parameters. Customized sequence assembly parameters are provided in Supplementary Table S1. The assembled sequences are available through the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/bioproject/271288).

Simple Sequence Repeat Detection and Primer Design

Contig assemblies and singleton sequences were analyzed for SSR identification with our in-house tool SiSeeR (http://bioinformatics.iyte.edu.tr/index.php?*n*=Softwares.SiSeeR). The minimum number of repeats required to identify perfect SSRs were 10 for mononucleotide, four for dinucleotide, and three for motifs comprised of three or more nucleotides. Primer design was performed with the Primer3 (Rozen and Skaletsky, 2000; http://frodo.wi.mit.edu/) console application. A total of 5054 contig sequences yielding 5727 nonredundant SSRs were converted from FASTA to the default Primer3 input format Boulder-IO. The Primer3 settings, customized to meet the requirements of SSR primer design, are provided in Supplementary Table S1. To produce primers flanking the SSR sequences, values for the start and end positions of each SSR were generated by enabling the SEQUENCE_TEMPLATE switch of the software.

Simple Sequence Repeat Amplification

Simple sequence repeat alleles were amplified in 20 μL reaction mixtures containing 1× PCR buffer, 1.5 mM $MgCl_2$, 0.25 mM of each deoxyribonucleotide triphosphate (Promega Corp.), 1 U Taq polymerase, 0.25 μM of each primer, and 50 ng template DNA. Thermal cycling conditions consisted of one cycle of initial denaturation for 10 min at 94°C, followed by 35 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 45 sec, with a final extension step of 10 min at 72°C. Polymerase chain reaction products were then run on a Fragment Analyzer (Advanced Analytical) capillary electrophoresis system using the DNF-900 dsDNA Reagent Kit (Advanced Analytical) according to the manufacturer's instructions. Simple sequence repeat alleles were visualized and scored using the PROSize 2.0 software version 1.2.1.1 (Advanced Analytical) (Supplementary Figure S2).

Simple Sequence Repeat Data Analysis

Simple sequence repeat alleles were scored as present (1) or absent (0). Average gene diversity (Nei, 1973) was calculated for each SSR marker according to the formula: average gene diversity = $\left(\sum_{i=1}^{n} 2 f_i \left(1 - f_i\right)\right) / n$ (Roldan-Ruiz et al., 2000), where f_i is the frequency of band presence for the ith allele and n is the number of alleles. Marker data were used to infer population structure and analyze molecular genetic diversity. Using the Structure computer program (Pritchard et al., 2000), models with 1 to 20 subpopulations (K) were tested for 20 iterations. Burn-in period and number of Monte Carlo Markov Chain repeats were 50,000 and 300,000, respectively. Structure Harvester program (Earl and Von Holt, 2012) was used to calculate ΔK values for each model based on posterior probabilities. The model with the highest ΔK was selected as the best. Inferred ancestry threshold was set as $\geq$0.60, to assign the accessions to subpopulations. Accessions with lower probabilities were assigned to the admixed group. The DARwin (http://darwin.cirad.fr/product.php) computer program was used to generate a Dice coefficient dissimilarity matrix that was then used to construct an unweighted neighbor-joining dendrogram of the accessions. Correlation of the dissimilarity matrix and the dendrogram was demonstrated with a Mantel test.

RESULTS

Sequence Assembly and Simple Sequence Repeat Identification

A total of 1,094,317 sequence reads, covering more than 623 megabases (Mb), were obtained from the sesame cultivar Muganli 57, using the Roche 454 GS-FLX (Roche Diagnostics) sequencing system. Removal of the adaptor and linker sequences resulted in a total cleaned sequence length of nearly 381 Mb. The average length of the raw reads (569 ± 76.5 nucleotides) was reduced to 348 ± 125.6 nucleotides after cleaning. After adaptor and linker removal, 616,210 reads (56.3% of the cleaned reads) could be assembled into 136,257 contigs (Table 2). The assembly encompassed nearly 65 Mb of the sesame genome, corresponding to 19.3% genome coverage, based on genome size estimation by flow cytometry (337 Mb) (Wang et al., 2014). The weighted median contig length, N50, of the assembly was 671 nucleotides.

Contigs and singletons were mined for SSRs, resulting in the identification of 5727 and 14,089 non-redundant SSRs in contigs and singleton sequences, respectively. SSR length ranged between 8 and 394 nucleotides, with an average length of 20.4 ± 0.17 nucleotides. Among the 19,816 SSRs identified, mononucleotide repeats were

Table 2. Sequence preprocessing and assembly statistics.

Parameter	Raw sequences	Cleaned sequences	Contigs
Total number of sequences	1,094,317	1,094,317	616,210 (136,257 contigs)
Minimum sequence length (nt)	47	40	40
Maximum sequence length (nt)	1200	900	53,745
Average sequence length (nt)	569 ± 76.5	348 ± 125.6	474 ± 680
Total number of bases	623,365,931	380,862,690	64,674,100

Table 3. Simple sequence repeat types in sesame genome.

