DEVELOPMENT OF SEQUENCE BASED MARKERS FOR MOLECULAR GENETIC ANALYSIS IN SESAME (*Sesamum indicum* L.)

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Molecular Biology and Genetics

by
Ayşe Özgür UNCU

June 2015
İZMİR

We approve the thesis of **Ayşe Özgür UNCU**

Examining Committee Members:

_____

Prof. Dr. Anne FRARY
Department of Molecular Biology and Genetics, Izmir Institute of Technology

_____

Prof. Dr. Sami DOĞANLAR
Department of Molecular Biology and Genetics, Izmir Institute of Technology

_____

Prof. Dr. Hülya İLBİ
Department of Horticulture, Ege University

_____

Assoc. Prof. Dr. Ali Ramazan ALAN
Department of Biology, Pamukkale University

_____

Assoc. Prof. Dr. Ekrem DÜNDAR
Department of Molecular Biology and Genetics, Balıkesir University

15 June 2015

_____

Prof. Dr. Anne FRARY
Supervisor, Department of Molecular Biology and Genetics,
Izmir Institute of Technology

_____

Prof. Dr. Ahmet KOÇ
Head of the Department of
Molecular Biology and Genetics

_____

Prof. Dr. Bilge KARAÇALI
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisors Professor Dr. Anne Frary and Prof. Dr. Sami Doğanlar. They generously shared their wisdom, knowledge and experience with me throughout my studies, and guided and encouraged me to be productive and develop as a scientist. It has been an honour to be a PhD student of Dr. Frary and Dr. Doğanlar and I will feel this honour throughout my lifetime.

I am grateful to İbrahim Çelik for being a supportive and generous colleague and a brother. I am also grateful to my colleagues Saniye Elvan Öztürk and Süleyman Can Öztürk for their sincere support and friendship.

I would like to thank my dear husband and colleague Ali Tevfik Uncu for supporting me at all times, and being my inspiration with his creative mind and gentle heart.

ABSTRACT

DEVELOPMENT OF SEQUENCE BASED MARKERS FOR MOLECULAR GENETIC ANALYSIS IN SESAME (*Sesamum indicum* L.)

Sesame (*Sesamum indicum* L.) is an orphan crop with most molecular genetic research work done in the last decade. In this study, pyrosequencing was used for the development of genomic SSR (Simple Sequence Repeat) markers in sesame. The approach proved successful in identifying 19,816 SSRs, 5727 of which were identified in a contig assembly that covers 19.29% of the sesame genome. As a result of this work, 933 experimentally validated sesame specific markers were introduced, 849 of which are applicable in *Sesamum mulayanum*, the wild progenitor of cultivated sesame. Using a subset of SSR markers, molecular genetic diversity and population structure of a collection of world accessions were analyzed. Results of the analyses revealed a pattern of gene flow among sesame diversity centers. Taken together with the high rate of genomic marker transferability between *S. indicum* and *S. mulayanum*, the results provide molecular genetic evidence for designating the two taxa as cultivated and wild forms of the same species.In related work, a Genotyping By Sequencing (GBS) approach was applied on recombinant inbred lines for single nucleotide polymorphism (SNP) identification and mapping in the sesame genome. As a result, 15,521 SNPs were identified and a high-resolution genetic linkage map was constructed using a core set of selected SNPs (781 SNPs) appropriate for use in linkage analysis. The 15,521 putative SNP markers represent a substantial contribution to the existing pool of sesame-specific markers. The genetic linkage map constructed in this work will enable the identification of loci involved in the genetic control of agriculturally important traits in sesame.

ÖZET

SUSAM'DA (*Sesamum indicum* L.) MOLEKÜLER GENETİK ANALİZLER İÇİN DİZİ TEMELLİ MARKÖRLERİN GELİŞTİRİLMESİ

Susam (*Sesamum indicum* L.) ihmal edilmiş bir tarım ürünü olup, nispeten az sayıda moleküler genetik çalışmaya konu olmuştur. Bu çalışmada, susam genomuna özel SSR markörleri geliştirmek üzere bir yeni nesil dizileme analiz yöntemi kullanılmış, 5727 adeti montajlanmış dizilere denk düşen toplam 19,816 SSR tanımlanmıştır. Çalışma sonucunda susam genomundan SSR bandı çoğalttığı deneysel olarak gösterilmiş toplam 933 SSR markörü tanıtılmıştır. Bu markörlerin 849 adedinin, kültüre alınmış susamın atası olan *Sesamum mulayanum*'a uygulanabilir olduğu gösterilmiştir. Yeni geliştirilen bir kısım SSR markörü, bir susam koleksiyonunda genetik çeşitlilik ve popülasyon yapısının araştırılmasında kullanılmıştır. Çalışma sonuçları, dünya genelindeki susam çeşitlilik merkezleri arasındaki mevcut ilişkileri ortaya koymuştur. SSR markörlerinin *S. indicum* ve *S. mulayanum* arasında yüksek oranda transfer edilebilir oluşu ile birlikte, genetik çeşitlilik ve popülasyon yapısı analiz sonuçları, *S. indicum* ve *S. mulayanum*'un aynı türün kültüre alınmış ve yabani formları olduğuna delil teşkil etmektedir. Çalışma kapsamında ayrıca, susam genomuna özel SNP markörleri geliştirmek ve bir genetik bağlantı haritası oluşturmak üzere GBS analiz yöntemi bir rekombinant saf hat popülasyonuna uygulanmıştır. Bu sayede susam genomunda 15,521 adet SNP tanımlanmış ve bu SNP'lerin 781 adedi, GBS analizinin uygulandığı rekombinant saf hat popülasyonunda gerçekleştirilecek bağlantı analizlerinde kullanılmak üzere tanımlanmıştır. Çalışma kapsamında yüksek sayıda SNP ve SSR markörünün geliştirilmesi ile, susam genomuna özel kısıtlı sayıdaki mevcut markörlere hatırı sayılır bir katkı yapılmıştır. GBS analizi sonucu oluşturulan genetik bağlantı haritası, zirai öneme sahip özelliklerin genetik temellerinin rekombinant saf hat popülasyonu kullanılarak araştırılmasına olanak sağlayacaktır.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Sesame (*Sesamum indicum* L.)

Sesame (*S. indicum* L. syn. *S.orientale* L.) (2n = 26), is a member of the Pedaliaceae family and is considered as the oil seed crop with the most ancient cultivation history (Bedigian, 2003). Based on archaeological evidence found in the Indus Valley, sesame cultivation can be traced back to 3050-3500 B.C. (Bedigian and Harlan, 1986). 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, 2013).

Sesame is an annual plant that displays erect growth with a plant height that ranges between 40 and 200 cm. The stem can be unbranched or, intermediately or heavily branched, depending on the genotype. The plant has tubular flowers and the corolla can be white, yellow or purple in color. Sesame is primarily an autogamous species, however, high outcrossing rates were identified in areas where insect activity is high or other vegetation that attracts pollinators is limited (Morris, 2009).

S. indicum produces capsules as fruits that range in size between 1.5 and 4 cm. While capsules are mostly bicarpellate, there are also genotypes that bear tetracarpellate capsules. Seed color is highly variable with white, black and several intermediate colors present (Bedigian, 1986; Bedigian and Harlan, 1986). As a crop, sesame is not nutrient demanding, therefore, it can be cultivated in rotation with soil-exhaustive crops. Sesame tolerates high temperatures and drought, but is susceptible to frost and waterlogging, thus, sesame is definitely a warm-season crop (Bedigian and Harlan, 1986).

Among the oil seed crops, sesame is the least productive and is mainly grown by small holders (Bhat et al., 1999; Bisht et al., 2004). Two growth habit traits, seed shattering and indeterminate growth constitute the major constraints on seed yield. As a result of indeterminate growth, flowering continues as long as conditions are favorable,

leading to the presence of capsules at different maturity stages on the same plant (Ashri, 1998). Because capsules dehisce upon maturity, seeds in mature capsules can be scattered and lost while there are still flowers and unripe capsules along the stem (Day, 2000). A natural mutant with indehiscent capsules was isolated in 1943, however, it could not be efficiently utilized in breeding programs due to adverse pleiotropic effects observed in the derived hybrids (Çağırgan, 2001). Similarly, negative side effects such as semi-sterility and low yield accompanied the closed-capsule trait in gamma-ray mutants (Çağırgan, 2001; Uzun et al., 2003). Irradiation mutants with determinate growth habit have also been isolated (Çağırgan, 2001; Uzun ve Çağırgan, 2009). Determinate growth was observed to result in synchronous flowering and lodging resistance, but negatively affected seed yield (Uzun ve Çağırgan, 2009).

In addition to the growth habit traits of capsule dehiscence and indeterminate growth, susceptibility to diseases caused by the bacterium *Pseudomonas syringae* pv. and the fungi *Cerocospora sesame* and *Alternaria sesame* reduce seed yield significantly. Incorporation of wild relatives into sesame breeding programs as a source of biotic stress resistance alleles is recommended in order to establish disease resistance in cultivated sesame (Kawase, 2000).

1.2. Nutritional value and uses

Sesame seeds are very nutritious with almost 50% oil and up to 25% protein content (Anilakumar et al., 2010). Sesame is cultivated primarily for its oil. An edible oil of exceptional quality with a very high degree of resistance against oxidative deterioration (Namiki, 2007) and a high unsaturated fatty acid content is obtained from sesame. Moreover, polyunsaturated fatty acids constitute more than half of the unsaturated fatty acid fraction in seeds (Anilakumar et al., 2010). Therefore, sesame deserves its reputation as ‘Queen of the oil seeds’ (Bedigian and Harlan, 1986). 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, anti-aging, anti-hypertensive, anti-cancer, cholesterol lowering and anti-mutagenic properties are reported by several authors (reviewed by Anilakumar, et al., 2010). The two major lignans, sesamin and sesamolin are proantioxidants. Once ingested, they turn into potent antioxidants as a result of

intramolecular rearrangements. Sesaminol is a newly identified sesame lignan and displays a strong antioxidant activity in its free form. Upon ingestion, free sesaminol is released as a result of deglucosylation by the action of β-glucosidases and intestinal microorganisms (Namiki 2007).

Sesame is indeed a versatile crop which serves as a leafy vegetable in addition to its common use as an oil seed. Leaves have a rich carotenoid, ascorbic acid, iron and calcium content and contain good amounts of protein (Bedigian, 2003). Therefore, sesame is a potential food-security crop in rural regions of Africa and Asia, where its cultivation is most common. Sesame seeds and oil are not only utilized for human consumption but also have uses in pharmaceutical and cosmetic industries. Recently, the potential of sesame oil for biodiesel production was evaluated (Ahmad et al., 2010). However, given that sesame is the least productive among the oil seed crops, feasibility of large-scale biodiesel production from sesame oil is questionable.

1.3. Domestication

Several reports exist that propose Africa as the domestication origin of cultivated sesame (*S. indicum*) (Ihlenfeldt and Grabow- Seidensticker, 1979; Kobayashi, 1986; Morris, 2009). This is probably due to the fact that most of the wild species that belong to the *Sesamum* genus are exclusive to Africa (Bedigian, 2003). Ihlenfeldt and Grabow- Seidensticker (1979) proposed the wild species *S. latifolium* Gillett from Africa as the progenitor of cultivated sesame. However, reciprocal crosses failed to yield fertile progeny from *S. indicum* (2n = 26) and *S. latifolium* (2n = 32), as a result of the mismatch between the chromosome numbers of the two species (Bedigian 2003; Bedigian, 2010). Because fertility of F_1 hybrids is the basic criterion to define the wild progenitor of a cultivated species (Harlan, 1992), *S. latifolium* cannot be considered as the putative progenitor of cultivated sesame. Indeed, there is sufficient evidence to assign not Africa but the Indian subcontinent as the domestication origin. The taxon *S. mulayanum* (syn. *S. orientale* var. *malabaricum* Nar.) is native to India. *S. mulayanum* displays morphological and physiological resemblance with a section of Indian sesame accessions. There is also biochemical evidence to assign *S. mulayanum* as the putative progenitor of cultivated sesame. 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. Most importantly, *S. indicum and S. mulayanum* share the same chromosome number (2n = 26) and their reciprocal crosses yield fully fertile progeny (Bedigian 2003), meeting the basic diagnostic criterion to assign the wild ancestor of a domesticated species.

Interestingly, sesame has retained certain characters such as seed shattering, indeterminate growth and, late and long maturity, which usually are associated with wild plants and are selected against by human activity. Therefore, sesame is not a stereotypical domesticated plant. Except for the capsule dehiscence trait which was probably retained as a result of the harvest strategies that were adapted by growers to the plant's seed dispersal habit, seed characters associated with yield and oil content have been domesticated. For example, smooth seeds with higher oil content have been selected against rough seeds. Larger capsules have been selected and tetracarpellate mutants have been kept for improved seed yields. Diversity in seed color and taste has been maintained due to a preference for different seed characteristics for different uses (Bedigian, 2003).

1.4. Molecular genetic research in sesame

1.4.1. Germplasm characterization in sesame using random DNA marker systems

While integration of molecular marker technologies has 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 (RAPD) technique for sesame germplasm characterization (Bhat et al., 1999). Results of the diversity analysis based on amplification profiles obtained with RAPD primers identified a higher molecular genetic diversity within a collection from the Indian subcontinent compared to accessions from outside India. This was an expected result, given that physiological, biochemical and genetic evidence indicates India as the domestication origin of cultivated sesame.

Because sequence knowledge is not a prerequisite for the utilization of random DNA markers, early molecular genetic work in sesame involved the use of such marker systems due to a lack of available sesame genome/transcriptome sequences. Kim et al. (2003) used Inter Simple Sequence Repeat (ISSR) markers in order to assess the genetic diversity in a collection representing different geographical origins. The clustering of sesame accessions was not correlated with geographical origin and the authors suggested that genetic resemblance of sesame accessions from different geographical locations could be the consequence of limited introduction and exchange of material between diverse locations. Laurentin and Karlovsky (2006) investigated the association between geographical origin and genetic diversity based on Amplified Fragment Length Polymorphisms (AFLPs). Similar with the results obtained with ISSRs (Kim et al., 2003), genetic similarity and geographical proximity were not found to be correlated. Taken together, the results of diversity assessment with ISSRs and AFLPs implied that sesame breeding schemes should not solely depend on selecting parents from diverse geographical locations unless genetic dissimilarity is confirmed prior to hybridizations. Correlation between metabolic and genetic diversity in sesame was evaluated by comparing the metabolic diversity assessed with HPLC analysis and molecular genetic diversity assessed with AFLPs (Laurentin et al., 2007). Overall, metabolic diversity among sesame accessions was higher than the molecular genetic diversity. Seed metabolic profiles and geographical origins did not display a correlation and patterns of metabolic and molecular genetic diversity did not overlap. Hence, breeding for metabolic traits such as lignan composition and oil content cannot solely rely on molecular genetic diversity assessments but should involve metabolic measurements. However, identification of DNA markers that are tightly linked to important metabolic characters will be useful for the precise assessment of metabolic diversity based on genotype, without the requirement for metabolic measurements.

1.4.2. Development of sesame-specific Simple Sequence Repeat markers

It was not until 2005 that the first set of sequence-based markers, sesame-specific SSRs were generated (Dixit et al., 2005). Since then, sequence-based marker development efforts in sesame have mostly focused on the development of SSR

markers. 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., 2014a). Simple sequence repeats are widely used in plant molecular genetic studies because they are co-dominant, hypervariable, reproducible, relatively abundant, and provide extensive genome coverage (Powell et al., 1996).

In the first report of SSR marker development in sesame, Dixit et al. (2005) introduced a total of ten SSR primers that amplify mononucleotide/dinucleotide repeats from the sesame genome using a microsatellite enrichment strategy. Spandana et al. (2012) also followed an enrichment-based approach for the development of 111 genomic SSR markers in sesame. SSR marker development through protocols that involve microsatellite-enriched library preparation is a costly and labor intensive approach, yielding in return, a relatively small number of markers. In addition, the resultant markers are biased toward the motifs incorporated in the oligonucleotide probes (Castoe et al., 2010). Therefore, a more productive marker development strategy would be either mining for SSRs in publicly available sesame sequences or *de novo* production of genomic or EST (Expressed Sequence Tag) sequences via Next Generation Sequencing (NGS) approaches for SSR surveys.

Data-mining or NGS approaches were employed for the development of sesame-specific SSR markers in a limited number of studies. Wei et al. (2008) searched for SSRs in publicly available sesame EST data and identified 155 non-redundant SSRs. A total of 50 SSR primers were tested for their transferability to other oil seed crops, cotton, soybean and sunflower. As a result, 44 out of 50 primers yielded PCR amplicons from sesame and two, three and four markers were transferable to cotton, soybean and sunflower, respectively. The pattern of molecular genetic diversity assessed with the SSR markers was not correlated with the geographical distribution of sesame accessions. Yepuri et al. (2013) also performed a database search in order to identify SSRs in available sesame EST sequences. Their analyses identified 156 SSR bearing loci amenable to primer design, 50 of which were experimentally validated. Following a NGS-based approach, an SSR survey in the sesame transcriptome resulted in the identification of 7702 potential SSR loci (Wei et al., 2011). When 50 SSR primers were tested experimentally, 40 of them resulted in successful PCR amplifications. Zhang et al. (2012) identified 2164 SSR loci by transcriptome sequencing and experimentally

tested 300 of the newly identified SSR primers. The authors were able to validate a total of 276 markers out of the tested 300 primers. Results obtained from the molecular genetic diversity analysis of a sesame collection using a total of 32 newly developed SSR markers were in agreement with formers reports of Kim et al. (2003), Laurentin and Karlovsky (2006) and Wei et al. (2008), displaying an incongruence between geographical location and genetic similarity for the analyzed sesame accessions.

Efforts for sequencing the sesame genome also served for SSR marker development in sesame. In 2014, the first draft genome assembly comprising 16 pseudomolecules that encompass approximately 80 % of the estimated sesame genome size was published (Wang et al., 2014). The draft assembly was surveyed for SSRs (Wei et al., 2014), resulting in the identification of 23,438 potential SSR loci, 218 of which were experimentally validated as genomic SSR markers.

1.4.3. High-throughput Single Nucleotide Polymorphism (SNP) identification and mapping in the sesame genome

SNPs are single base pair positions in the genome at which different sequence alternatives exist in normal individuals of a population (Brookes, 1999). SNPs are distinguished from point mutations by setting a frequency threshold of 1 % for the rare allele observed in the polymorphic nucleotide locus. While in principle up to four alternative alleles can be present in a SNP locus, SNPs are mostly bi-allelic, displaying only two alternative nucleotides. However, the high abundance of SNPs in the genome compensates for their bi-allelic nature (Rafalski, 2002).

Owing to the continuous progress in next-generation sequencing technologies, high-throughput SNP identification and simultaneous genotyping by sample multiplexing became a fast and cost-efficient route for generating thousands of species-specific markers and large quantities of genotypic data. However, the complexity of plant genomes due to a high abundance of repetitive sequences stands as a challenge to sequence alignment and SNP identification (Zou et al., 2014). Luckily, a variety of protocols such as RRS (Reduced Representation Shotgun sequencing) (Altshuler et al., 2000), CRoPS (Complexity Reduction of Polymorphic Sequences) (van Orsouw et al., 2007), RAD (Restriction-site Associated DNA) tag sequencing (Baird et al., 2008), GBS (Genotyping By Sequencing) (Elshire et al., 2011) and SLAF-seq (Specific Length

Amplified Fragment-Sequencing) (Sun et al., 2013), have been established that enable the reduction of genome complexity. All protocols take advantage of restriction enzymes for the avoidance of repeat rich sequences in genomes and for increasing the abundance of low copy regions in sequencing libraries. As a result, sequence alignment problems are minimized, improving the accuracy of polymorphism identification.

The first report of high-throughput SNP identification and genotyping in sesame describes the application of the SLAF-seq approach (Zhang et al., 2013). The authors identified a total of 3673 polymorphic loci, including SNPs and indels (insertions/deletions). Genotypic data of an F_2 population were used to construct a genetic linkage map of the sesame genome. Out of the 3673 loci, 1233 (1079 SNPs and 154 indels) were mapped into 15 linkage groups. More recently, RAD tag sequencing was applied on a sesame RIL (Recombinant Inbred Line) population, yielding a total of 3769 SNP markers (Wu et al., 2014b). Among 3769 SNPs, 1327 were suitable for use in linkage analysis and were used in conjunction with SSR and indel markers. As a result, a total of 1230 markers (1190 SNPs, 22 SSRs, 18 indels) were mapped into 14 linkage groups.

1.4.4. Genotyping By Sequencing (GBS)

RAD (Restriction site Associated DNA) tag sequencing (Baird et al., 2008) represents the first application of integrated SNP identification and genotyping by multiplex reduced-representation next-generation sequencing. The GBS protocol is a refined version of RAD tag sequencing. The approach was first demonstrated by Elshire et al. (2011) on maize (*Zea mays* L.) and barley (*Hordeum vulgare* L.). Compared to RAD sequencing, GBS has fewer sample preparation steps and library preparation does not involve a fragment size selection procedure. Sample multiplexing is highly simplified by simultaneously ligating barcode and common adapters prior to sample pooling. Because GBS introduced simplicity and cost-efficiency into multiplex reduced-representation sequencing protocols, the approach was readily adopted by the plant genetics and breeding community (Poland and Rife, 2012; He et al., 2014). Utilization of a preexisting marker panel frequently fails to identify the molecular genetic diversity among genotypes of interest. In such cases, genome mapping becomes especially problematic, since the diversity between parental genotypes of a segregating population

should be sufficiently represented by the predefined markers. Because the GBS protocol is intended for *de novo* polymorphism identification and genotyping, it is significantly more efficient in genetic map construction compared to conventional approaches. Single gene mapping through GBS is extremely straightforward, since genotype data are directly co-analyzed with phenotypic scores in order to identify linked markers (Poland et al., 2012). GBS is the most cost and time-efficient approach for generating large quantities of genomic data for mapping quantitative traits in experimental populations or well-defined association panels.

Recently, the usefulness of the GBS approach for high-throughput SNP marker development was shown on allotetraploid Upland cotton (*Gossypium hirsutum* L.) (Islam et al., 2015). Out of the 5617 SNPs identified through GBS, 2294 could be assigned to Upland cotton subgenomes. In another recent study, a Genome Wide Association Study (GWAS) based on 7530 SNP loci identified by GBS proved successful in identifying SNPs highly associated with Bean Yellow Mosaic Virus (BYMV) resistance phenotype in common bean (Phaseolus vulgaris L.) (Hart and Griffiths, 2015). A preexisting linkage map of Chickpea (*Cicer arietinum* L.) was extended with 743 SNP loci identified by GBS, allowing the saturation of a QTL hot-spot for drought tolerance with new SNP markers (Jaganathan et al., 2015). The efficiency of GBS in genome mapping and QTL analysis was also shown on a woody species with limited genomic resources. When applied on a blackcurrant (*Ribes nigrum* L.) F_1 mapping population, GBS enabled the construction of a high-resolution linkage map and identification of QTLs controlling berry weight and soluble sugar content (Russel et al., 2014).

1.5. Aim of the study

The available pool of sesame-specific molecular markers is limited due to the relatively recent onset of molecular genetic research in this valuable oil crop species. Therefore, no significant progress has been achieved in molecular breeding in the crop to establish high-yielding, superior cultivars. The aim of this work was to develop novel molecular genetic tools in sesame, in order to aid the accumulation of knowledge for understanding the genetic control of important agronomic and metabolic traits. Toward this aim, pyrosequencing and GBS approaches were employed for high-throughput SSR

and SNP marker development, respectively. Molecular genetic diversity and population structure of a collection of world accessions were analyzed using the newly identified SSR markers. It is expected that valuable information will be produced from the combined analysis of the diversity and ancestral inference, which will be useful for parental selection in future sesame breeding schemes. An intraspecific sesame RIL population, whose parents were identified as distinct genotypes and were shown to display variation in metabolite composition, was used in GBS analysis for SNP identification and mapping. Genotypic data obtained from GBS were used for the construction of an intraspecific genetic linkage map of the sesame genome. Genetic linkage map and the genotypic data of the RIL population are expected to serve as the necessary tools for mapping important metabolic and agronomic traits in the sesame genome.

CHAPTER 2

MATERIALS AND METHODS

2.1. Plant material and DNA isolation

A collection of *S. indicum* accessions, consisting of 78 landraces and 15 cultivars, was used as plant material in genetic diversity and population structure analyses (Table 2.1). 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. *S. mulayanum* acc. COL/INDIA/1992/MAFF/0161 seeds were obtained from Dr. Makoto Kawase, National Institute of Agrobiological Sciences (NIAS), Japan.

An intraspecific recombinant inbred line population (F_6-RILs) derived from the cross *S. indicum* (Acc. No. 95-223) x *S. indicum* (Acc. No. 92-3091) was used as plant material for GBS analysis. Seeds of the F_6-RIL population were contributed by Dr. Petr Karlovsky, University of Gottingen, Gottingen, Germany.

Ten seeds per sesame accession (three seeds per F_6-RIL) were planted and grown in soil containing peat moss, perlite and natural fertilizer in the greenhouse facility at Izmir Institute of Technology. Genomic DNA from each accession was isolated from liquid nitrogen-frozen ground leaf tissue pooled from ten plants harvested at the two to four-leaf stage. DNA extraction from the world accession collection was done using the Wizard Magnetic 96 Plant System (Promega Corp., Madison,WI, USA) with the Beckman Coulter Biomek NX Workstation (Beckman Coulter, Brea, CA,

USA) according to the manufacturer's instructions. Genomic DNA from 91 RILs and parental accessions was isolated from liquid nitrogen-frozen ground leaf tissue using the NucleoSpin Plant II Kit (Macherey Nagel, Duren, Germany), according to the manufacturer's instructions.

Table 2.1. Sesame material used in the study. Accessions are listed in order of analysis.

Genotype[†]	Plant introduction	Origin	Landrace/Cultivar
1 ‡	PI167115	Turkey, Adana	Landrace
2 ‡	PI161385	Korea	Landrace
3 ‡	PI154298	Mexico	Landrace
4 ‡	PI250099	Egypt	Landrace
5 ‡	PI543241	Bolivia	Landrace
6 ‡	PI229668	Argentina	Landrace
7 ‡	PI263441	Japan, Honshu	Landrace
8 ‡	PI304259	Thailand	Landrace
9 ‡	PI207665	Morocco	Landrace
10 ‡	PI490024	Thailand	Landrace
11 ‡	PI234427	China	Landrace
12 ‡	PI433863	Nigeria	Landrace
13 ‡	PI239001	Greece, Rhodes	Landrace
14 ‡	PI323306	Pakistan	Landrace
15 ‡	PI251294	Jordan	Landrace
16 ‡	PI254698	South America	Landrace
17 ‡	PI198158	Former USSR	Landrace
18 ‡	PI179485	Iraq	Landrace
19 ‡	PI158769	Venezuela	Landrace
20 ‡	PI226567	Ethiopia	Landrace
21 ‡	PI601234	United States	Cultivar
22 ‡	PI198156	Iraq	Landrace
23 ‡	PI561704	Mexico	Cultivar
24 ‡	PI200428	Pakistan	Landrace
25 ‡	PI490114	Sudan	Cultivar
26 ‡	PI186511	Nigeria	Landrace
27 ‡	PI211627	Afghanistan	Landrace
28 ‡	PI231033	Mozambique	Landrace
29 ‡	PI164142	India	Landrace
30 ‡	PI184671	Liberia	Landrace
31	PI306695	India	Landrace
32	PI207664	Morocco	Landrace

(Cont. on next page)

Table 2.1 (cont.)

Genotype[†]	PI	Origin	Landrace/Cultivar
33 [‡]	PI250029	Iran	Landrace
34 [‡]	PI229667	Argentina	Landrace
35 [‡]	PI250030	Iran	Landrace
36 [‡]	PI153509	Venezuela	Landrace
37 [‡]	PI158038	China	Landrace
38 [‡]	PI203150	Jordan	Landrace
39 [‡]	PI189082	Cameroon	Landrace
40 [‡]	PI643459	Tajikistan	Landrace
41 [‡]	PI258372	Former USSR	Cultivar
42 [‡]	PI200427	Pakistan	Landrace
43 [‡]	PI209965	Ethiopia	Landrace
44 [‡]	PI599444	United States	Cultivar
45 [‡]	PI234424	China	Landrace
46 [‡]	PI195122	China	Landrace
47 [‡]	PI254705	United States	Landrace
48 [‡]	PI157155	India	Landrace
49 [‡]	PI207667	Morocco	Landrace
50 [‡]	PI198155	Egypt	Landrace
51 [‡]	PI211088	Afghanistan	Landrace
52 [‡]	PI231034	Mozambique	Landrace
53 [‡]	PI156618	China	Landrace
54 [‡]	PI490072	Korea, South	Cultivar
55 [‡]	PI253984	Syria	Landrace
56 [‡]	PI186509	Nigeria	Landrace
57 [‡]	PI210687	Somalia	Landrace
58 [‡]	PI189081	Cameroon	Landrace
59 [‡]	PI238988	Greece, Rhodes	Landrace
60 [‡]	PI189229	Belgian Congo	Landrace
61 [‡]	PI163595	Guatemala	Landrace
62 [‡]	PI321096	Kenya	Landrace
63 [‡]	PI253424	Israel	Landrace
64 [‡]	PI251704	Former USSR	Landrace
65 [‡]	PI224663	Libya	Landrace
66 [‡]	PI288852	Nepal	Landrace
67 [‡]	PI238430	Turkey, Izmir	Landrace
68 [‡]	PI200106	Myanmar	Landrace
69 [‡]	PI254703	Venezuela	Landrace
70[§]	TR38356	Turkey, Tekirdag	Landrace
71[§]	Golmarmara	Turkey	Landrace
72[§]	Ozberk	Turkey	Landrace

(Cont. on next page)

Table 2.1 (cont.)