Motif length	Number of occurrences	Frequency (%)
Mononucleotide	9611	48.5
Dinucleotide	8924	45
Trinucleotide	492	2.5
Tetranucleotide	86	0.4
Pentanucleotide	72	0.4
Hexanucleotide	378	1.9
Heptanucleotide	157	0.8
Octanucleotide	96	0.5
Total	19,816	100

the most abundant, representing 48.5% of all SSRs (Table 3). Dinucleotide repeats were the second most common SSR type and represented 45.0% of all SSRs. The sum of mono- and dinucleotide repeats alone constituted 93.5% of all SSRs. The percentage of abundance of the remaining repeat types ranged between 0.4 (tetra- and pentanucleotide repeats) and 2.5% (trinucleotide repeats). While A/T was the predominant mononucleotide repeat (81.7%), AT was the most abundant dinucleotide repeat (32.5%), followed by TA (19.2%). Also, AT-rich repeats were prevalent for tri- and tetra- nucleotide repeats with AAT/ATT (27.2%) and AAAT/ATTT (11.6%) representing the most abundant trinucleotide and tetranucleotide repeats, respectively (Table 4).

Primer Design and Simple Sequence Repeat Validation

For successful SSR primer design, flanking sequences of sufficient length should be present. Compared with singletons, contigs usually provide longer flanking sequences and, therefore, allow greater flexibility in the primer design process. Thus, to improve the efficiency of primer design and ensure a high rate of successful PCR amplification, primers were designed only for the SSRs identified in contigs. Of the 5727 SSRs identified in contigs, 2465 SSRs met the requirements for primer design. We tested 1000 of the designed primers for their amplification efficiency, using a cultivated (Muganli 57) and a wild (*S. mulayanum*) sesame accession. A total of 933 primers (93.3%) successfully amplified PCR products from *S. indicum* while 849 (84.9%) amplified products from both genotypes. Among the 849 markers,

Table 4. Most abundant simple sequence repeat (SSR) motifs.

SSR motif	Number of SSRs	Motif frequency† (%)
A/T	7852	81.7
C/G	1759	18.3
AT	2900	32.5
TA	1716	19.2
AG/CT	1450	16.2
TC/GA	1185	13.3
AAT/ATT	134	27.2
ATA/TAT	114	23.2
TTA/TAA	79	16.1
AAAT/ATTT	10	11.6
ATAC/GTAT	9	10.5
AAACCCT/AGGGTTT	36	22.9
CCCTAAA/TTTAGGG	21	13.4
GGGTTTA/TAAACCC	16	10.2

† Motif frequencies are relative to SSR types. Only motifs with a frequency ≥0.10 are listed.

228 (26.9%) were polymorphic between *S. indicum* and *S. mulayanum*. To validate the presence of the expected SSR motifs within the amplicons, eight amplicons from Muganli 57 were sequenced with the dye-terminator method (data not shown). All eight sequences contained the expected SSR motifs, proving the identity of our primers as SSR markers. Primer and transferability information for the SSR markers is available at http://plantmolgen.iyte.edu.tr/data/.

Genetic Diversity and Population Structure Analyses of a Sesame Collection using Simple Sequence Repeat Markers

A total of 50 SSR markers, selected according to their amplification efficiency based on the peak heights of their capillary electropherograms, were applied to 94 sesame accessions from throughout the world. Asia, the Indian subcontinent, the Middle East, Africa, and the Americas were represented by 15, 8, 34, 19, and 13 accessions, respectively (Table 1). Because 24 Turkish accessions were included in analyses, the Middle East had the highest number of representative accessions. The proposed progenitor of cultivated sesame *S. mulayanum* (Bedigian, 2003) was also included in analyses. Except for one marker, which was subsequently excluded from analysis, all markers yielded high-quality, reproducible fragments. When applied on the sesame accessions, these 49 markers produced a total of 219 alleles. Only two of 49 markers were monomorphic, while the remaining 47 markers were polymorphic and amplified a total of 217 alleles, 215 (99%) of which were polymorphic (Table 5). The average number of alleles produced by the SSR markers was 4.5, with the highest number of alleles (17 alleles) produced by marker siSSR-621. The average gene diversity value of the markers was intermediate (0.20), with the highest value calculated for siSSRg-575 (0.48 ± 0.02), and the lowest (zero) calculated for the two non-polymorphic markers (siSSRg-48 and siSSRg-933) (Table 5).

Table 5. Simple sequence repeat (SSR) markers used for the molecular genetic analysis.

SSR marker	Repeat motif (5′ to 3′)	Number of alleles	Gene diversity†
SiSSRg-1	(GTG/CAC)4	5	0.30 ± 0.08
SiSSRg-4	(ATTT/AAAT)4	3	0.18 ± 0.05
SiSSRg-17	(ATG/CAT)7	2	0.32 ± 0.02
SiSSRg-42	(GAG/CTC)5	4	0.18 ± 0.09
SiSSRg-45	(TGG/CCA)4	2	0.14
SiSSRg-47	(CT/AG)7	3	0.08 ± 0.02
SiSSRg-48	(AGA/TCT)4	1	0
SiSSRg-51	(TATG/CATA)3	4	0.28 ± 0.13
SiSSRg-111	(AATT)3	2	0.39 ± 0.03
SiSSRg-178	(TTG/CAA)5	7	0.17 ± 0.06
SiSSRg-223	(TTA/TAA)6	3	0.30 ± 0.14
SiSSRg-236	(TAAA/TTTA)3	3	0.34 ± 0.14
SiSSRg-346	(CCA/TGG)5	4	0.21 ± 0.06
SiSSRg-392	(CCCCA/TGGGG)4	4	0.22 ± 0.09
SiSSRg-393	(CAA/TTG)4	5	0.21 ± 0.10
SiSSRg-410	(CT/AG)11	5	0.18 ± 0.09
SiSSRg-422	(AT)11	6	0.19 ± 0.08
SiSSRg-437	(GTTTT/AAAAC)3	7	0.18 ± 0.07
SiSSRg-485	(TG/CA)8	6	0.20 ± 0.07
SiSSRg-491	(TTAT/ATAA)3	4	0.22 ± 0.07
SiSSRg-549	(TACA/TGTA)3	4	0.26 ± 0.11
SiSSRg-575	(ATGT/ACAT)5	2	0.48 ± 0.02
SiSSRg-606	(GGAGTA/TACTCC)4	6	0.21 ± 0.08
SiSSRg-621	(TA)6	17	0.11 ± 0.03
SiSSRg-634	(GGGGT/ACCCC)3	5	0.23 ± 0.11
SiSSRg-635	(TGATT/AATCA)3	2	0.01 ± 0.01
SiSSRg-640	(AT)6	4	0.26 ± 0.13
SiSSRg-654	(TTTC/GAAA)3	7	0.14 ± 0.06
SiSSRg-666	(ATGA/TCAT)3	9	0.12 ± 0.05
SiSSRg-670	(TG/CA)7	4	0.31 ± 0.08
SiSSRg-679	(CTTTT/AAAAG)3	3	0.20 ± 0.09
SiSSRg-692	(GCA/TGC)4	9	0.13 ± 0.07
SiSSRg-708	(AATT)3	6	0.23 ± 0.08
SiSSRg-733	(GT/AC)8	3	0.06 ± 0.02
SiSSRg-767	(AT)7	2	0.02
SiSSRg-786	(TAG/CTA)4	4	0.11 ± 0.06
SiSSRg-801	(TGAAA/TTTCA)4	3	0.34 ± 0.16
SiSSRg-825	(CTCCGC/GCGGAG)3	5	0.04 ± 0.01
SiSSRg-859	(ACTCAC/GTGAGT)3	2	0.04 ± 0.04
SiSSRg-863	(TAT/ATA)4	4	0.20 ± 0.07
SiSSRg-892	(AT)10	10	0.16 ± 0.04
SiSSRg-924	(AT)7	4	0.29 ± 0.07
SiSSRg-925	(ATC/GAT)4	5	0.22 ± 0.11
SiSSRg-933	(TAC/GTA)4	1	0
SiSSRg-945	(TGATCA)4	2	0.34 ± 0.06
SiSSRg-949	(GAA/TTC)4	4	0.14 ± 0.04
SiSSRg-975	(TC/GA)7	3	0.32 ± 0.14
SiSSRg-985	(TA)7	7	0.13 ± 0.06
SiSSRg-991	(TC/GA)7	2	0.31 ± 0.04

† For each marker, average gene diversity ± standard error is presented.

The average gene diversity value of tetranucleotide SSRs (0.26) was higher than that of di- (0.19), tri- (0.17), penta- (0.20), and hexanucleotide (0.16) SSRs. However, no statistically significant correlation was observed between gene diversity and repeat lengths of the markers.

Simple sequence repeat marker data were used to assess the population structure and molecular genetic diversity of the sesame accessions using the computer programs Structure and DARwin, respectively. Results from the two analyses were compared to determine the relatedness and diversity of the accessions. Population structure analysis suggested a model that assigned the accessions to two subpopulations (Supplementary Figure S1). As a result, 57 and 25 accessions were assigned to Subpopulations 1 and 2, respectively, and 12 accessions were considered as admixed (Table 1). While all of the Turkish accessions and the majority of accessions from the Middle East (nine of 10 accessions) were assigned to the same cluster (Subpopulation 1), Asian accessions were distributed to both subpopulations and the admixed accessions. With the exception of one Indian accession in Subpopulation 1, all accessions from the Indian subcontinent, the domestication origin of cultivated sesame (Bedigian, 2003), were assigned to Subpopulation 2. *Sesamum mulayanum*, the proposed progenitor of *S. indicum* (Bedigian, 2003), was also assigned to the same cluster. Seventeen of the 19 African accessions were almost equally shared between Subpopulation 2 (nine accessions) and the group of admixed accessions (eight accessions).

A dendrogram displaying the molecular genetic relationships of the accessions was drawn using the Dice coefficient and the unweighted neighbor-joining algorithm (Fig. 1). A strong correlation between the distance matrix and the neighbor-joining dendrogram was evident by the Mantel test result ($r = 0.961$). Average pairwise dissimilarity among accessions was 0.39, with the highest value (0.78) calculated between accessions from Turkey (PI205229) and Mozambique (PI231033), and the lowest (0.07) calculated between a Turkish accession (PI 205229) and a Turkish registered cultivar (Cumhuriyet 99). *Sesamum indicum* accessions fell into three clusters (Clusters A, B, and C) in the dendrogram (Fig. 1). Clustering patterns of the dendrogram and population structure analysis were largely in agreement (Table 1).