Genotype[†]	PI	Origin	Landrace/Cultivar
73[§]	Tan99	Turkey	Cultivar
74[§]	Cumhuriyet99	Turkey	Cultivar
75[§]	Osmanli99	Turkey	Cultivar
76[§]	Kepsut99	Turkey	Cultivar
77[§]	Orhangazi99	Turkey	Cultivar
78[‡]	PI177072	Turkey, Eskisehir	Landrace
79[‡]	PI170753	Turkey, Canakkale	Landrace
80[§]	PI238431	Turkey, Manisa	Landrace
81[‡]	PI167248	Turkey, Adana	Landrace
82[‡]	PI205229	Turkey, Izmir	Landrace
83[§]	PI238481	Turkey, Adiyaman	Landrace
84[§]	PI238420	Turkey, Izmir	Landrace
85[§]	PI238445	Turkey, Manisa	Landrace
86[§]	PI238450	Turkey, Manisa	Landrace
87[§]	PI238433	Turkey, Mersin	Landrace
88[§]	PI240844	Turkey, Mersin	Landrace
89[‡]	PI205225	Turkey, Antalya	Landrace
90[§]	PI238453	Turkey, Canakkale	Landrace
91[¶]	95-223	Africa	Landrace
92[¶]	92-3091	Korea	Landrace
93[§]	Muganli57	Turkey	Landrace
94[#]	*S.mulayanum*	India	*S. mulayanum*

[†]Seed sources are coded. [‡]USDA: U.S. Department of Agriculture; [§]AARI: Aegean Agricultural Research Institute; [¶]UG: University of Gottingen; [#]NIAS: National Institute of Agrobiological Sciences.

2.2. DNA sequencing and SSR validation

For SSR identification, total genomic DNA of *S. indicum* cv. Muganli 57 was subjected to pyrosequencing. Pyrosequencing was done with a Roche 454 GS-FLX sequencer and performed by 454 Lifesciences Corp. (Branford, CT, USA). SSR validation was done using the dye-terminator sequencing method. PCR products, purified with the DNA Clean & Concentrator^{TM}-5 Kit (Zymo Research, Irvine, CA, USA), 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 consisted of 30 cycles of 96 °C 20 sec, 50 °C 20 sec, 60 °C 4 min. The reaction mixture for each SSR amplicon

was purified using ZR DNA Sequencing Clean-up Kit™ (Zymo Research) and eluted in 30 μL of sample loading solution (Beckman Coulter). Sequencing PCR products were run on a Beckman CEQ8800 capillary electrophoresis system 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).

2.3. Pyrosequencing data processing and sequence assembly

Adapter and linker sequences were removed from the raw sequence reads to facilitate genome assembly. Because most assembly tools cannot directly process SFF files, Standard Flowgram Format (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, available at 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, a whole genome shotgun and EST sequence assembler (Chevreux et al., 2004) Version 3.4, 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 non-default parameters. Customized sequence assembly parameters are provided in (Table 2.1). Raw sequence processing, assembly, SSR detection and primer design were carried out by the Allmer Lab at Izmir Institute of Technology. The assembled sequences are available on NCBI at http://www.ncbi.nlm.nih.gov/bioproject/271288.

2.4. SSR detection and primer design

Contig assemblies and singleton sequences were analyzed for SSR identification with SiSeeR software (http://bioinformatics.iyte.edu.tr/index.php?n=Softwares.SiSeeR). The minimum number of repeats required to identify perfect SSRs were: ten 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 non-redundant 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 Table 2.2. In order 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.

Table 2.2. Customized parameters of MIRA and Primer3 softwares.

MIRA Parameters		**Primer3 Parameters**	
Number of passes	10	Primer task	Generic
SKIM each pass	10	Primer pick left primer	1
Spoiler detection last pass only	Yes	Primer pick internal oligo	1
Bases per hash	16	Primer pick right primer	1
Percent required	73	Primer optimum size	21
Minimum relative score	73	Primer minimum size	18
Nasty repeat ratio	2000	Primer maximum size	27
Possible vector leftover clip	10	Primer max ns accepted	1
Advanced contig editing	On	Primer product size range	100 - 300
Minimum overlap	20	Primer minimum Tm	50˚C
Minimum neighbor quality	20	Primer optimum Tm	55˚C
Minimum repeat coverage	300	Primer maximum Tm	60˚C
Number of bases		Primer pair max difference Tm	5˚C
free of sequencing errors	18	Primer minimum GC	40%
		Primer optimum GC percentage	50%
		Primer maximum GC percentage	60%
		Primer num return	1
		Primer pick anyway	0

2.5. SSR amplification

SSR alleles were amplified in 20 µL reaction mixtures containing 1X PCR buffer, 1.5 mM $MgCl_2$, 0.25 mM of each deoxyribonucleotide triphosphate (dNTP) (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. PCR products were then run on a

Fragment Analyzer™ (Advanced Analytical, Ames, IA, USA) capillary electrophoresis system using the DNF-900 dsDNA Reagent Kit (Advanced Analytical) according to the manufacturer's instructions. SSR alleles were visualized and scored using the PROSize 2.0™ software version 1.2.1.1 (Advanced Analytical).

2.6. SSR data analysis

SSR 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 = $(\sum_{i=1}^{n} 2fi(1-fi))/n$ (Roldan-Ruiz et al., 2000),

where *fi* is the frequency of band presence for the i^{th} 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 MCMC (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 ≥ 0.60, in order 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 which 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.

2.7. GBS library preparation and sequencing

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

2.8. Sequence alignment and SNP calling

Raw sequence processing, alignment and SNP calling were performed at the University of Wisconsin-Madison Biotechnology Center. Raw sequence reads were converted to a FASTQ file by CASAVA 1.8.2 (Illumina Inc.) for further processing. To initiate the data analysis with the GBS Discovery Pipeline (Glaubitz et al., 2014) of TASSEL Version 3.0 (Bradbury et al., 2007), the FASTQ file and barcode key file that lists the plate layout and barcodes for each sample were used as input files. Using the FastqToTagCountPlugin of the pipeline, reads that began with the expected barcodes followed by an *Ape*KI cut site remnant (CWGC) were trimmed to 64 bases. Sequence reads with N (unidentified base) in the first 64 bases after the barcode were eliminated. Reads with an intact enzyme cut site or the beginning of the common adapter were truncated and padded to 64 bases with poly A. The reads were then sorted to merge the redundant reads into single tags and resultant tags were listed as a tagCount file by the plugin. MergeMultipleTagCountPlugin produced the merged tagCount file, the file was converted to FASTQ format by the TagCountToFastqPlugin to be used as the input file for tag alignment to the draft genome assembly by bowtie2 plugin. Pseudomolecule sequences of the assembly were downloaded from the Sinbase database (Wang et al., 2014). The genome assembly comprised of 16 pseudomolecules of assembled scaffolds

and one group of concatenated contigs (group 17) with stretches of 100 N as the spacer. The output of the alignment was converted to a TOPM (Tags On Physical Map) file by SAMConverterPlugin for SNP calling from the alignment. Sequence reads sorted and demultiplexed according to their barcode adapters by the FastqToTBTPlugin were kept as a TBT file (Tags By Taxa). TOPM and TBT files were used by the TagsToSNPByAlignmentPlugin for SNP calling. Non-default parameters used for SNP calling were: Minimum value of F (inbreeding coefficient = 1-Ho/He) [mnF]: 0.8, Minimum minor allele count (default: 10) [mnMAC]: 100000. SNPs that pass the mnMAC threshold were kept in the output HapMap file for each sesame pseudomolecule. Duplicate SNPs in the HapMap files were merged by the MergeDuplicateSNPsPlugin. In order to allow heterozygosity detection in SNP loci, callHets option of the plugin was switched to True. Threshold for genotypic mismatch rate (misMat) was set as 0.1. Merged SNPs were filtered with the GBSHapMapFiltersPlugin. Non-default SNP filtering parameters were: Minimum taxon coverage [mnTCov]: 0.01, Minimum site coverage [mnSCov]: 0.5, filtering for SNPs in statistically significant LD (Linkage Disequilibrium) with at least one neighboring SNP [hLD]: True, Minimum R^2 value for the LD filter [-mnR2]: 0.2, Minimum Bonferroni-corrected p-value for the LD filter [-mnBonP]: 0.005.

2.9. Genetic linkage map construction

SNP genotype data obtained from GBS analysis were used for the construction of a genetic linkage map. SSR markers were used in conjunction with SNPs for map construction. SSR markers to be used for linkage analysis were determined by performing a parental survey using our newly developed markers. SSR markers found polymorphic between the parental genotypes were genotyped in the F_6-RIL population. SSR amplification mixtures and thermal cycling conditions were applied as described in Section 2.5.

A genetic linkage map was constructed using the JoinMap 4.0 (Van Ooijen 2006) computer program. Marker order was determined with the regression mapping algorithm using a maximum recombination frequency threshold of 0.40. Minimum logarithm of odds (LOD) threshold was set as 6 and a goodness-of-fit jump threshold for loci removal was set as 5. The ripple command was adjusted to confirm marker

order after the addition of each locus. Map distances were calculated with the Kosambi mapping function (Kosambi, 1943). Linkage groups were visualized with the MapChart 2.3 computer program (Voorrips, 2002).

CHAPTER 3

RESULTS AND DISCUSSION

3.1. High-throughput Simple Sequence Repeat (SSR) marker development in sesame

3.1.1. SSR 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 work, a pyrosequencing approach was used for high-throughput SSR marker development in sesame.

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 sequencing system. Removal of the adapter 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 adapter and linker removal, 616,210 reads (56.3% of the cleaned reads) could be assembled into 136,257 contigs (Table 3.1). 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). Thus, a good portion of the sesame genome was covered by the sequence assembly. The weighted median contig length (N50) of the assembly was 671 nucleotides.

Table 3.1. 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

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 density in the contig assembly was 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 kb 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 kb 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* genome was reported as one SSR per 6 kb, 1.1 kb 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.) 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.

The length of the identified SSRs 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 the most abundant, representing 48.5% of all SSRs (Table 3.2). 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 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 3.2). 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.

Table 3.2. Simple sequence repeat types in sesame genome.

Motif Length	**Number of Occurrence**	**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

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 Arabidopsis, purple false brome (*Brachypodium distachyon* L.), sorghum, rice (*Oryza sativa* L.), barrel clover (*Medicago truncatula*) and poplar (*Populus trichocarpa*). While A/T was the predominant mononucleotide repeat (81.7%), AT was the most abundant dinucleotide repeat (32.5%), followed by TA (19.2%) in the pool of identified SSRs. AT-rich repeats were also 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 3.3). 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., 2014a). In addition, AG/CT was consistently found as the predominant motif in sesame genic SSRs. Thus, it appears that the relative abundance of simple sequence repeat types differ between genic and genomic SSRs in sesame.

Table 3.3. Most abundant 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

Evolutionary forces apply differently to coding and non-coding 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 over-represented 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).

For successful SSR primer design, flanking sequences of sufficient length should be present. Compared to singletons, contigs usually provide longer flanking sequences and, therefore, allow greater flexibility in the primer design process. Thus, in order 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 (*S. indicum* L. cv. Muganli 57) and a wild *Sesamum* (*S. mulayanum*) accession. A total of 933 primers (93.3%) successfully amplified PCR products from *S. indicum* while 849 (84.9%) amplified products from both genotypes. Thus, 91% of the experimentally validated SSRs were applicable to *S. mulayanum*, corresponding to a very high rate of marker transferability. 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 et al., 2003). Among the 849 markers that yielded successful PCR amplification from *S. mulayanum*, 221 (26%) were polymorphic between *S. indicum* and *S. mulayanum* (Table 3.4).

In order to validate the presence of the expected SSR motifs within the amplicons, eight amplicons from cv. Muganli 57 were sequenced with the dye-terminator method (Figure 3.1). Based on the data from pyrosequencing analysis, the repeat motifs expected in the amplicons from cv. Muganli 57 were $(ATTT)^3$ for SiSSR-104, $(AATT)^3$ for SiSSR-210, $(TAAT)^3$ for SiSSR-214, $(GAAG)^3$ for SiSSR-216, $(AT)^8$ for SiSSR-219, $(TAAA)^3$ for SiSSR-236, $(TTTA)^3$ for SiSSR-249 and $(TTC)^3$ for SiSSR-252. Sequence analysis revealed that all eight sequences contained the expected SSR motifs (Figure 3.1), proving the identity of our primers as SSR markers, thus

validating the reliability of our SSR marker design approach. Primer and transferability information for the SSR markers is available at http://plantmolgen.iyte.edu.tr/data/.

In this study, *S. mulayanum* was not detected as an outgroup in the neighbor-joining analysis (results shown in Section 3.2) 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 which 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.

Table 3.4. List of polymorphic Simple Sequence Repeat (SSR) markers between cultivated and wild sesame accessions.

Polymorphic markers between *S. indicum* and *S.mulayanum*					
SiSSRg-9	SiSSRg-165	SiSSRg-344	SiSSRg-466	SiSSRg-623	SiSSRg-774
SiSSRg-13	SiSSRg-178	SiSSRg-347	SiSSRg-471	SiSSRg-630	SiSSRg-776
SiSSRg-14	SiSSRg-181	SiSSRg-349	SiSSRg-472	SiSSRg-635	SiSSRg-780
SiSSRg-18	SiSSRg-182	SiSSRg-350	SiSSRg-475	SiSSRg-640	SiSSRg-781
SiSSRg-20	SiSSRg-183	SiSSRg-351	SiSSRg-477	SiSSRg-642	SiSSRg-792
SiSSRg-26	SiSSRg-187	SiSSRg-356	SiSSRg-482	SiSSRg-646	SiSSRg-797
SiSSRg-34	SiSSRg-199	SiSSRg-357	SiSSRg-483	SiSSRg-656	SiSSRg-799
SiSSRg-42	SiSSRg-202	SiSSRg-362	SiSSRg-490	SiSSRg-665	SiSSRg-806
SiSSRg-43	SiSSRg-205	SiSSRg-368	SiSSRg-491	SiSSRg-666	SiSSRg-820
SiSSRg-47	SiSSRg-209	SiSSRg-380	SiSSRg-495	SiSSRg-670	SiSSRg-823
SiSSRg-55	SiSSRg-211	SiSSRg-384	SiSSRg-496	SiSSRg-673	SiSSRg-830
SiSSRg-59	SiSSRg-222	SiSSRg-385	SiSSRg-497	SiSSRg-674	SiSSRg-847
SiSSRg-68	SiSSRg-223	SiSSRg-386	SiSSRg-499	SiSSRg-675	SiSSRg-859
SiSSRg-70	SiSSRg-226	SiSSRg-393	SiSSRg-502	SiSSRg-678	SiSSRg-863
SiSSRg-71	SiSSRg-232	SiSSRg-402	SiSSRg-506	SiSSRg-683	SiSSRg-864
SiSSRg-75	SiSSRg-233	SiSSRg-404	SiSSRg-507	SiSSRg-688	SiSSRg-867
SiSSRg-80	SiSSRg-236	SiSSRg-409	SiSSRg-509	SiSSRg-692	SiSSRg-876
SiSSRg-81	SiSSRg-249	SiSSRg-410	SiSSRg-510	SiSSRg-694	SiSSRg-901
SiSSRg-85	SiSSRg-252	SiSSRg-413	SiSSRg-513	SiSSRg-704	SiSSRg-902
SiSSRg-97	SiSSRg-253	SiSSRg-414	SiSSRg-514	SiSSRg-708	SiSSRg-907
SiSSRg-98	SiSSRg-263	SiSSRg-415	SiSSRg-519	SiSSRg-709	SiSSRg-910
SiSSRg-105	SiSSRg-268	SiSSRg-423	SiSSRg-524	SiSSRg-721	SiSSRg-912
SiSSRg-107	SiSSRg-269	SiSSRg-428	SiSSRg-526	SiSSRg-724	SiSSRg-913
SiSSRg-111	SiSSRg-280	SiSSRg-434	SiSSRg-531	SiSSRg-725	SiSSRg-915
SiSSRg-113	SiSSRg-283	SiSSRg-436	SiSSRg-549	SiSSRg-730	SiSSRg-916
SiSSRg-114	SiSSRg-286	SiSSRg-437	SiSSRg-556	SiSSRg-731	SiSSRg-918
SiSSRg-116	SiSSRg-295	SiSSRg-440	SiSSRg-566	SiSSRg-736	SiSSRg-919
SiSSRg-128	SiSSRg-301	SiSSRg-442	SiSSRg-568	SiSSRg-737	SiSSRg-930
SiSSRg-130	SiSSRg-302	SiSSRg-443	SiSSRg-575	SiSSRg-738	SiSSRg-933
SiSSRg-131	SiSSRg-313	SiSSRg-444	SiSSRg-582	SiSSRg-742	SiSSRg-942
SiSSRg-135	SiSSRg-315	SiSSRg-448	SiSSRg-583	SiSSRg-743	SiSSRg-944
SiSSRg-139	SiSSRg-316	SiSSRg-450	SiSSRg-584	SiSSRg-749	SiSSRg-945
SiSSRg-140	SiSSRg-318	SiSSRg-453	SiSSRg-590	SiSSRg-754	SiSSRg-968
SiSSRg-148	SiSSRg-326	SiSSRg-459	SiSSRg-592	SiSSRg-760	SiSSRg-975
SiSSRg-152	SiSSRg-329	SiSSRg-460	SiSSRg-606	SiSSRg-761	SiSSRg-977
SiSSRg-153	SiSSRg-333	SiSSRg-464	SiSSRg-613	SiSSRg-763	SiSSRg-985
SiSSRg-160	SiSSRg-338	SiSSRg-465	SiSSRg-621	SiSSRg-765	