Cluster A comprised 56 accessions, which coincided with Subpopulation 1, with the exception of a single accession (PI157155, India) that grouped with Cluster B according to the neighbor-joining analysis. Thus, with only one exception, molecular genetic diversity analysis perfectly reflected the subpopulation assignment pattern for Cluster A. The pairwise dissimilarity of accessions in Cluster A ranged between 0.07 and 0.55, with an average pairwise dissimilarity of 0.29 (data not shown). Cluster A consisted of two subclusters, Clusters A1 and A2. Cluster A1 contained the majority of the accessions (45 out of 56 accessions) with an average pairwise dissimilarity of 0.27. The minimum and maximum pairwise dissimilarity values for Cluster A1 were 0.07 and 0.50, respectively

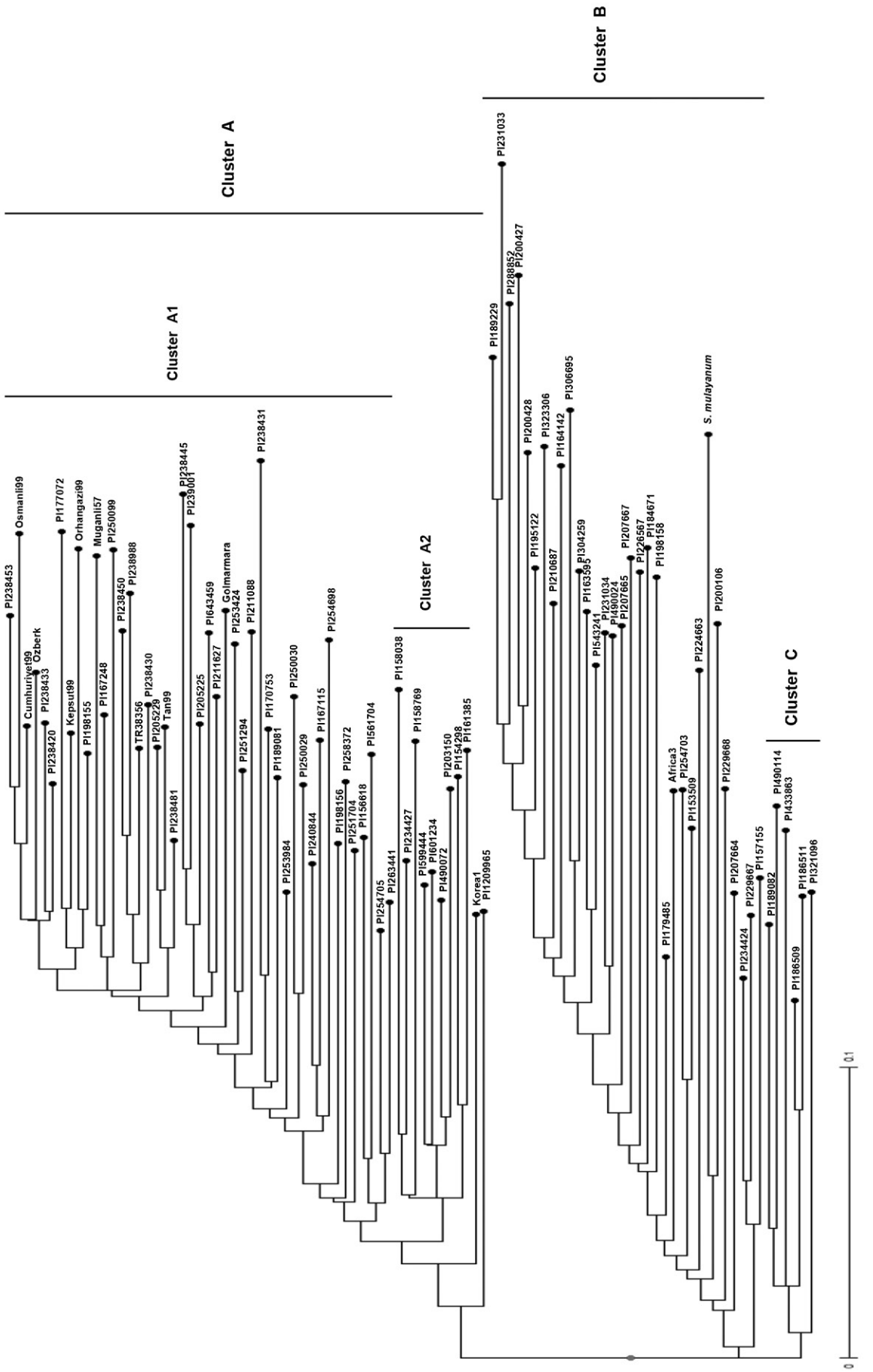

Figure 1. Unweighted neighbor-joining dendrogram of sesame accessions constructed using genomic simple sequence repeat markers.

(data not shown). All of the Turkish accessions were in Cluster A1. The average pairwise dissimilarity of the Turkish accessions was 0.24 (data not shown), indicating the presence of a moderate level of molecular genetic diversity within the germplasm. Registered cultivars were intermixed with the remaining Turkish accessions in the dendrogram. All accessions from the Middle East, except two, fell into Cluster A1. The remaining accessions in Cluster A1 were from diverse locations including five accessions from Central and East Asia, three accessions from the Americas, two accessions from Southern Europe, two accessions from the former Soviet Union, and one accession from Central Africa. Nine accessions, including four accessions from the Americas, four from East Asia, and one from the Middle East were grouped into Cluster A2. Average pairwise dissimilarity for Cluster A2 was 0.26, with minimum and maximum values of 0.16 and 0.40, respectively (data not shown). Two accessions in Cluster A, one from Ethiopia and one from Korea, were the most genetically distinct accessions in this cluster and fell into neither of the two subclusters.

Cluster B comprised 32 accessions. The wild *Sesamum* accession, *S. mulayanum*, fell into this cluster alongside seven accessions from the Indian subcontinent, five accessions from East and Southeast Asia, 11 accessions from Africa, and six accessions from the Americas. Two accessions, representing the former Soviet Union and the Middle East, also fell into Cluster B. With the exception of six admixed accessions and one Indian accession assigned to Subpopulation 1, Cluster B accessions coincided with Subpopulation 2. The pairwise dissimilarity among accessions in the cluster ranged between 0.16 and 0.65, while the highest average pairwise dissimilarity (0.40) was calculated for this cluster (data not shown). Thus, molecular genetic diversity was highest in Cluster B. Separate evaluation of the accessions from the Indian subcontinent gave a relatively high average pairwise dissimilarity value of 0.42 (data not shown), indicating a considerable level of molecular genetic diversity in the domestication center. Cluster C only contained six accessions from Africa. The pairwise dissimilarity among the six accessions ranged between 0.11 and 0.39, with an average pairwise dissimilarity of 0.27 (data not shown). Interestingly, all six accessions in Cluster C were in the admixed group according to population structure analysis.

DISCUSSION

Simple Sequence Repeat Development and Validation

De novo development of genomic SSRs by next-generation sequencing has several advantages over conventional microsatellite enrichment-based approaches. Apart from the higher cost and higher demand on time and labor, SSR development by sequencing microsatellite-enriched library clones results in the identification of a biased set of SSRs defined by the motifs incorporated in the oligonucleotide probes. In contrast, next-generation sequencing approaches yield a vast quantity of sequence data from which an unbiased search for all types of SSR motifs can be performed with much less labor and time devoted for the experimental process (Castoe et al., 2010). In this study, a genomic marker development approach using pyrosequencing data proved successful in identifying a large number of SSRs (19,816 SSRs) in the sesame genome. Our contig assembly encompassed a good portion of the sesame genome (19.3%), allowing the identification of 5727 nonredundant SSRs, corresponding to an average density of one SSR every 11.3 kb of genomic DNA. A similar finding was reported by Wei et al. (2014), who estimated the average distance between SSRs in sesame genome as 11.7 kb. However, our SSR density estimate is lower than those reported for other monocot and dicot plant genomes, which ranged between one SSR per 1 and 6 kb (Cardle et al., 2000; Lawson and Zhang, 2006; Cavagnaro et al., 2010; Sonah et al., 2011). The amount of analyzed sequence, SSR search parameters, and data mining algorithm all directly impact the resultant number and frequency of identified SSRs. As a result, there is often discrepancy among SSR density estimates reported for the same species by different authors. For example, while analyzing the same sequence data, the density of SSRs identified in sesame genic sequences decreased from one SSR per 6.6 to 10.8 kb, when the minimum SSR length was increased from 15 to 18 bases (Zhang et al., 2012). The average SSR density of the *Arabidopsis thaliana* (L.) Heynh. genome was reported as one SSR per 6, 1.1, and 2.4 kb, by Cardle et al. (2000), Lawson and Zhang (2006), and Sonah et al. (2011), respectively. Similarly, there is a dramatic difference between the two SSR density estimates for the sorghum (*Sorghum bicolor* L. Moench.) genome reported by Cavagnaro et al. (2010) (one SSR per 3.1 kb) and Sonah et al. (2011) (one SSR per 5.7 kb). All of these results highlight the fact that none of these estimates can be taken as the ultimate reference unless SSR mining criteria and algorithms are standardized across different studies.

While the common definition of SSRs specifies motif length as 1 to 6 nucleotides (Jones et al., 2009), we also included hepta- and octanucleotide repeats in our SSR survey and identified both SSR types at a higher frequency than tetra- and pentanucleotide SSRs (Table 3). Thus, it was valuable to expand our search with hepta- and octanucleotide repeats, as we found that the presence of these repeat types in the sesame genome is not negligible. Mononucleotide repeats were the most abundant repeat type, followed by dinucleotide SSRs (48.5 and 45.03% of all SSRs, respectively). Our results were in agreement with that of Cardle et al. (2000) and Sonah et al. (2011), who identified mononucleotide repeats as the predominant repeat type in several plant genomes including *A. thaliana*, purple false brome [*Brachypodium distachyon* (L.) P. Beauv], sorghum, rice (*Oryza sativa* L.), barrel clover (*Medicago truncatula* Gaertn.) and poplar (*Populus trichocarpa* Torr. & A. Gray). In concordance with our findings, A/T was the most abundant mononucleotide repeat in all

of the plant genomes examined by Sonah et al. (2011), and the sum of AT and TA repeats constituted more than 50% of the dinucleotide repeats in the genomes of dicot species. The trend also applied for trinucleotide SSRs with a predominance of AT-rich repeats, similar to our findings. Whether or not mononucleotide repeats are included in SSR surveys, all reports on genic SSR development in sesame indicate dinucleotide repeats as the predominant SSR type in coding sequences (Wei et al., 2008; Wei et al., 2011; Zhang et al., 2012; Yepuri et al., 2013; Wu et al., 2014). In addition, AG/CT was consistently found as the predominant motif in sesame genic SSRs.

Evolutionary forces apply differently to coding and noncoding sequences, therefore, identification of SSRs from genomic and genic datasets is likely to result in distinct patterns of repeat type and motif abundance. For example, the presence of mononucleotide repeats significantly increases the rate of insertion–deletion mutations and such mutations easily escape proofreading and mismatch repair mechanisms, leading to transcriptional and translational slippage and frameshift mutations. Thus, such repeats are selected against in coding sequences (Gu et al., 2010). In contrast, monucleotide repeats are overrepresented in genomic SSR sets because they do not introduce a constraint on the function of most genomic sequences. The difference in motif abundance between genomic and genic SSRs might be explained by the fact that nucleotide abundance is biased toward a higher GC content in coding sequences (Messeguer et al., 1991).