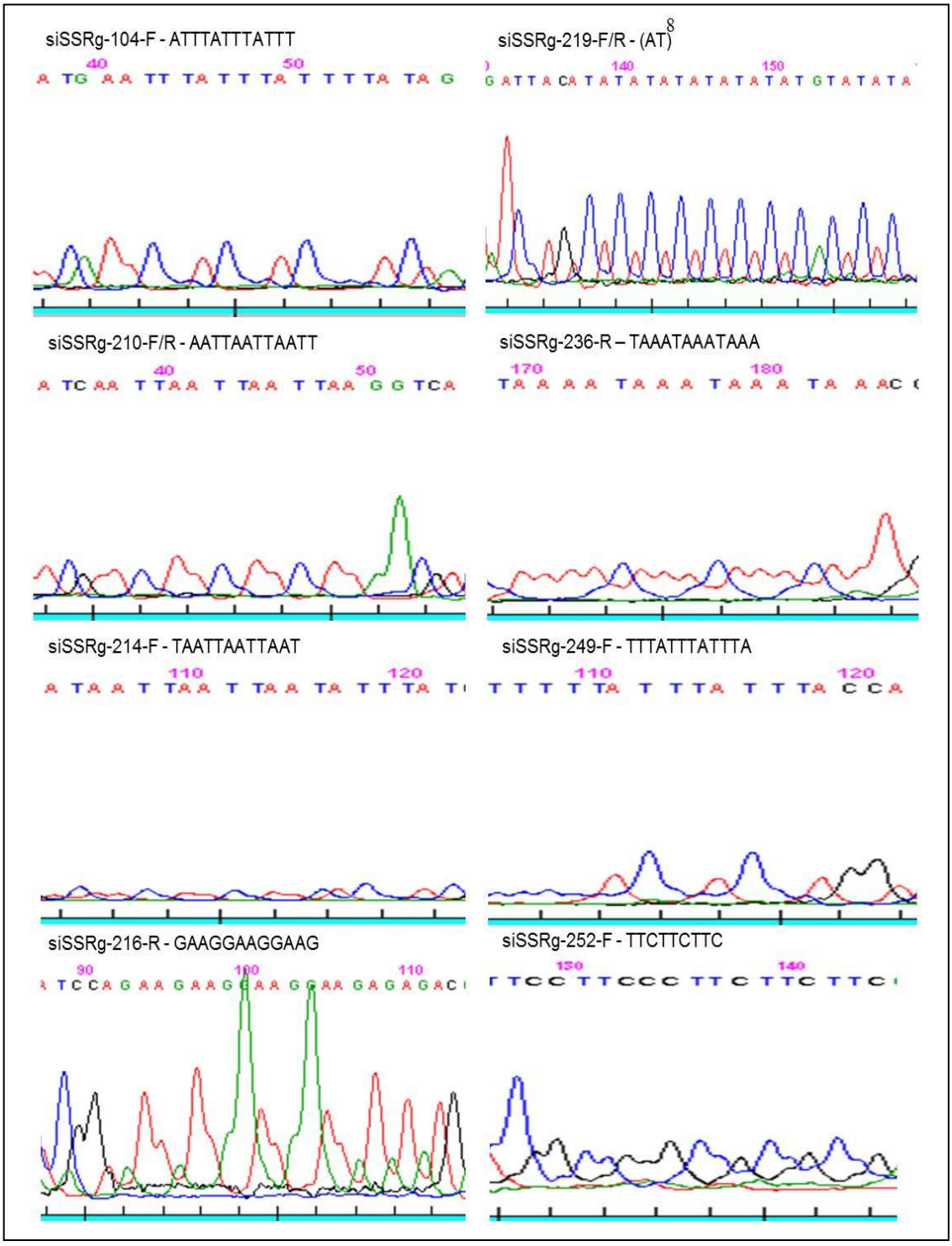

Figure 3.1. Sequencing electropherograms displaying the simple sequence repeats.

3.2. Assessment of the genetic diversity and population structure of a sesame world collection

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 2.1). 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 3.5). 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 (Figure 3.2). 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 (0) calculated for the two non-polymorphic markers (siSSRg-48 and siSSRg-933) (Table 3.5). 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.

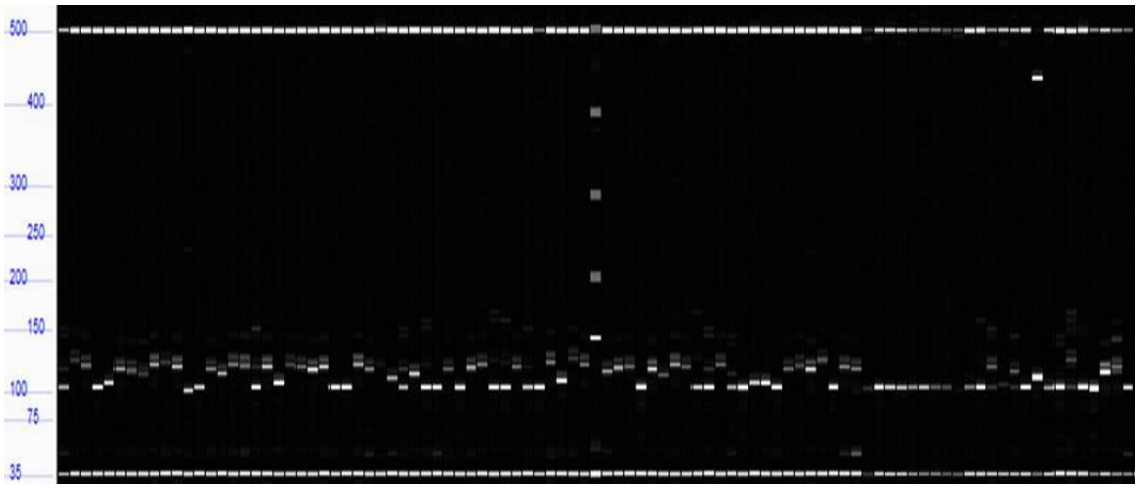

Figure 3.2. Virtual gel image produced by PROSize 2.0™ software displaying the SSR alleles amplified by the marker siSSR-621. SSR alleles are aligned by the software with respect to the standard upper and lower alignment markers which are 500 bp and 35 bp, respectively.

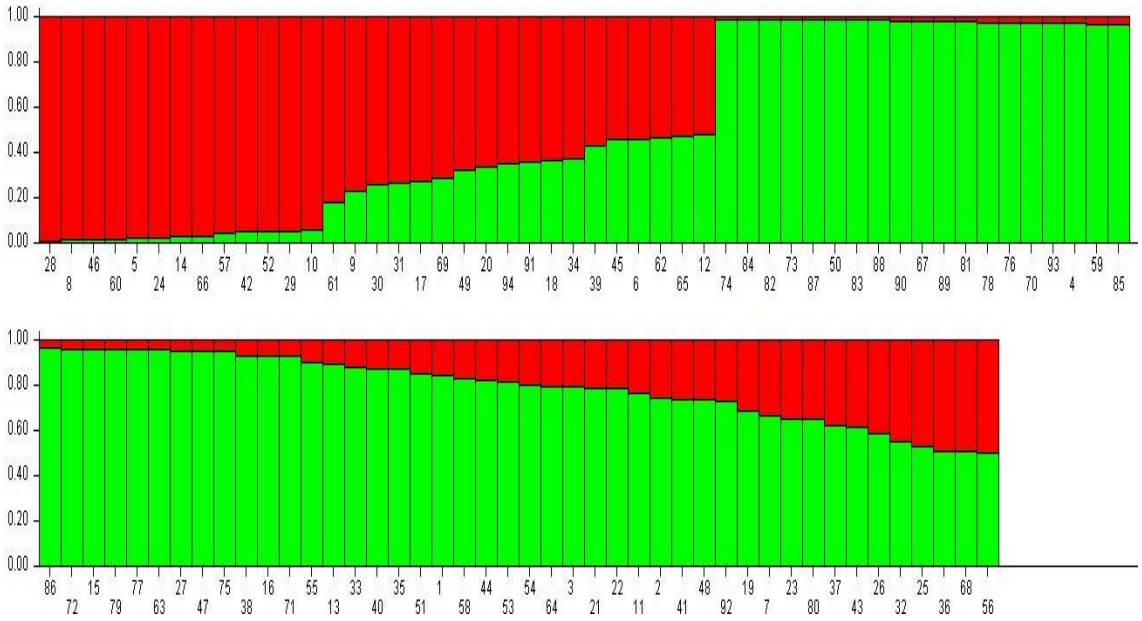

Figure 3.3. Bar plot displaying the estimated genetic structure of sesame accessions with two subpopulations. Each accession is represented by a vertical bar, partitioned according to estimated membership in the two subpopulations. Subpopulation 1 fraction is shown in green and Subpopulation 2 fraction is shown in red. Numbering of accessions follows the order in analysis.

SSR 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, in order to determine the relatedness and diversity of the accessions. Population structure analysis suggested a model that assigned the accessions to two subpopulations (Figure 3.3). As a result, 57 and 25 accessions were assigned to Subpopulations 1 and 2, respectively, and 12 accessions were considered as admixed (Table 3.6). While all of the Turkish accessions and the majority of accessions from the Middle East (nine of ten 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. *S. 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 (Figure 3.4). 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 (PI 205229) and Mozambique (PI231033), and the lowest (0.07), calculated between a Turkish accession (PI 205229) and a Turkish registered cultivar (Cumhuriyet 99). *S. indicum* accessions fell into three clusters (Clusters A, B and C) in the dendrogram (Figure 3.4).

Table 3.5. Simple Sequence Repeat (SSR) markers used for the molecular genetic analyses.

SSR Marker	Repeat Motif (5' to 3')	Number of Alleles	Gene diversity	SSR Marker	Repeat Motif (5' to 3')	Number of Alleles	Gene diversity
SiSSRg-1	(GTG/CAC)4	5	0.30 ± 0.08	SiSSRg-635	(TGATT/AATCA)3	2	0.01 ± 0.01
SiSSRg-4	(ATTT/AAAT)4	3	0.18 ± 0.05	SiSSRg-640	(AT)6	4	0.26 ± 0.13
SiSSRg-17	(ATG/CAT)7	2	0.32 ± 0.02	SiSSRg-654	(TTTC/GAAA)3	7	0.14 ± 0.06
SiSSRg-42	(GAG/CTC)5	4	0.18 ± 0.09	SiSSRg-666	(ATGA/TCAT)3	9	0.12 ± 0.05
SiSSRg-45	(TGG/CCA)4	2	0,14	SiSSRg-670	(TG/CA)7	4	0.31 ± 0.08
SiSSRg-47	(CT/AG)7	3	0.08 ± 0.02	SiSSRg-679	(CTTTT/AAAAG)3	3	0.20 ± 0.09
SiSSRg-48	(AGA/TCT)4	1	0	SiSSRg-692	(GCA/TGC)4	9	0.13 ± 0.07
SiSSRg-51	(TATG/CATA)3	4	0.28 ± 0.13	SiSSRg-708	(AATT)3	6	0.23 ± 0.08
SiSSRg-111	(AATT)3	2	0.39 ± 0.03	SiSSRg-733	(GT/AC)8	3	0.06 ± 0.02
SiSSRg-178	(TTG/CAA)5	7	0.17 ± 0.06	SiSSRg-767	(AT)7	2	0,02
SiSSRg-223	(TTA/TAA)6	3	0.30 ± 0.14	SiSSRg-786	(TAG/CTA)4	4	0.11 ± 0.06
SiSSRg-236	(TAAA/TTTA)3	3	0.34 ± 0.14	SiSSRg-801	(TGAAA/TTTCA)4	3	0.34 ± 0.16
SiSSRg-346	(CCA/TGG)5	4	0.21 ± 0.06	SiSSRg-825	(CTCCGC/GCGGAG)3	5	0.04 ± 0.01
SiSSRg-392	(CCCCA/TGGGG)4	4	0.22 ± 0.09	SiSSRg-859	(ACTCAC/GTGAGT)3	2	0.04 ± 0.04
SiSSRg-393	(CAA/TTG)4	5	0.21 ± 0.10	SiSSRg-863	(TAT/ATA)4	4	0.20 ± 0.07
SiSSRg-410	(CT/AG)11	5	0.18 ± 0.09	SiSSRg-892	(AT)10	10	0.16 ± 0.04
SiSSRg-422	(AT)11	6	0.19 ± 0.08	SiSSRg-924	(AT)7	4	0.29 ± 0.07
SiSSRg-437	(GTTTT/AAAAC)3	7	0.18 ± 0.07	SiSSRg-925	(ATC/GAT)4	5	0.22 ± 0.11
SiSSRg-485	(TG/CA)8	6	0.20 ± 0.07	SiSSRg-933	(TAC/GTA)4	1	0
SiSSRg-491	(TTAT/ATAA)3	4	0.22 ± 0.07	SiSSRg-945	(TGATCA)4	2	0.34 ± 0.06
SiSSRg-549	(TACA/TGTA)3	4	0.26 ± 0.11	SiSSRg-949	(GAA/TTC)4	4	0.14 ± 0.04
SiSSRg-575	(ATGT/ACAT)5	2	0.48 ± 0.02	SiSSRg-975	(TC/GA)7	3	0.32 ± 0.14
SiSSRg-606	(GGAGTA/TACTCC)4	6	0.21 ± 0.08	SiSSRg-985	(TA)7	7	0.13 ± 0.06
SiSSRg-621	(TA)6	17	0.11 ± 0.03	SiSSRg-991	(TC/GA)7	2	0.31 ± 0.04
SiSSRg-634	(GGGGT/ACCCC)3	5	0.23 ± 0.11				

Table 3.6. Cluster assignments of the world accession collection according to Structure and DARwin analyses. Subpopulation assignments are based on Structure analysis. Cluster assignments based on the neighbor-joining dendrogram are also displayed.

			Inferred Ancestry Subpopulation			
Genotype	**PI**	**Origin**	**1**	**2**	**Subpop. Assign.**	**Cluster Assign.**
1	PI167115	Turkey, Adana	0.843	0.157	1	A1
2	PI161385	Korea	0.749	0.251	1	A2
3	PI154298	Mexico	0.794	0.206	1	A2
4	PI250099	Egypt	0.972	0.028	1	A1
5	PI543241	Bolivia	0.026	0.974	2	B
6	PI229668	Argentina	0.462	0.538	Admixed	B
7	PI263441	Japan, Honshu	0.666	0.334	1	A1
8	PI304259	Thailand	0.015	0.985	2	B
9	PI207665	Morocco	0.231	0.769	2	B
10	PI490024	Thailand	0.064	0.936	2	B
11	PI234427	China	0.771	0.229	1	A2
12	PI433863	Nigeria	0.481	0.519	Admixed	C
13	PI239001	Greece, Rhodes	0.897	0.103	1	A1
14	PI323306	Pakistan	0.03	0.97	2	B
15	PI251294	Jordan	0.962	0.038	1	A1
16	PI254698	South America	0.933	0.067	1	A1
17	PI198158	Former USSR	0.275	0.725	2	B
18	PI179485	Iraq	0.369	0.631	2	B
19	PI158769	Venezuela	0.687	0.313	1	A2
20	PI226567	Ethiopia	0.336	0.664	2	B
21	PI601234	United States	0.792	0.208	1	A2
22	PI198156	Iraq	0.786	0.214	1	A1
23	PI561704	Mexico	0.655	0.345	1	A1
24	PI200428	Pakistan	0.028	0.972	2	B
25	PI490114	Sudan	0.533	0.467	Admixed	C
26	PI186511	Nigeria	0.588	0.412	Admixed	C
27	PI211627	Afghanistan	0.957	0.043	1	A1
28	PI231033	Mozambique	0.01	0.99	2	B
29	PI164142	India	0.054	0.946	2	B
30	PI184671	Liberia	0.264	0.736	2	B
31	PI306695	India	0.268	0.732	2	B
32	PI207664	Morocco	0.551	0.449	Admixed	B
33	PI250029	Iran	0.884	0.116	1	A1
34	PI229667	Argentina	0.376	0.624	2	B
35	PI250030	Iran	0.872	0.128	1	A1
36	PI153509	Venezuela	0.513	0.487	Admixed	B

(Cont. on next page)

Table 3.6 (cont.)

			Inferred Ancestry Subpopulation			
Genotype	**PI**	**Origin**	**1**	**2**	**Subpop. Assign.**	**Cluster Assign.**
37	PI158038	China	0.626	0.374	1	A2
38	PI203150	Jordan	0.935	0.065	1	A2
39	PI189082	Cameroon	0.434	0.566	Admixed	C
40	PI643459	Tajikistan	0.873	0.127	1	A1
41	PI258372	Former USSR	0.742	0.258	1	A1
42	PI200427	Pakistan	0.051	0.949	2	B
43	PI209965	Ethiopia	0.616	0.384	1	A
44	PI599444	United States	0.823	0.177	1	A2
45	PI234424	China	0.46	0.54	Admixed	B
46	PI195122	China	0.015	0.985	2	B
47	PI254705	United States	0.954	0.046	1	A1
48	PI157155	India	0.736	0.264	1	B
49	PI207667	Morocco	0.328	0.672	2	B
50	PI198155	Egypt	0.987	0.013	1	A1
51	PI211088	Afghanistan	0.851	0.149	1	A1
52	PI231034	Mozambique	0.052	0.948	2	B
53	PI156618	China	0.817	0.183	1	A1
54	PI490072	Korea, South	0.801	0.199	1	A2
55	PI253984	Syria	0.904	0.096	1	A1
56	PI186509	Nigeria	0.502	0.498	Admixed	C
57	PI210687	Somalia	0.046	0.954	2	B
58	PI189081	Cameroon	0.834	0.166	1	A1
59	PI238988	Greece, Rhodes	0.971	0.029	1	A1
60	PI189229	Belgian Congo	0.017	0.983	2	B
61	PI163595	Guatemala	0.18	0.82	2	B
62	PI321096	Kenya	0.467	0.533	Admixed	C
63	PI253424	Israel	0.958	0.042	1	A1
64	PI251704	Former USSR	0.796	0.204	1	A1
65	PI224663	Libya	0.476	0.524	Admixed	B
66	PI288852	Nepal	0.03	0.97	2	B
67	PI238430	Turkey, Izmir	0.98	0.02	1	A1
68	PI200106	Myanmar	0.512	0.488	Admixed	B
69	PI254703	Venezuela	0.287	0.713	2	B
70	TR38356	Turkey, Tekirdag	0.973	0.027	1	A1
71	Golmarmara	Turkey	0.933	0.067	1	A1
72	Ozberk	Turkey	0.964	0.036	1	A1
73	Tan99	Turkey	0.988	0.012	1	A1
74	Cumhuriyet99	Turkey	0.991	0.009	1	A1

(Cont. on next page)

Table 3.6 (cont.)

			Inferred Ancestry Subpopulation			
Genotype	**PI**	**Origin**	**1**	**2**	**Subpop. Assign.**	**Cluster Assign.**
75	Osmanli99	Turkey	0.953	0.047	1	A1
76	Kepsut99	Turkey	0.976	0.024	1	A1
77	Orhangazi99	Turkey	0.961	0.039	1	A1
78	PI177072	Turkey, Eskisehir	0.978	0.022	1	A1
79	PI170753	Turkey, Canakkale	0.962	0.038	1	A1
80	PI238431	Turkey, Manisa	0.652	0.348	1	A1
81	PI167248	Turkey, Adana	0.979	0.021	1	A1
82	PI205229	Turkey, Izmir	0.989	0.011	1	A1
83	PI238481	Turkey, Adiyaman	0.987	0.013	1	A1
84	PI238420	Turkey, Izmir	0.991	0.09	1	A1
85	PI238445	Turkey, Manisa	0.971	0.029	1	A1
86	PI238450	Turkey, Manisa	0.966	0.034	1	A1
87	PI238433	Turkey, Mersin	0.988	0.012	1	A1
88	PI240844	Turkey, Mersin	0.986	0.014	1	A1
89	PI205225	Turkey, Antalya	0.98	0.02	1	A1
90	PI238453	Turkey, Canakkale	0.985	0.015	1	A1
91	95-223	Africa	0.361	0.639	2	B
92	92-3091	Korea	0.732	0.268	1	A
93	Muganli57	Turkey	0.973	0.027	1	A1
94	*S.mulayanum*	India	0.354	0.646	2	B

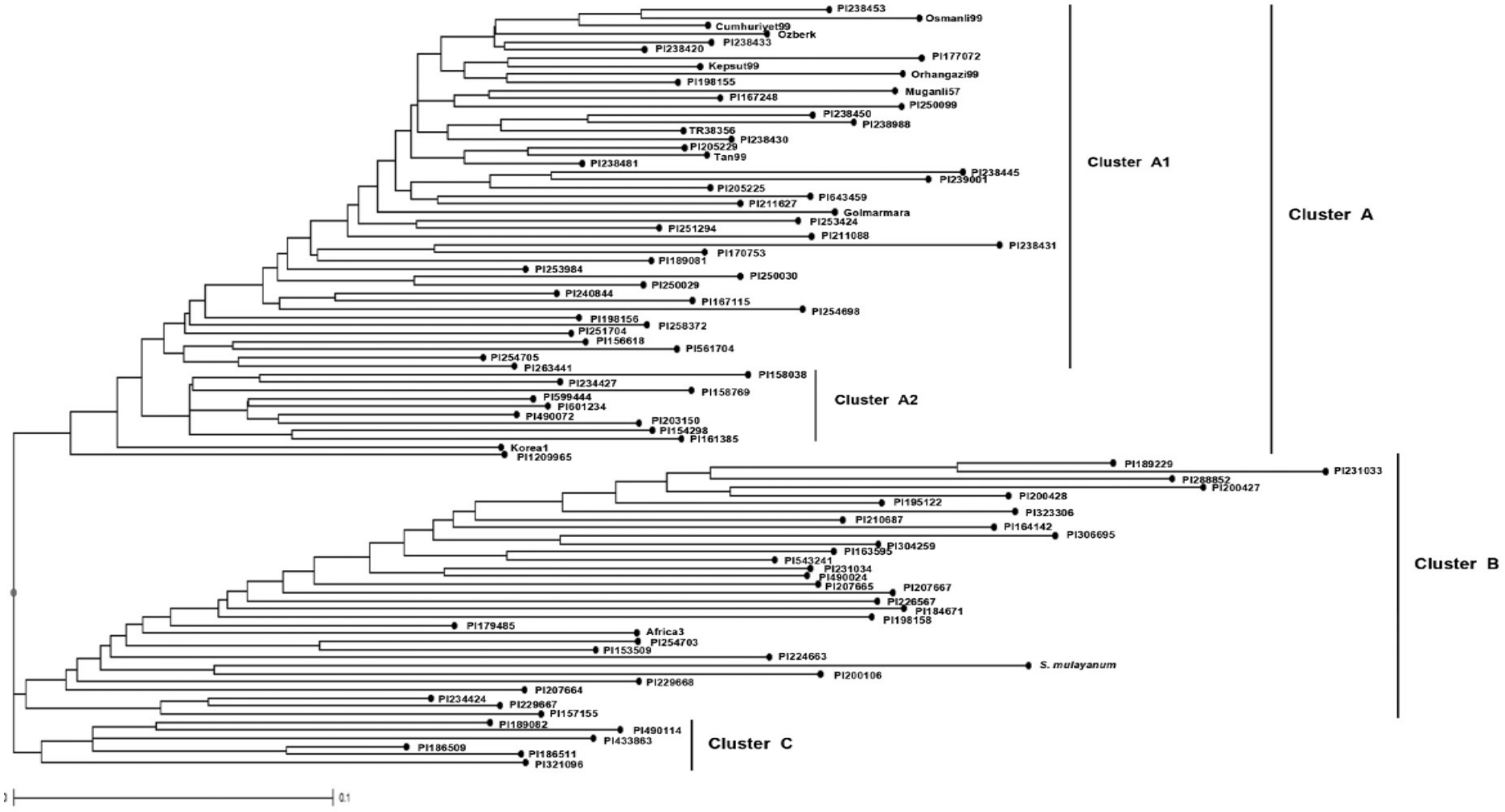

Figure 3.4. Unweighted neighbor-joining dendrogram of sesame accessions constructed using genomic simple sequence repeat markers.

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. 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.

Cluster A comprised 56 accessions, which coincided with Subpopulation 1 with the exception of a single accession (PI157155, India) which 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. 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. All of the Turkish accessions were in Cluster A1. The average pairwise dissimilarity of the Turkish accessions was 0.24, 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. 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. 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, 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.

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.

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, 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 which 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.

3.3. Construction of a genetic linkage map of the sesame genome by GBS analysis

A low rate of marker polymorphism has been consistently reported for sesame by several authors (Dixit et al., 2005; Wei et al., 2009; Wang et al., 2012, Zhang et al., 2013). In agreement with the relevant literature, only 50 out of the 933 (5.36%) experimentally validated SSR markers were found to be polymorphic between the parental accessions of the F_6-RIL population (Table 3.7). In addition, as sesame is a neglected crop species, the number of available sesame-specific markers is limited. Taken together, the low marker polymorphism rate and the limited pool of genome-specific markers stand as obstacles for the construction of a high-resolution linkage map in sesame (Zhang et al., 2013). Therefore, it is necessary to utilize novel approaches that allow simultaneous polymorphism identification and genotyping if construction of a linkage map with sufficient marker resolution is intended. In this work, a GBS approach (Elshire et al., 2011) was employed for high-throughput SNP marker development and genotyping in a F_6-RIL mapping population.

Table 3.7. List of the SSR markers polymorphic between the parental accessions of the F_6-RIL population.

	Parental alleles (bp)			Parental alleles (bp)	
SSR marker	**95-223†**	**92-3091†**	**SSR marker**	**95-223**	**92-3091**
SiSSRg-42	125	135	SiSSRg-679	208	205
SiSSRg-48	234	238	SiSSRg-682	170-173	170
SiSSRg-51	237	224	SiSSRg-686	191	184
SiSSRg-66	222	230	SiSSRg-694	120	127
SiSSRg-75	165	157	SiSSRg-708	276	265
SiSSRg-114	222	229	SiSSRg-709	191	187
SiSSRg-216	143	138-143	SiSSRg-721	170	177
SiSSRg-236	262	269	SiSSRg-738	210	202
SiSSRg-259	383	394	SiSSRg-763	168	161
SiSSRg-265	171	166	SiSSRg-791	188	176
SiSSRg-299	243	260	SiSSRg-812	195	182
SiSSRg-302	162	169	SiSSRg-844	202	206
SiSSRg-362	218	229	SiSSRg-852	229	221
SiSSRg-382	189	180	SiSSRg-855	193	190
SiSSRg-393	148	152	SiSSRg-880	192	217
SiSSRg-410	190	195	SiSSRg-892	232	239
SiSSRg-450	140	145	SiSSRg-908	152	148
SiSSRg-485	180	162	SiSSRg-924	147	139
SiSSRg-526	188	180	SiSSRg-925	182	186
SiSSRg-538	166	166-174	SiSSRg-934	195	209
SiSSRg-566	189	193	SiSSRg-949	213	217
SiSSRg-567	134	130	SiSSRg-963	239	243
SiSSRg-621	100	113	SiSSRg-968	196	209
SiSSRg-666	195	200	SiSSRg-972	162	157
SiSSRg-670	203	199	SiSSRg-991	173	163

† Plant introduction numbers correspond to the parental accessions of the F_6-RIL population.

3.3.1. Sequence filtering and tag alignment

A total of 343,970,622 raw sequence reads were obtained from sequencing of a GBS library that represents 91 F_6-RILs and two parental accessions. The number of accepted reads containing the expected barcodes and the enzyme cut site remnant was 164,433,076, comprising 47.8% of the total reads (Table 3.8). Sesame_RIL 45 was not represented in the pool of accepted reads and Sesame_RIL 117 was the least represented genotype with only 168 sequences. The most highly represented genotype was

Sesame_RIL 79 with 3,296,818 appropriately barcoded reads. The average number of reads obtained from RILs was 1,758,127. The two parental accessions, *S. indicum* Acc. No. 95-223 and Acc. No. 92-3091 yielded 1,523,490 and 2,920,019 good quality reads, respectively (Table 3.9). Merged tagCount file, generated by collapsing reads into a set of unique sequence tags called "merged tags" contained 416,048 tags (Table 3.8). Elshire et al. (2011) indicates that inter-sample variation is valid for all multiplex sequencing protocols and sources from the accuracy problems of available DNA quantification methods. The authors highlight the necessity of developing high-throughput methods with improved precision for quantifying high molecular weight DNA for multiplex sequencing analyses. Yet, in our work, only two out of 91 RILs had insufficient sequence data and were excluded from the analysis.