Primers were designed to amplify the SSRs identified in contigs. Dye-terminator sequencing of PCR products of randomly selected primers proved successful in identifying the expected perfect SSR motifs within amplicons, validating the reliability of our SSR marker design approach (data not shown). When our SSR primers were tested on a *S. mulayanum* accession, a very high rate of marker transferability (91%) was detected. This result was anticipated because experimental evidence suggests that *S. mulayanum* is the wild progenitor of cultivated sesame (Kawase, 2000; Bedigian, 2003). Indeed, *S. indicum* and *S. mulayanum* are proposed as the domesticated and wild forms of the same biological species, since they share the same chromosome number ($2n = 26$) and their reciprocal crosses produce fertile progeny. In addition, while members of the *Sesamum* genus may have differences in their seed lignan compositions, seeds of both *S. indicum* and *S. mulayanum* accumulate the two major lignans, sesamin and sesamolin (Bedigian, 2003). In this study, *S. mulayanum* was not detected as an outgroup in the neighbor-joining analysis and was clustered together with *S. indicum* accessions. When the high rate of marker transferability between *S. indicum* and *S. mulayanum* is also taken into account, our results provide additional support for designating *S. indicum* and *S. mulayanum* as the cultivated and wild forms of the same species. This close relationship and the potential of *S. mulayanum* germplasm to harbor disease resistance and abiotic stress tolerance traits (Kawase, 2000) makes it essential to incorporate *S. mulayanum* accessions into breeding programs, if substantial improvement of disease and stress tolerance related characters is intended. Bisht et al. (2004) demonstrated that resistance to phyllody disease and insect pests, drought tolerance, and improved yields could be achieved through selections from crosses between *S. mulayanum* and cultivated accessions. Here, we introduce more than 800 markers that efficiently amplify SSR fragments from both *S. indicum* and *S. mulayanum*. These markers constitute the necessary tools for mapping agriculturally important traits using populations derived from hybrids of the two subspecies and for introgression of those traits into cultivated germplasm via marker-assisted breeding.

Assessment of the Genetic Diversity and Population Structure of a Sesame World Collection

Incongruence between geographical proximity and genetic distance has been reported for sesame germplasm by several authors (Bhat et al., 1999; Kim et al., 2003; Laurentin and Karlovsky, 2006; Zhang et al., 2012). In many locations, the genetic basis of sesame is narrow and based on the allelic pool derived from limited introductions (Bhat et al., 1999). Kim et al. (2003) suggested that genetic resemblance of sesame accessions from diverse geographical locations could be the consequence of limited introduction and exchange of material between diverse locations. In our study, neither molecular genetic diversity nor population structure analysis yielded a topology strictly defined by geographical location. However, our results displayed certain patterns of association between genetic similarity and geographical proximity.

When the results of molecular genetic diversity and population structure analyses were compared, the clustering patterns of the neighbor-joining dendrogram and subpopulation assignment analyses overlapped almost perfectly. Structure analysis suggested a model that assigns sesame accessions into two subpopulations. With only a few exceptions, the two subpopulations, Subpopulations 1 and 2, corresponded to Cluster A and B accessions in the neighbor-joining dendrogram, respectively. Accessions that were assigned to neither of the subpopulations in population structure analysis constituted the admixed group, which corresponded to Cluster C accessions of the dendrogram with the addition of six Cluster B accessions.

The Indian subcontinent is proposed as the domestication origin of cultivated sesame (Bedigian, 2003). In our analysis, Indian subcontinent accessions displayed a high average pairwise dissimilarity (0.42) as expected. With the exception of a single Indian accession, all accessions from the Indian subcontinent, including the putative progenitor *S. mulayanum*, grouped together in Subpopulation 2. None of the Turkish accessions fell into Subpopulation 2, suggesting that Turkish germplasm is genetically quite distinct from the germplasm in sesame's origin of domestication. In agreement with our results, Ashri (1998) indicated that visual inspection is sufficient to distinguish Turkish and Indian sesame accessions. A

similar case applied for Middle Eastern accessions, with a majority of accessions clustered together with Turkish accessions in the same subpopulation and the same cluster (Cluster A1) of the neighbor-joining dendrogram. Thus, to a certain extent, genetic diversity seemed to correlate with geographical proximity in the Middle East region, including Turkey. Turkish registered cultivars were intermixed with landraces in the neighbor-joining dendrogram. This was not an unexpected result since genic sequences, which are presumably under the pressure of artificial selection, are not represented at a high rate in genomic marker sets. Thus, a high proportion of markers developed from genomic sequences would be phenotypically neutral. Therefore, genomic SSR markers would not necessarily reflect artificial selection events and would not distinguish cultivars from landraces.

In contrast to Indian subcontinent accessions that were almost exclusively found in Subpopulation 2, more than half (seven accessions) of the East and Southeast Asian (China–Korea–Japan region) accessions were clustered in Subpopulation 1 with the rest (five accessions) distributed to Subpopulation 2 and the group of admixed accessions. Average pairwise dissimilarity among East and Southeast Asian accessions (0.36) indicated a relatively high level of molecular genetic diversity. These results suggest that sesame germplasm in East and Southeast Asia diversified from that in the domestication origin and that this material harbors a high level of genetic diversity.