Table 3.8. Sequencing and tag alignment statistics.

Parameter	**Number of sequence**
Raw reads	343,970,622
Accepted reads	164,433,076
Merged tags	416,048
Tags aligned to unique locations	350,151
Tags aligned to multiple locations	33,739
Tags mapped to genome assembly	333,439
Tags mapped to unanchored contigs	50,421
Unaligned tags	32,158

Haploid chromosome number of the sesame genome is 13. The available sesame draft genome assembly (Wang et al., 2014) comprises 16 pseudomolecules of assembled scaffolds that represent the 13 chromosomes and a 17^{th} group of unanchored, concatenated contigs. The output of tag alignment to the draft assembly (TOPM file) contained 416,048 tags (merged tags) (Table 3.8) which represented 157,667,909 tag sequences from the RIL population and the parental accessions (Table 3.9). As a result of the alignment, 383,890 tags (92.27%) were aligned to the assembly, with 350,151 tags being mapped to unique locations (91.21%) and the rest (33,739 tags) aligning to multiple loci. Out of the 383,890 GBS tags located in the draft assembly, 333,439 were mapped to the 16 pseudomolecules of the draft genome assembly and 50,421 tags were mapped to the group of concatenated sequence of unanchored contigs (Table 3.8).

Table 3.9. Sequencing and tag alignment statistics per genotype.

Genotype	Accepted reads	Frequency in accepted reads (%)	Tags represented in TOPM†	Frequency in TOPM†
S. indicum (Acc. No. 95-223)	1,523,490	0.93	1,459,290	0.93
S. indicum (Acc. No. 92-3091)	2,920,019	1.78	2,818,155	1.79
Sesame_RIL 1	930,521	0.57	889,707	0.56
Sesame_RIL 2	1,121,172	0.68	1,077,899	0.68
Sesame_RIL4	1,191,025	0.72	1,144,867	0.73
Sesame_RIL 5	1,829,666	1.11	1,745,654	1.11
Sesame_RIL 7	2,814,520	1.71	2,704,172	1.72
Sesame_RIL 8	2,513,107	1.53	2,423,901	1.54
Sesame_RIL 9	1,503,664	0.91	1,446,923	0.92
Sesame_RIL 12	1,872,386	1.14	1,799,663	1.14
Sesame_RIL 14	1,769,800	1.08	1,703,906	1.08
Sesame_RIL 15	1,574,061	0.96	1,508,007	0.96
Sesame_RIL 16	2,242,974	1.36	2,161,678	1.37
Sesame_RIL 17	2,854,960	1.74	2,747,197	1.74
Sesame_RIL 19	867,654	0.53	830,83	0.53
Sesame_RIL 20	1,192,380	0.73	1,150,462	0.73
Sesame_RIL 21	1,858,041	1.13	1,795,363	1.14
Sesame_RIL 22	2,089,046	1.27	2,013,498	1.28
Sesame_RIL 23	1,169,249	0.71	1,080,049	0.69
Sesame_RIL 24	2,512,737	1.53	2,418,574	1.53
Sesame_RIL 26	2,223,914	1.35	2,144,321	1.36
Sesame_RIL 27	1,694,244	1.03	1,613,159	1.02
Sesame_RIL 28	1,575,259	0.96	1,500,639	0.95
Sesame_RIL 29	1,442,658	0.88	1,372,247	0.87
Sesame_RIL 30	1,573,969	0.96	1,505,593	0.95
Sesame_RIL 31	1,674,155	1.02	1,606,546	1.02
Sesame_RIL 32	545,782	0.33	526,711	0.33
Sesame_RIL 34	963,206	0.59	892,612	0.57
Sesame_RIL 35	2,215,059	1.35	2,133,783	1.35
Sesame_RIL 36	2,046,845	1.24	1,967,143	1.25
Sesame_RIL 38	1,406,305	0.86	1,319,989	0.84
Sesame_RIL 39	2,227,729	1.35	2,146,122	1.36
Sesame_RIL 41	2,288,544	1.39	2,175,938	1.38
Sesame_RIL 42	1,341,709	0.82	1,288,585	0.82
Sesame_RIL 43	2,590,831	1.58	2,498,700	1.58
Sesame_RIL 44	1,898,107	1.15	1,825,912	1.16
Sesame_RIL 45	0	0.00	0	0.00

(Cont. on next page)

Table 3.9 (cont.)

Genotype	Accepted reads	Frequency in accepted reads (%)	Tags represented in TOPM†	Frequency in TOPM†
Sesame_RIL 46	617,52	0.38	594,498	0.38
Sesame_RIL 50	1,076,138	0.65	1,030,184	0.65
Sesame_RIL 51	1,084,699	0.66	1,044,511	0.66
Sesame_RIL 52	2,871,237	1.75	2,767,442	1.76
Sesame_RIL 54	1,707,486	1.04	1,643,525	1.04
Sesame_RIL 56	1,090,323	0.66	1,050,122	0.67
Sesame_RIL 58	1,496,295	0.91	1,443,600	0.92
Sesame_RIL 59	2,068,711	1.26	1,896,724	1.20
Sesame_RIL 60	1,777,004	1.08	1,707,021	1.08
Sesame_RIL 61	2,053,130	1.25	1,969,416	1.25
Sesame_RIL 63	1,453,287	0.88	1,390,052	0.88
Sesame_RIL 64	2,740,553	1.67	2,607,339	1.65
Sesame_RIL 65	1,615,166	0.98	1,550,859	0.98
Sesame_RIL 67	13,592	0.01	12,942	0.01
Sesame_RIL 68	1,270,425	0.77	1,188,588	0.75
Sesame_RIL 69	1,462,313	0.89	1,403,616	0.89
Sesame_RIL 70	2,003,814	1.22	1,928,678	1.22
Sesame_RIL 71	1,873,527	1.14	1,799,440	1.14
Sesame_RIL 73	1,838,067	1.12	1,770,137	1.12
Sesame_RIL 74	2,245,467	1.37	2,165,417	1.37
Sesame_RIL 77	1,805,192	1.10	1,737,633	1.10
Sesame_RIL 79	3,296,818	2.00	3,143,459	1.99
Sesame_RIL 80	1,493,636	0.91	1,433,339	0.91
Sesame_RIL 82	312,635	0.19	295,303	0.19
Sesame_RIL 83	1,675,021	1.02	1,610,413	1.02
Sesame_RIL 84	1,756,462	1.07	1,690,308	1.07
Sesame_RIL 85	796,029	0.48	766,34	0.49
Sesame_RIL 87	2,181,717	1.33	2,102,007	1.33
Sesame_RIL 88	2,300,381	1.40	2,220,268	1.41
Sesame_RIL 90	189,834	0.12	182,57	0.12
Sesame_RIL 91	2,114,361	1.29	2,039,021	1.29
Sesame_RIL 92	2,785,509	1.69	2,683,105	1.70
Sesame_RIL 94	2,368,541	1.44	2,286,626	1.45
Sesame_RIL 95	2,216,317	1.35	2,130,608	1.35
Sesame_RIL 96	1,460,288	0.89	1,415,153	0.90
Sesame_RIL 97	2,192,066	1.33	2,111,784	1.34
Sesame_RIL 98	1,667,886	1.01	1,603,238	1.02
Sesame_RIL 99	1,158,348	0.70	1,107,359	0.70
Sesame_RIL 100	1,266,588	0.77	1,222,804	0.78

(Cont. on next page)

Table 3.9 (cont.)

Genotype	Accepted reads	Frequency in accepted reads (%)	Tags represented in TOPM†	Frequency in TOPM†
Sesame_RIL 101	1,512,015	0.92	1,453,686	0.92
Sesame_RIL 102	1,995,333	1.21	1,925,407	1.22
Sesame_RIL 103	2,068,642	1.26	1,991,262	1.26
Sesame_RIL 104	2,085,082	1.27	2,009,840	1.27
Sesame_RIL 105	1,944,313	1.18	1,875,859	1.19
Sesame_RIL 106	2,100,441	1.28	2,016,417	1.28
Sesame_RIL 107	3,250,325	1.98	3,094,714	1.96
Sesame_RIL 108	2,381,024	1.45	2,286,562	1.45
Sesame_RIL 109	2,650,728	1.61	2,554,100	1.62
Sesame_RIL 110	2,461,600	1.50	2,355,441	1.49
Sesame_RIL 111	1,714,846	1.04	1,652,910	1.05
Sesame_RIL 112	1,508,189	0.92	1,434,095	0.91
Sesame_RIL 113	2,577,714	1.57	2,344,124	1.49
Sesame_RIL 114	2,169,364	1.32	2,099,110	1.33
Sesame_RIL 115	2,226,903	1.35	2,153,749	1.37
Sesame_RIL 116	2,331,218	1.42	2,237,228	1.42
Sesame_RIL 117	168	0.00	161	0.00
TOTAL	164,433,076	100	157,667,909	100

† TOPM: Tags On Physical Map

3.3.2. SNP calling and filtering

Sequence reads sorted by taxa were used in conjunction with the TOPM file for SNP calling from the tag alignment. As a result, a total of 16,731 raw SNPs were identified, 15,929 (95.21%) of which were mapped to the 16 pseudomolecules of the draft genome assembly and 802 SNPs (4.79%) mapped to unanchored contigs (Table 3.10). Raw SNPs were further processed by merging duplicates (redundant SNP loci identified in reads from both directions). When duplicates were merged, resultant number of unique SNPs was 15,521, with 14,786 SNPs (95.26%) mapped in the draft genome assembly and 735 SNPs (4.73%) mapped to the 17^{th} group consisting of unanchored contigs (Table 3.10). Location and allelic variant information for the merged SNPs can be accessed at http://plantmolgen.iyte.edu.tr/data/. With 2361 raw and 2192 merged SNPs, the highest number of SNPs were mapped to the pseudomolecule 3

of the draft assembly. Conversely, the lowest number of SNPs (137 raw and 116 merged SNPs) were mapped to the pseudomolecule 16 of the assembly (Table 3.10).

Presence of a reference genome introduces important advantages to GBS analysis. Conversely, linkage maps constructed by marker identification and genotyping through GBS assist the efforts for the improvement of genome assemblies by correcting the order and orientation of contigs. In case a well-established reference genome with contigs assembled and ordered based on a verified consensus map exists, marker order can be determined without the requirement of linkage analysis (Poland and Rife, 2012). However, such an advantage does not hold for sesame, since not all sequences within the available genome assembly (Wang et al., 2014) could be oriented and ordered, due to the insufficient number of markers used for anchoring sequences to a linkage map. In addition, the assembly consists of 16 pseudomolecules instead of 13 (the haploid chromosome number of *S. indicum*). As a result, linkage analysis is a prerequisite for grouping and ordering identified markers. Yet, the presence of a genome assembly was highly beneficial for us, since it complemented an important shortcoming of GBS analysis. GBS produces sequence tags of maximum 64 bp length and the library preparation protocol, which involves the use of a single restriction enzyme (ApeKI), does not yield overlapping fragments which would enable sequence assembly. Therefore, unless SNP markers identified by GBS are located in a sequence context, they cannot be further utilized for genotyping in other populations and their utilization remains exclusive to the population genotyped in GBS analysis. In this work, it was feasible to locate the 15,521 putative SNPs in draft assembly sequences, which will enable primer/probe design for the further utilization of these large number of SNP markers for genotyping any sesame population.

In order to determine a core set of SNPs appropriate for use in segregation analyses for gene/QTL mapping in the genotyped F_6-RIL population, merged SNPs were filtered for minimum taxa coverage, minimum locus coverage, and LD. Because GBS analysis yielded an excess of SNPs, a sufficient number of SNPs (781 SNPs) were still retained as a result of filtering and used in linkage analysis for the construction of a genetic linkage map of the sesame genome.

Table 3.10. SNP calling and filtering statistics.

Pseudomolecule	Raw SNPs	Merged SNPs	Filtered SNPs
1	1612	1496	11
2	908	844	31
3	2361	2192	29
4	1071	995	53
5	1221	1116	58
6	1053	970	61
7	1015	960	40
8	1061	991	19
9	960	906	5
10	1010	939	30
11	1078	1023	28
12	1044	986	5
13	430	366	201
14	202	182	12
15	766	704	7
16	137	116	12
17	802	735	179
TOTAL	16,731	15,521	781
% in assembly†	95.21	95.26	77.08
% in contigs‡	4.79	4.73	22.92

† Percentage of SNPs mapped to reference genome pseudomolecules.
‡ Percentage of SNPs mapped to unassembled contigs.

3.3.3. Construction of a genetic linkage map

In this study, GBS proved superior in terms of polymorphism identification (15,521 SNPs) (Table 3.10), compared to SLAF-seq (Zhang et al., 2013) and RAD tag sequencing (Wu et al., 2014b) approaches, which identified 3673 and 3769 polymorphic loci in the sesame genome, respectively. The percentage of SNP markers used for linkage analysis in the RIL population was 5.03% (781 markers), lower than those reported by Zhang et al. (2013) and Wu et al. (2014b), who incorporated 34.63% (1272 markers) and 35.21% (1327 markers) of the markers they identified into their linkage analyses, respectively. Presumably, this was the result of the high stringency SNP filtering protocol applied in this work, which included a high LD filter for RIL populations that eliminated SNPs with high rates of genotyping errors. Despite this higher stringency, a core set of 781 markers proved sufficient for the construction of a high-resolution genetic linkage map of the sesame genome using the JoinMap 4.0 program (Figure 3.5).

A total of 831 markers (781 SNPs and 50 SSRs) were used to construct the linkage map (Figure 3.5). Out of the 831 markers, 730 (716 SNPs, 14 SSRs) were mapped into 15 linkage groups (LGs) that span a total genetic distance of 999 cM (Table 3.11). Among mapped SNPs, 178 (24.86%) were identified in the unanchored contigs of the draft genome assembly and the rest (75.14%) were identified in the 16 pseudomolecule sequences. Average marker density of the map was 1.37 cM per marker interval (Table 3.11), similar to that reported by Zhang et al. (2013) (1.21 cM), who did the first work of high-throughput marker development and mapping in the sesame genome by multiplex next-generation sequencing. Among the 15 linkage groups, two were designated as LG8A and L8B, since both groups harbored SNP loci located in the same pseudomolecule (Pseudomolecule 3) of the draft genome assembly. Similarly, the two linkage groups with SNP markers located in Pseudomolecule 10 were designated as LG10A and LG10B (Table 3.11). As a result, the expected number of linkage groups that corresponds to the haploid chromosome number of sesame (13) was obtained by linkage analysis.

Table 3.11. Distribution of markers in linkage groups.

Linkage group	Number of markers†	Size (cM)	Marker density (cM per marker interval)
LG1	204 (13, 17)	87.35	0.43
LG2	167 (4, 6, 16, 17)	188.89	1.13
LG3	85 (7, 17)	153.96	1.81
LG4	61 (5, 17)	55.58	0.91
LG5	36 (17)	91.75	2.55
LG6	35 (11, 17)	114.69	3.28
LG7	27 (2)	21.71	0.80
LG8A	20 (3)	85.23	4.26
LG8B	20 (3, 17)	61.46	3.07
LG9	14 (17)	11.79	0.84
LG10	13 (17)	5.90	0.45
LG11A	12 (10)	22.28	1.86
LG11B	13 (10)	36.86	2.84
LG12	10 (17)	12.25	1.22
LG13	13 (1, 17)	49.29	3.79
Total	**730**	**998.99**	**1.37**

†SNP locations in draft genome assembly pseudomolecules are given in parantheses.
Pseudomolecule 17 corresponds to unanchored sequences.

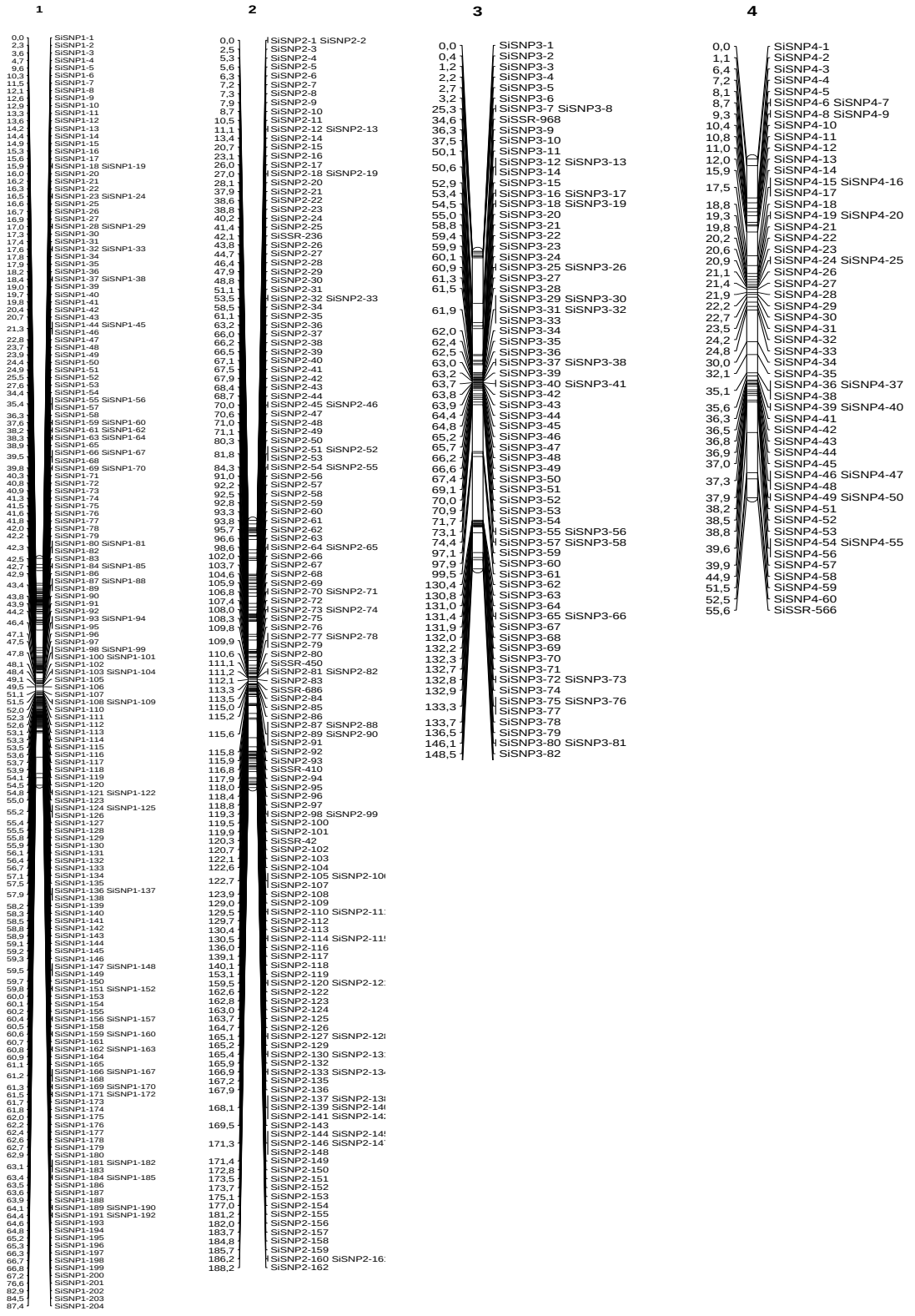

Figure 3.5. Genetic linkage map of the sesame genome constructed by GBS analysis. Marker locations (cM) are displayed on the left side of each linkage group.

(Cont. on next page)

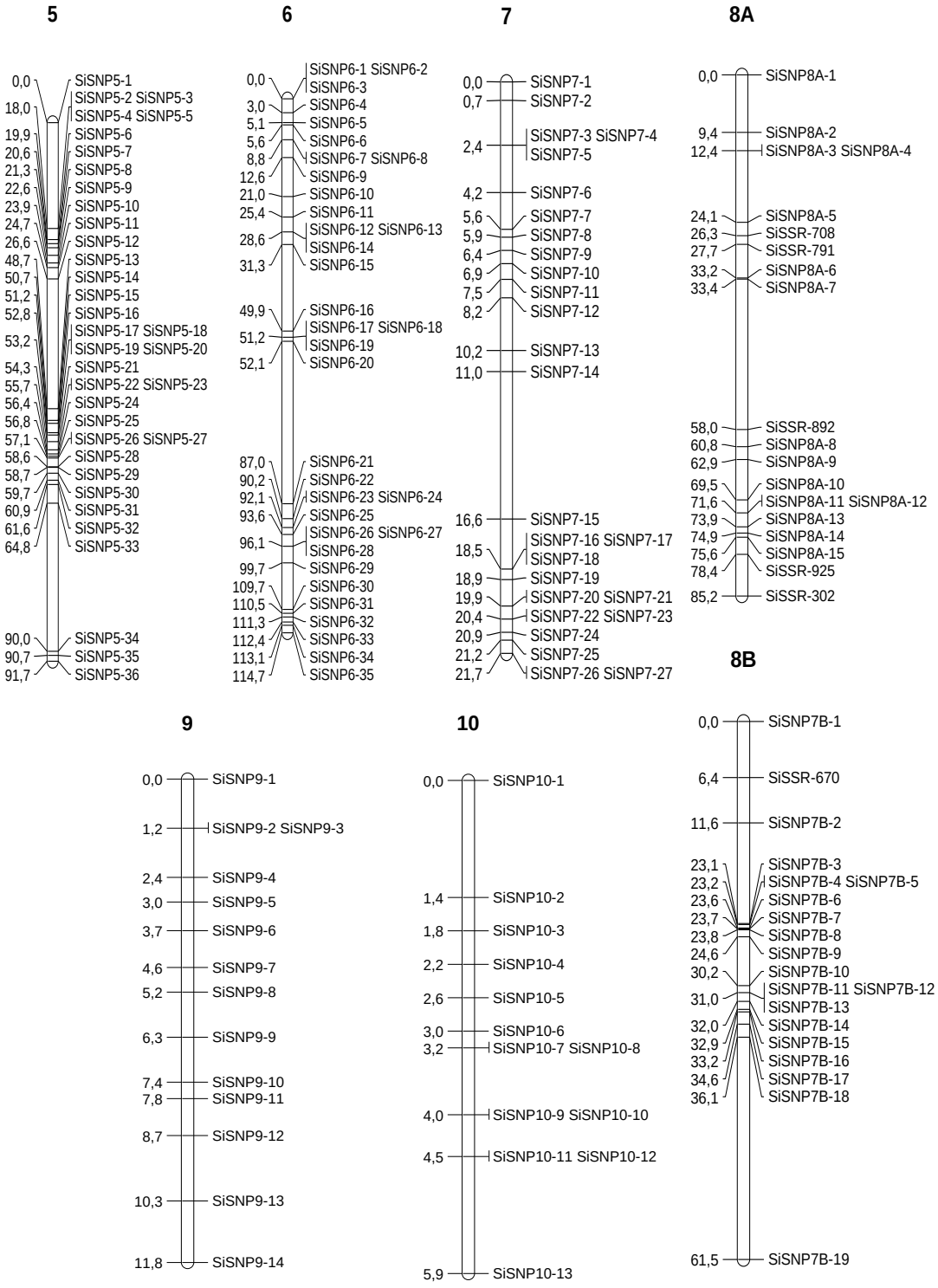

Figure 3.5 (cont.)

(Cont. on next page)

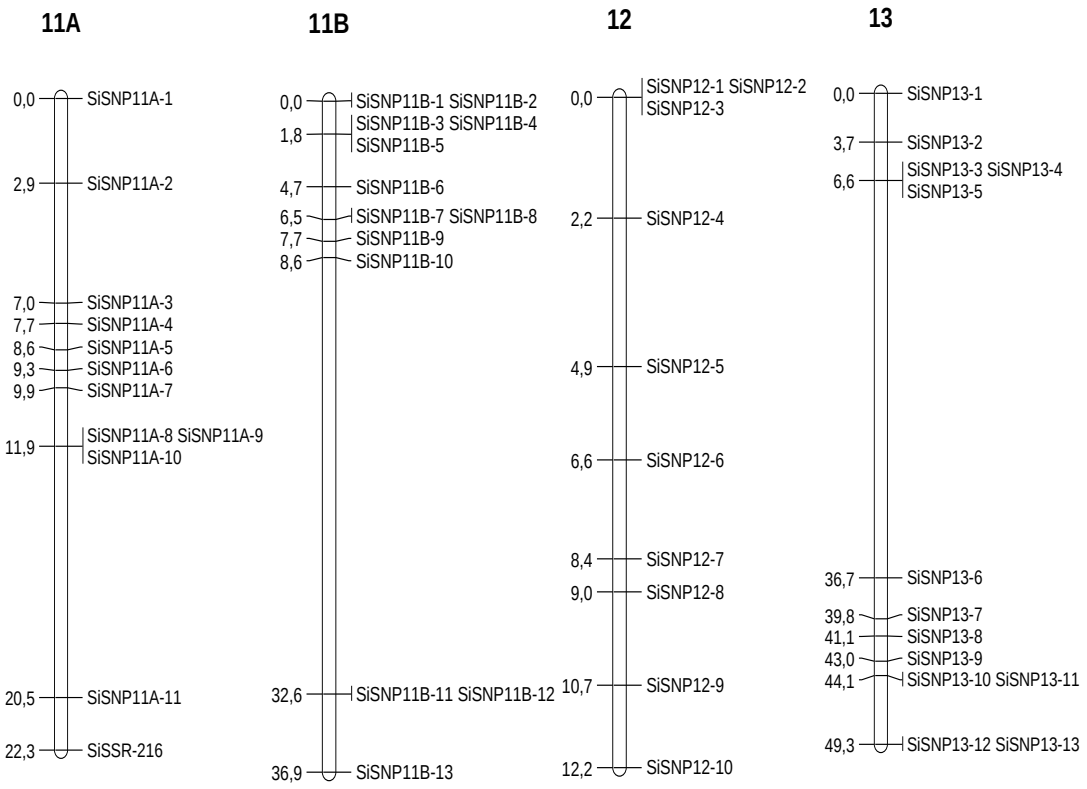

Figure 3.5 (cont.)

LG1 and LG12 had the highest and lowest number of markers, with 204 and 10 SNPs, respectively. LG1 also displayed the lowest average distance between adjacent markers (0.43 cM). LG2 encompassed a genetic distance of 188.89 cM, which was the largest genetic distance represented by the linkage groups. LG2 had 167 markers (162 SNPs and 5 SSRs) and displayed a marker density of 1.13 cM per marker interval (Table 3.11). The linkage group with the lowest number of markers was LG12 with 10 markers. The shortest genetic distance spanned by a linkage group was 5.90 cM (LG10). LG8A displayed the lowest marker density of 4.26 cM with 20 markers (15 SNPs and 5SSRs) encompassing a genetic distance of 85.23 cM (Table 3.10). Transition mutations were predominant (57.4%) among mapped SNPs, as expected (Table 3.12).

Table 3.12. Statistics of nucleotide substitution types mapped to linkage groups.

Mutation type	Substitution	Number	Frequency (%)
Transition	A↔G	213	29.75
Transition	C↔T	198	27.65
Transversion	A↔C	70	9.78
Transversion	G↔T	73	10.20
Transversion	A↔T	79	11.03
Transversion	G↔C	83	11.59
	Total	**716**	**100**

Out of the 50 SSR markers that were used in linkage analysis, only 14 (28%) were mapped into linkage groups. Similarly, Wu et al. (2014b) were able to map 22 out of 54 SSRs, when they used SSR markers in conjunction with SNP and indel polymorphisms identified by RAD tag sequencing of a sesame RIL population. We attribute the low rate of SSR incorporation into the linkage groups to the fact that reduced representation library protocols are optimized for eliminating repeat-rich sequences to establish a high genic-to-repetitive DNA ratio (Elshire et al., 2011; He et al., 2014), resulting in the under-representation of SSR bearing genomic regions in the tags.

The linkage map had a total of 16 gaps larger than 10 cM located in linkage groups 3, 5, 6, 8A, 8B, 11B and 13 (Figure 3.5). In parallel with the results of this work, Wu et al. (2014b) reported 16 gaps (>10 cM) in their sesame genetic linkage map constructed through a RAD tag sequencing approach. When reduced-representation sequencing protocols are employed, it is reasonable to expect that certain regions of the genome are under-represented. Indeed, the degree of uniformity in genome representation is unknown for sequence tags generated through GBS analysis (Poland et al., 2012).

The major advantage of using recombinant inbred lines in genome mapping studies is that RILs constitute permanent mapping populations. Because segregation is totally or almost complete in RILs, the genotypes can be propagated unlimitedly. Therefore once a linkage map is established, it can continuously be improved by reanalyzing preexisting genotypic data in conjunction with data from new analyses. In trait mapping studies, RIL populations can be confidently evaluated over subsequent years and in many different environments. Therefore, RILs improve the precision of

detecting the genetic component of variance while analyzing quantitative traits (Burr et al. 1988). The parents of the sesame F_6-RIL population used in this work were identified as genetically distinct genotypes by Laurentin and Karlovsky (2006). Moreover, the parental genotypes produced distinct profiles when evaluated for their secondary metabolite content (personal communication with Dr. Petr Karlovsky, University of Gottingen, Germany), and therefore, were used for generating a RIL population in order to study the inheritance of secondary metabolite accumulation in sesame. Thus, the genotypic data and the genetic linkage map produced through GBS provide the essential molecular tools to enable further work to identify the genetic control of agronomic and metabolic traits in sesame.

CHAPTER 4

CONCLUSION

Sesame is an ancient oil seed crop which deserves the reputation as 'Queen of the oil seeds' due to the exceptional quality attributes of its oil (Bedigian and Harlan, 1986). Because the crop is not nutrient demanding, tolerant to high temperatures and drought, and can be utilized as a leafy vegetable, it is a potential food-security crop, especially for the rural regions of Africa (Bedigian, 2003). However, sesame is a neglected crop with very little progress achieved in systematic breeding of the crop for disease resistance and improved yield traits. Molecular genetic research in sesame is relatively recent, therefore, molecular breeding in the crop is also restricted.

In this work, next-generation sequencing approaches were employed for high-throughput SSR and SNP marker development in sesame. Pyrosequencing of the sesame accession *S. indicum* L. cv. Muganli 57 enabled the identification of 5727 SSR loci in a contig assembly. Validity of the SSR development approach was successfully verified by dye-terminator sequencing of SSR fragments. A total of 1000 SSR primers were tested for their amplification efficiency and inter-species transferability. As a result, 933 experimentally validated SSR markers were introduced with 91% of those markers transferable to the putative progenitor of cultivated sesame (*S. mulayanum*). Analysis of the genetic diversity and population structure in a collection of world accessions clustered *S. mulayanum* with accessions from the domestication origin, supporting its assignment as the wild progenitor of cultivated sesame. Moreover, patterns of gene flow among diversity centers were revealed by the analyses, providing useful hints for designing future breeding schemes.