African accessions were mainly shared between Subpopulation 2 and the group of admixed accessions with a high average pairwise dissimilarity of 0.40 calculated for these accessions. Out of 19 African accessions included in the analysis, nearly half (eight accessions) fell into the group of admixed accessions, six of which constituted a separate cluster (Cluster C) in the neighbor-joining dendrogram, indicating that these genotypes were highly distinct from the rest of the analyzed accessions. These results suggest intense interbreeding activity between the two sesame subpopulations in Africa, resulting in the emergence of a germplasm distinct from that in Asia and the Indian subcontinent. Of the 13 accessions from the Americas, 11 were distributed to the two subpopulations with seven accessions in Subpopulation 1 and four accessions in Subpopulation 2, implying that the genetic basis of sesame in the Americas is constituted by introductions from both subpopulations. In support of our conclusion, a high average pairwise dissimilarity of 0.34 was calculated for these accessions.

Overall, in agreement with the results of Laurentin and Karlovsky (2006), our results suggest that the well-recognized diversity centers of sesame, Africa and the China–Korea–Japan region, harbor almost as much genetic diversity as the domestication origin. As stated above, clustering of accessions did not follow a strict correlation with geographical location but provided hints for better exploiting the breeding potential of sesame by evaluating the genetic resemblance of the analyzed accessions. Results of the subpopulation assignment analysis display the pattern of gene flow among sesame diversity centers and should be useful for selection of parents with diverse genetic backgrounds while designing breeding schemes. Thus, our SSR markers proved successful in identifying molecular diversity and resolving genetic relationships in a set of accessions from throughout the world. These markers will be of great use for genome mapping, core collection establishment, germplasm enhancement, and marker-assisted breeding studies.

Acknowledgments

This study was supported by grant 108O478 from the Scientific and Technological Research Council of Turkey (TUBITAK). We are grateful to Dr. Makoto Kawase, National Institute of Agrobiological Sciences, Genetic Resources Center, for providing *Sesamum mulayanum* seed samples. We also thank Dr. Ahmet Semsettin Tan from AARI for seeds of Turkish sesame cultivars and Ali Tevfik Uncu from Izmir Institute of Technology for help with plant growth and DNA extraction.

References

Abou-Gharbia, H.A., A.A.Y. Shehata, and F. Shahidi. 2000. Effect of processing on oxidative stability and lipid classes of sesame oil. Food Res. Int. 33:331–340. doi:10.1016/S0963-9969(00)00052-1

Anilakumar, K.R., A. Pal, F. Khanum, and A.S. Bawa. 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: J. Janick, editor, Plant breeding reviews. Vol. 16. John Wiley & Sons, Oxford. p. 179–228.

Bedigian, D. 2003. Evolution of sesame revisited: Domestication, diversity and prospects. Genet. Resour. Crop Evol. 50:779–787. doi:10.1023/A:1025029903549

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

Bhat, K.V., P.P. Babrekar, and S. Lakhanpaul. 1999. Study of genetic diversity in Indian and exotic sesame (*Sesamum indicum* L.) germplasm using random amplified polymorphic DNA (RAPD) markers. Euphytica 110:21–33. doi:10.1023/A:1003724732323

Bisht, I.S., K.V. Bhat, S. Lakhanpaul, B.K. Biswas, M. Pandiyan, and R.R. Hanchinal. 2004. Broadening the genetic base of sesame (*Sesamum indicum* L.) through germplasm enhancement. Plant Genet. Resour. 2:143–151. doi:10.1079/PGR200445

Brown, S.M., M.S. Hopkins, S.E. Mitchell, M.L. Senior, T.Y. Wang, R.R. Duncan, S. Gonzales Candelas, and S. Kresovich. 1996. Multiple methods for the identification of polymorphic simple sequence repeats (SSRs) in sorghum [*Sorghum bicolor* (L.) Moench]. Theor. Appl. Genet. 93:190–198. doi:10.1007/BF00225745

Cardle, L., L. Ramsay, D. Milbourne, M. Macaulay, D. Marshall, and R. Waugh. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156:847–854.

Castoe, T.A., A.W. Poole, W. Gu, A.P.J. de Koning, J.M. Daza, E.N. Smith, and D.D. Pollock. 2010. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Mol. Ecol. Resour. 10:341–347. doi:10.1111/j.1755-0998.2009.02750.x

Cavagnaro, P.F., D.A. Senalik, L. Yang, P.W. Simon, T.T. Harkins, C.D. Kodira, S. Huang, and Y. Weng. 2010. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genomics 11:569. doi:10.1186/1471-2164-11-569

Chevreux, B., T. Pfisterer, B. Drescher, A.J. Driesel, W.E.G. Müller, T. Wetter, and S. Suhai. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14:1147–1159. doi:10.1101/gr.1917404

Dixit, A., M.H. Jin, J.W. Chung, J.W. Yu, H.K. Chung, K.H. Ma, Y.J. Park, and E.G. Cho. 2005. Development of polymorphic microsatellite markers in sesame (*Sesamum indicum* L.). Mol. Ecol. Notes 5:736–738. doi:10.1111/j.1471-8286.2005.01048.x

Earl, D.A., and B.M. Von Holt. 2012. Structure Harvester: A website and program for visualizing Structure output and implementing the Evanno method. Conserv. Genet. Resour. 4:359–361. doi:10.1007/s12686-011-9548-7

FAOSTAT. 2014. Food and Agriculture Organization of the United Nations. http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor (accessed 31 Aug. 2014).

Frary, A., P. Tekin, I. Celik, S. Furat, B. Uzun, and S. Doganlar. 2014. Morphological and molecular diversity in Turkish sesame germplasm and selection of a core set for inclusion in the national collection. Crop Sci. 54:1–10. doi:10.2135/cropsci2012.12.0710

Gu, T., S. Tan, X. Gou, H. Araki, and D. Tian. 2010. Avoidance of long mononucleotide repeats in codon pair usage. Genetics 186:1077–1084. doi:10.1534/genetics.110.121137

Jannati, M., R. Fotouhi, A.P. Abad, and Z. Salehi. 2009. Genetic diversity analysis of Iranian citrus varieties using micro satellite (SSR) based markers. J. Hortic. For. 1:120–125.

Jones, N., H. Ougham, H. Thomas, and I. Pasakinskiene. 2009. Markers and mapping revisited: Finding your gene. New Phytol. 183:935–966. doi:10.1111/j.1469-8137.2009.02933.x

Kawase, M. 2000. Genetic relationships of the ruderal weed type and the associated weed type of *Sesamum mulayanum* NAIR distributed in the Indian subcontinent to cultivated sesame, *S. indicum* L. Jpn. J. Trop. Agr. 44:115–122.

Kim, D.H., G. Zur, Y. Danin-Poleg, S.W. Lee, K.B. Shim, C.W. Kang, and Y. Kashi. 2003. Genetic relationships of sesame germplasm collection as revealed by inter-simple sequence repeats. Plant Breed. 121:259–262. doi:10.1046/j.1439-0523.2002.00700.x

Laurentin, H.E., and P. Karlovsky. 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

Lawson, M.J., and L. Zhang. 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol. 7:R14. doi:10.1186/gb-2006-7-2-r14

Lipman, D.J., and W.R. Pearson. 1985. Rapid and sensitive protein similarity searches. Science 227:1435–1441. doi:10.1126/science.2983426

Messeguer, R., M.W. Ganal, J.C. Steffens, and S.D. Tanksley. 1991. Characterization of the level, target sites and inheritance of cytosine methylation in tomato nuclear DNA. Plant Mol. Biol. 16:753–770. doi:10.1007/BF00015069

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70:3321–3323. doi:10.1073/pnas.70.12.3321

Powell, W., G.C. Machray, and J. Provan. 1996. Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1:215–222. doi:10.1016/1360-1385(96)86898-1

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

Roldan-Ruiz, I., J. Dendauw, E.V. Bockstaele, A. Depicker, and M.D. Loose. 2000. AFLP markers reveal high polymorphic rates in ryegrasses (*Lolium* spp.). Mol. Breed. 6:125–134. doi:10.1023/A:1009680614564

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. In: S. Krawetz and S. Misener, editors, Methods in molecular biology: Bioinformatics methods and protocols. Vol. 132. Humana Press, Totowa, NJ. p. 365–386.

Sonah, H., R.K. Deshmukh, A. Sharma, V.P. Singh, D.K. Gupta, R.N. Gacche, J.C. Rana, N.K. Singh, and T.R. Sharma. 2011. Genome-wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium*. PLoS ONE 6:e21298. doi:10.1371/journal.pone.0021298

Spandana, B., V.P. Reddy, G.J. Prasanna, G. Anuradha, and S. Sivaramakrishnan. 2012. Development and characterization of microsatellite markers (SSR) in *Sesamum* (*Sesamum indicum* L.) species. Appl. Biochem. Biotechnol. 168:1594–1607. doi:10.1007/s12010-012-9881-7

Surapaneni, M., V. Yepuri, L.R. Vemireddy, A. Ghanta, and E.A. Siddiq. 2014. Development and characterization of microsatellite markers in Indian sesame (*Sesamum indicum* L.). Mol. Breed. 34:1185–1200. doi:10.1007/s11032-014-0109-0

Wang, L., S. Yu, C. Tong, Y. Zhao, Y. Liu, C. Song, Y. Zhang, X. Zhang, Y. Wang, W. Hua, D. Li, D. Li, F. Li, J. Yu, C. Xu, X. Han, S. Huang, S. Tai, J. Wang, X. Xu, Y. Li, S. Liu, R.K. Varshney, J. Wang, and X. Zhang. 2014. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. Genome Biol. 15:R39. doi:10.1186/gb-2014-15-2-r39

Wei, L.B., H.Y. Zhang, Y.Z. Zheng, W.Z. Guo, and T.Z. Zhang. 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, W., X. Qi, L. Wang, Y. Zhang, W. Hua, D. Li, H. Lv, and X. Zhang. 2011. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics 12:451. doi:10.1186/1471-2164-12-451

Wei, X., L. Wang, Y. Zhang, X. Qi, X. Wang, X. Ding, J. Zhang, and X. Zhang. 2014. Development of simple sequence repeat (SSR) markers of sesame (*Sesamum indicum*) from a genome survey. Molecules 19:5150–5162. doi:10.3390/molecules19045150

Wu, K., M. Yang, H. Liu, Y. Tao, J. Mei, and Y. Zhao. 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

Yepuri, V., M. Surapaneni, V.S.R. Kola, L.R. Vemireddy, B. Jyothi, V. Dineshkumar, G. Anuradha, and E.A. Siddiq. 2013. Assessment of genetic diversity in sesame (*Sesamum indicum* L.) genotypes, using EST-derived SSR markers. J. Crop Sci. Biotechnol. 16:93–103. doi:10.1007/s12892-012-0116-9

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