In this work, the GBS approach, which enables simultaneous high-throughput SNP marker development and genotyping, was applied for the first time in sesame. Using an intraspecific F_6-RIL population, a total of 15,521 unique SNPs were identified and genotyped in the population through GBS analysis. Filtering for SNPs with high accuracy of genotype calls in the population resulted in 781 SNPs designated as a core set for genome mapping and QTL analyses in the RIL population. As a result of linkage analysis, a high-resolution genetic linkage map comprising 13 linkage groups that span

a genetic distance of 999 cM with 730 markers was constructed. Genotypic data obtained from GBS analysis and the high-resolution genetic linkage map constitute valuable molecular genetic tools for dissecting the genetic control over important metabolic and agronomic traits in sesame. Overall, the outcomes of this work represent a significant contribution to the existing molecular genetic tools specific for the sesame genome and will enable the acceleration of molecular breeding in this orphan crop species.

The genotypic data for the RIL population and the molecular genetic linkage map constitute the basis of the future work that will involve mapping metabolic and agronomic traits in the sesame genome. Because RILs are immortal genotypes that can be propagated infinitely, it will be feasible to evaluate the population for secondary metabolite accumulation and agronomic traits and analyze the data in conjunction with the genotypic data obtained in this work through GBS analysis. To this end, RILs grown under field conditions will be characterized for the accumulation of major lignans (sesamin and sesamolin) and lignan glucosides in their seeds. QTL analysis will be performed to identify and locate genomic regions that harbor loci controlling the synthesis/accumulation of these health beneficial phenolic compounds. Agronomic characters that will be evaluated will include traits that facilitate mechanized harvesting such as capsule non-dehiscence, plant height and branching pattern. Seed yield will be evaluated on the basis of 1000 seed weight and seed weight per plant. Because seed coat color is associated with oil yield, RILs will also be characterized for this trait. Physiological characters including days to flowering and maturity, and morphological characters such as flower color, anthocyanin pigmentation in stem, leaves and capsules, and stem and leaf hairiness will be scored in RILs. Genomic regions that contribute to the control of agronomic traits will be identified through QTL analysis and linked SNP loci will be determined. As a result, it is anticipated that genetic control over antioxidant lignan accumulation, yield and mechanized harvesting associated traits will be dissected and systematic breeding of these important characters will be feasible through molecular breeding. Therefore, future utilization of the outcomes of this work will be of significant importance in facilitating the development of new sesame cultivars with improved seed yield and superior oil quality due to elevated concentrations of health beneficial antioxidant compounds.

REFERENCES

Altshuler, D.; Pollara, V. J.; Cowles, C. R.; Van Etten, W.J.; Baldwin, J.; Linton, L.; Lander, E. S. An SNP map of the human genome generated by reduced representation shutgun sequencing. *Nature* **2000**, *407*, 513–516.

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

Ashri, A. Sesame Breeding. In *Plant Breeding Reviews*, Janick, J., Eds.; John Wiley and Sons: Oxford, 1998; Vol. *16*, pp 179–228.

Baird, N. A.; Etter, P. D.; Atwood, T. S.; Currey, M. C.; Shiver, A. L.; Lewis, Z. A.; Selker, E. U.; Cresko, W. A.; Johnson, E. A. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **2008**, *3*, e3376.

Bedigian, D.; Harlan, J. R. Evidence for the cultivation of sesame in the ancient world. *Econ. Bot*. **1986**, *40*, 137–154.

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

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

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

Bradbury, P. J.; Zhang, Z.; Kroon, D. E.; Casstevens, T. M.; Ramdoss, Y.; Buckler, E. S. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635.

Brookes, A. J. The essence of SNPs. *Gene* **1999**, *234*, 177–186.

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

Burr, B.; Burr, F. A.; Thompson, K. H.; Albertson, M. C.; Stuber, C. W. Gene mapping with recombinant inbreds in maize. *Genetics* **1988**, *118*, 519–526.

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

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

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

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

Çağırgan, M. I. Mutation techniques in sesame (Sesamum indicum L.) for intensive management: confirmed mutants. In *Sesame improvement by induced mutations*, IAEA-TECDOC-1195, IAEA, Vienna, 2001, pp 31–40.

Day, J. S. Development and maturation of sesame seeds and capsules. *Field Crops Res*. **2000**, *67*, 1–9.

Dixit, A.; Jin, M. H.; Chung, J. W.; Yu, J. W.; Chung, H. K.; Ma, K. H.; Park, Y. J.; Cho, E. G. Development of polymorphic microsatellite markers in sesame (*Sesamum indicum* L.). *Mol. Ecol. Notes* **2005**, *5*, 736–738.

Earl, D. A.; Von Holt, B. M. Structure Harvester: a website and program for visualizing Structure output and implementing the Evanno method. *Conserv. Genet. Resour*. **2012**, *4*, 359–361.

Elshire, R. J.; Glaubitz, J. C.; Sun, Q.; Polland, J. A.; Kawamoto, K.; Buckler, E. S.; Mitchell, S. E. A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS One* **2011**, *6*, e19379.

Glaubitz, J. C.; Casstevens, T. M.; Lu, F.; Harriman, J; Elshire, R. J.; Sun, Q; Buckler, E. S. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **2014**, *9*, e90346.

Gu, T.; Tan, S.; Gou, X.; Araki, H.; Tian, D. Avoidance of long mononucleotide repeats in codon pair usage. *Genetics* **2010**, *186*, 1077–1084.

Harlan J. R. Crops and Man. 2nd edition. Agronomy Society of America, Madison, WI, 2010.

Hart, J. P.; Griffiths, P. D. Genotyping-by-Sequencing enabled mapping and marker development for the *By-2* Potyvirus resistance allele in common bean. *The Plant Genome* **2015**, 10.3835/plantgenome2014.09.0058.

He, J.; Zhao, X.; Laroche, A.; Lu, Z. X.; Liu, H. K.; Li. Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelarate plant breeding. *Frontiers in Plant Science* **2014**, *5*, 484.

Ihlenfeldt H. D.; Grabow-Seidensticker, U. The genus Sesamum and the origin of the cultivated sesame. In *Taxonomic Aspects of African Economic Botany*, Kunkel, G., Eds.; Excmo Ayuntamiento, Las Palmas de Gran Canaria, 1979; pp 53–60.

Islam, M. S.; Thyssen, G. N.; Jenkins, J. N.; Fang, D. D. Detection, validation, and application of Genotyping-by-Sequencing based single nucleotide polymorphisms in Upland cotton. *The Plant Genome* **2015**, 10.3835/plantgenome2014.07.0034.

Jaganathan, D.; Thudi, M.; Kale, S.; Azam, S.; Roorkiwal, M.; Gaur, P. M.; Kishor, P. B. K.; Nguyen, H.; Sutton, T.; Varshney, R. K. Genotyping-by-sequencing based intra-specific genetic map refines a "QTL-hotspot" region for drought tolerance in chickpea. *Mol. Genet. Genomics* **2015**, *290*, 559–571.

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

Jones, N.; Ougham, H.; Thomas, H.; Pasakinskiene, I. Markers and mapping revisited: finding your gene. *New Phytol.* **2009**, *183*, 935–966.

Kawase, M. 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.* **2000**, *44*, 115–122.

Kim, D. H.; Zur, G.; Danin-Poleg, Y.; Lee, S. W.; Shim, K. B.; Kang, C. W.; Kashi, Y. Genetic relationships of sesame germplasm collection as revealed by inter-simple sequence repeats. *Plant Breeding* **2003**, *121*, 259–262.

Kobayashi, T. The Path of Sesame. Iwanami Shoten, Japan, 1986.

Kosambi, D. D. The estimation of map distances from recombination values. *Ann Eugen*, **1943**, *12*, 172–175.

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

Laurentin, H.; Ratzinger, A.; Karlovsky, P. Relationship between metabolic and genomic diversity in sesame (*Sesamum indicum* L.). *BMC Genomics* **2008**, *9*, 250.

Lawson, M. J.; Zhang, L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol.* **2006**, *7*, R14.

Lipman, D. J.; Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435–1441.

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

Morris, J. B. Characterization of sesame (*Sesamum indicum* L.) germplasm regenerated in Georgia, USA. *Genet. Resour. Crop. Evol.* **2009**, *56*, 925–936.

Namiki M. Nutraceutical functions of sesame: A review. *Crit. Revs. In Food Sci. Nutr.* **2007**, *47*, 651–673.

Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Nat. Acad. Sci.* **1973**, *70*, 3321–3323.

Orsouw, N. J.; Hogers, R. C. J.; Janssen, A.; Yalçın, F.; Snoeijers, S.; Verstege, E.; Schneiders, H.; Van der Poel, H., Van Oeveren, J.; Verstegen, H.; Van Eijk, M. J. T. Complexity reduction of polymorphic sequences (CROPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* **2007**, *11*, e1172.

Poland, J. A.; Rife, T. W. Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome* **2012**, *5*, 92–102.

Powell, W.; Machray, G. C.; Provan, J. Polymorphism revealed by simple sequence repeats. *Trends Plant Sci*. **1996**, *1*, 215–222.

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

Rafalski, J. A. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci.* **2002**, *162*, 329–333.

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

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

Russell, J.; Hackett, C.; Hedley, P.; Liu, H.; Milne, L.; Bayer, M.; Marshall, D.; Jorgensen, L.; Gordon, S.; Brennan, R. The use of genotyping by sequencing in blackcurrant (Ribes nigrum): developing high-resolution linkage maps in species without reference genome sequences. *Mol. Breed.* **2014**, *33*, 835–849.

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

Spandana, B.; Reddy, V. P.; Prasanna, G. J.; Anuradha, G.; Sivaramakrishnan, S. Development and characterization of microsatellite markers (SSR) in *Sesamum* (*Sesamum indicum* L.) species. *Appl. Biochem. Biotechnol.* **2012**, *168*, 1594–1607.

Sun X.; Liu, D.; Zhang, X.; Li, W.; Liu, H.; Hong, W., Jiang, C.; Guan, N.; Ma, C.; Zeng, H.; Xu, C.; Song, J.; Huang, L.; Zheng, H. SLAF-seq: an efficient method of large-scale De novo SNP discovery and genotyping using high-throughput sequencing. *PLoS One* **2013**, *8*, e58700.

Uzun B.; Çağırgan, M. I. Identification of molecular markers linked to determinate growth habit in sesame. *Euphytica* **2009**, *166*, 379–384.

Uzun B.; Lee, D.; Donini, P.; Çağırgan, M. I. Identification of a molecular marker linked to closed capsule mutant trait in sesame using AFLP. *Plant Breeding* **2003**, *122*, 95–95.

Van Ooijen, J. W. JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B. V., Wageningen, Netherlands, 2006.

Voorrips, R. E. MapChart: software for the graphical presentation of linkage maps and QTLs. *The Journal of Heredity*, **2002**, *93*, 77–78.

Wang L.; Zhang, Y., Qi, X.; Gao, Y.; Zhang, X. Development and characterization of 59 polymorphic cDNA-SSR markers for the edible oil crop *Sesamum indicum* (Pedaliaceae). *Am. J. Bot.* **2012**, *99*, e394–398.

Wang, L.; Yu, S.; Tong, C.; Zhao, Y.; Liu, Y.; Song, C.; Zhang, Y.; Zhang, X.; Wang, Y.; Hua, W.; Li, D.; Li, D.; Li, F.; Yu, J.; Xu, C.; Han, X.; Huang, S.; Tai, S.; Wang, J.; Xu, X.; Li, Y.; Liu, S.; Varshney, R. K.; Wang, J.; Zhang, X. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol.* **2014**, *15*, R39.

Wei, L. B.; Zhang, H. Y.; Zheng, Y. Z.; Guo, W. Z.; Zhang, T. Z. Developing EST-Derived Microsatellites in Sesame (*Sesamum indicum* L.). *Acta Agron. Sin*. **2008**, *34*, 2077–2084.

Wei L. B; Zhang, H. Y.; Zheng, Y. Z.; Miao, H. M.; Zhang, T. Z.; Guo, W. Z. A genetic linkage map construction for sesame (*Sesamum indicum* L.). *Genes Genom.* **2009**, *31*, 199–208.

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

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

Wu, K.; Yang, M.; Liu, H.; Tao, Y.; Mei, J.; Zhao, Y. Genetic analysis and molecular characterization of Chinese sesame (*Sesamum indicum* L.) cultivars using Insertion-Deletion (InDel) and Simple Sequence Repeat (SSR) markers. *BMC Genet*. **2014a**, *15*, 35.

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

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

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

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

Zou, X.; Shi, C.; Austin, R. S.; Merico, D.; Munholland, S.; Marsolais, F.; Navabi, A.; Crosby, W. L.; Pauls, K. P.; Yu, K.; Cui, Y. Genome-wide single nucleotide polymorphism and insertion-deletion discovery through next-generation sequencing of reduced representation libraries in common bean. *Mol. Breed.* **2014**, *33*, 769–778.

VITA

Ayşe Özgür UNCU

Nationality: Turkish

Place of birth: Izmir, Turkey

Marital status: Married

E-mail: aysedulger@yahoo.com.tr

EDUCATION

PhD, Department of Molecular Biology and Genetics, Izmir Institute of Technology, Turkey, 2015

MSc, Department of Horticultural Genetics and Biotechnology, Mediterranean Agronomic Institute of Chania, Greece, 2010

DSPU (Degree of Postgraduate Specialization), Department of Horticultural Genetics and Biotechnology, Mediterranean Agronomic Institute of Chania, Greece, 2008

BSc, Deparment of Food Engineering, Celal Bayar University, Turkey, 2007

PUBLICATIONS

Uncu A.T.[1], **UNCU, A.O.[1]**, Doğanlar, S., Frary, A. 2015. Authentication of Botanical Origin in Herbal Teas by Plastid Non-coding DNA Length Polymorphisms. Journal of Agricultural and Food Chemistry, doi: 10.1021/acs.jafc.5b01255

UNCU, A.O., Uncu, A.T., Celik, I., Doganlar, S., Frary, A. 2015. A Primer to Molecular Phylogenetic Analysis in Plants. Critical Reviews in Plant Sciences, (In Press)

UNCU, A.O., Gultekin, V., Allmer, J., Frary, A., Doğanlar, S. 2015. Genomic SSR Markers Reveal Patterns of Genetic Relatedness and Diversity in Sesame. The Plant Genome, doi:10.3835/plantgenome2014.11.0087.

UNCU, A.O., Doğanlar, S., Frary, A. 2013. Biotechnology for enhanced nutritional quality in Plants. Critical Reviews in Plant Sciences, 32: 321-343.

Bazakos, C[1]., **DULGER, A.O.[1]**, Uncu, A.T.[1], Spaniolas, S., Spano, T., Kalaitzis, P. 2012. A SNP-based PCR–RFLP capillary electrophoresis analysis for the identification of the varietal origin of olive oils. Food Chemistry, 134: 2411-2418.

AWARDS AND HONORS

Mediterranean Agronomic Institute of Chania, CIHEAM scholar, 2007-2010.